A TRUST-CENTRIC FRAMEWORK FOR SECURE, USER-CENTRIC AI SYSTEMS WITH ACTIONABLE AND PERSONALIZED RECOURSE

by

OLUWAFEYISAYO O. OYENIYI

A dissertation submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE AND INFORMATICS

2026

Oakland University
Rochester, Michigan

Doctoral Advisory Committee:

Amartya Sen, Ph.D., Chair
Lanyu Xu, Ph.D.
Sunny Raj, Ph.D.
Eralda Caushaj, Ph.D.

To my mother, who chose to invest in her daughters and, through her courage and vision, showed us that dreaming big is always worth the leap. To my sisters, whose laughter, encouragement, and unwavering presence have been a constant light beside me. And to the water walkers, who step bravely into uncertain tides, wading through the waves even when the path is unclear, trusting that something greater lies just beyond the horizon.

ACKNOWLEDGMENTS

This dissertation's completion marks the culmination of years of learning, perseverance, and personal growth. I am deeply grateful for the guidance, encouragement, and support of the many individuals and institutions who have accompanied me on this journey.

I extend my heartfelt gratitude to my academic advisor, Dr. Amartya Sen, for his mentorship, presence, and unwavering support throughout this research journey. His insight, encouragement, intentionality, and patience have been instrumental in shaping both this dissertation and my development as a researcher. I am also grateful to my committee members, Dr. Lanyu Xu, Dr. Eralda Caushaj, and Dr. Sunny Raj, whose thoughtful feedback, constructive critiques, and encouragement strengthened this work in countless ways.

I would like to thank Dr. Guangzhi Qu and the Department of Computer Science and Engineering for their financial support through teaching and research assistantships, which made this research possible. Their investment enabled me to pursue ambitious goals and access the resources necessary to advance my work. I am also grateful to external collaborators Dr. Kenneth Fletcher and Dr. Ayan Roy, as well as to Shreyansh Sandip Dhandhukia, for their enriching contributions. My sincere thanks go to my research group members, Victorine Wakam Younang and Dr. Damilola Alao, for providing a supportive environment, intellectual camaraderie, thoughtful proofreading, and friendship.

This journey would have been impossible without the steadfast love and support of my family. To God, the author and finisher of my faith, thank you for life, presence, guidance, and all your blessings. To my mother, Mrs. Bolanle Oyeniyi, thank you for your enduring love, prayers, and for teaching me to dream boldly. To my sisters, Oluwabusayo and Similoluwa Oyeniyi, I am eternally grateful for your love, humor, encouragement, and

presence throughout this journey. To Benedicta Abodunrin, thank you for your unwavering commitment, support, and continuous encouragement. I also extend my gratitude to the Taiwo, Akinbobuyi and Adeniyi families, Mr. Damilola Alao, and Dr. Pfeiffer for their steadfast support.

I am profoundly thankful for the invaluable emotional support from friends: Olaoluwa Adeogo, Yetunde Fawehinmi, and Oluseyi Olanihun for constant encouragement; Ahmed Adenrele, Olubunmi Ayo-Ogunmoko, Naomi Mogridge, Promice Mosely, Eucharia Tagoe, Lovelyn Epelle, Ayomide, and Oluwakemi Yusuf, for their listening ears, motivation, and encouragement.

Finally, to all those I have interacted with in professional and personal circles, your companionship has been a colorful addition to my journey. I am deeply grateful for your support and the role each of you has played in bringing this work to fruition.

Oluwafeyisayo O. Oyeniyi

ABSTRACT

A TRUST-CENTRIC FRAMEWORK FOR SECURE, USER-CENTRIC AI SYSTEMS WITH ACTIONABLE AND PERSONALIZED RECOURSE

by

Oluwafeyisayo O. Oyeniyi

Artificial Intelligence (AI) applications increasingly mediate high-stakes decisions in domains such as finance, healthcare, and social welfare. Yet their deployment often exposes users and institutions to risks arising from bias, opacity, and technical vulnerabilities. High predictive accuracy alone fails to ensure trust, particularly in contexts where user engagement, recourse, and accountability are critical. These challenges motivate a principled, layered approach to operationalizing trust in AI.

This dissertation presents a trust-centric AI framework that formalizes trust as a multi-layer architectural objective. Trust is structured across three interdependent layers: *technical trust*, ensuring model integrity, security, and resilience against adversarial attacks; *cognitive and practical trust*, realized through causally grounded, feasible, cost-aware, and interpretable counterfactual explanations; and *user-aligned trust*, achieved by leveraging preference-aware, lexicographic optimization that personalizes actionable recourse. By aligning these layers through secure deployment mechanisms, causal verification, and multi-objective optimization, the framework shows how failures in security, explainability, or personalization can be systematically mitigated within a coherent trust-centric design

The framework is evaluated within tabular and structured decision-making environments, providing a testbed for secure model deployment, causally grounded explanations, and preference-aware actionable recourse.These contexts illustrate how layered trust can reconcile predictive accuracy, interpretability, and user agency under uncertainty and high

personal stakes. By embedding verifiable, actionable, and user-aligned trust into AI-driven systems, this work supports responsible deployment in high-stakes domains and provides technical guidance for policymakers, organizations, and system architects seeking robust, transparent, and user-centered AI.

TABLE OF CONTENTS

TABLE OF CONTENTS—Continued

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES—Continued

CHAPTER ONE

INTRODUCTION

In many domains, such as finance [2], healthcare [3], and recruitment [4], Artificial Intelligence (AI) applications have evolved from experimental tools to core infrastructure governing high-stakes decisions. This shift is characterized by embedding predictive models and automated decision rules into operational workflows that directly affect individuals' life opportunities and institutional outcomes. Growing confidence in the accuracy and consistency of AI-driven systems has enabled organizations to operate at a scale and efficiency unattainable by human decision-makers alone, positioning these systems as critical infrastructure for high-stakes decision-making [5, 6, 7]. As a result, automated decision-making is now deeply embedded in social, economic, and institutional systems, increasing societal reliance on AI-driven systems at unprecedented scale [8].

However, the expansion of AI into high-stakes domains has exposed systemic limitations that challenge its legitimacy and safe deployment. These challenges emerge across multiple dimensions. First, AI-driven systems often lack transparency, making it difficult for users, regulators, and affected individuals to understand how decisions are generated, thereby constraining accountability [9]. Second, biases embedded in historical training data can be amplified, resulting in discriminatory outcomes in domains such as credit scoring and hiring, disproportionately affecting marginalized groups [10]. Third, AI-driven systems are vulnerable to technical threats, including adversarial attacks, data poisoning, and model manipulation, which can compromise system integrity and user safety [11]. Fourth, the integration of automated decision rules into institutional processes frequently reduces user agency, limiting individuals' ability to contest or meaningfully respond to decisions that affect their lives [12]. Finally, responsibility for errors or harms is often diffused among

developers, deployers, users, and automated systems themselves, complicating legal and ethical accountability [13].

These challenges are not isolated shortcomings but reflect distinct dimensions along which AI-driven systems may fail to earn or sustain trust. As a result, *trust* has emerged as a foundational requirement for the effective deployment of AI systems, particularly in high-stakes and user-facing contexts [9, 13]. Yet trust is frequently presumed to emerge from model performance alone rather than deliberately engineered as a system-level objective [14, 15, 16]. Such assumptions risk rejection, misuse, or uncritical over-reliance, undermining both practical utility and societal legitimacy. Trust in AI is inherently multidimensional. It encompasses interrelated properties including reliability, security, transparency, fairness, accountability, and user agency. These properties can be organized into four complementary layers: *technical trust*, concerning system integrity and robustness; *cognitive trust*, reflecting transparency and interpretability; *practical trust*, capturing the feasibility and real-world actionability of outputs; and *user-aligned trust*, ensuring alignment with individual goals, constraints, and preferences.

Failure in any one of these dimensions can erode confidence regardless of predictive accuracy. An AI-driven healthcare model, for example, may generate statistically accurate risk assessments yet remain opaque, embed inequities, or lack meaningful recourse tools, each sufficient to undermine trust. Accordingly, trust must be embedded as a core design principle throughout the AI lifecycle, spanning secure deployment, explanation, recourse, and user interaction.

1.1 The Need For A Trust-Centric Framework in AI-Driven Systems

The preceding discussion reveals a fundamental challenge: *How can trust be intentionally designed, enforced, and operationalized across the lifecycle of an AI-driven system, rather than presumed as a byproduct of predictive performance?* Addressing this question requires

reconceptualizing trust not as a singular property, but as a progressive and multi-layered construct. No single technique suffices to establish trust across these dimensions. Instead, trust must be systematically supported through coordinated mechanisms spanning system design, deployment, explanation, and user interaction, as depicted in Fig. 1.1.

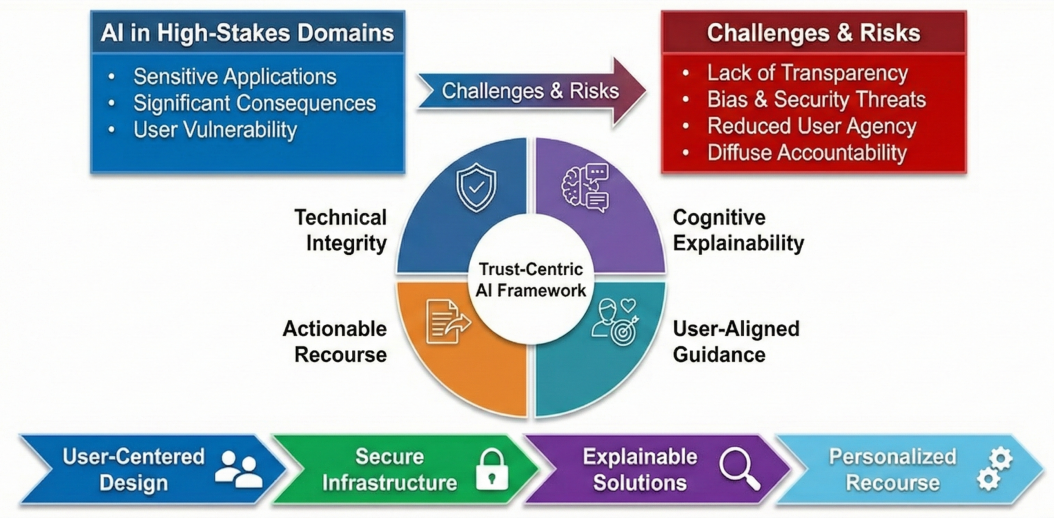


Figure 1.1: Building Trust Through Integrated, Multi-Layered AI Approaches (Image generated using Google Gemini)

1.1.1 AI in Sensitive and User-Facing Domains

This work is grounded in AI applications deployed in sensitive and high-impact domains, where the consequences of error, misunderstanding, or unrealistic recommendations are substantial. In such settings, trust is essential to meaningful engagement. Mental health diagnosis exemplifies this class of domains. Unlike many physical illnesses, mental disorders such as Major Depressive Disorder (MDD) lack definitive biomarkers and are characterized by overlapping, subjective, and context-dependent symptoms [17]. These diagnostic complexities are further compounded by limited access to trained professionals

[18] and persistent social stigma, which often discourages individuals from seeking timely support [19].

Prior research has demonstrated the potential of leveraging Machine Learning (ML) models to identify patterns in behavioral, linguistic, and physiological data relevant to mental health assessment. However, in domains of this nature, accuracy alone does not ensure meaningful engagement. AI systems must also be accessible, dependable, and user-centric to ensure meaningful engagement. This work establishes the broader problem space for a trust-centric framework, i.e., designing AI systems that individuals are willing to rely upon in contexts where vulnerability, uncertainty, and personal consequence are high. Conventional performance-driven evaluation metrics, such as accuracy, fail to capture whether systems are understandable, secure, actionable, and aligned with user needs.

1.1.2 System-Level Trust in Distributed Environments

Having established the importance of user-centered engagement, it is equally critical to ensure that the underlying infrastructure can be trusted. While user-centric design is necessary in sensitive domains, it is insufficient if the underlying infrastructure is vulnerable to compromise. Trust at the user interface layer cannot compensate for weaknesses in the technical foundation producing AI-driven decisions. As AI systems increasingly migrate from centralized cloud infrastructures to decentralized and resource-constrained edge environments, ensuring model integrity has become a critical systems challenge [20].

The distributed nature of edge computing introduces expanded attack surfaces, exposing AI models to threats such as data poisoning, collusion, and inference attacks [21]. These vulnerabilities can silently degrade model behavior, which undermines reliability without immediate detection. Consequently, sustaining trust requires extending beyond conceptual assurances toward verifiable system-level guarantees. Technical trust must therefore be

engineered through mechanisms that protect model integrity, secure communication, and ensure robustness in distributed deployments. This layer establishes the foundational assurance upon which higher-level forms of trust depend.

1.1.3 Cognitive Trust Through Actionable Explainability

Even when system integrity is secured, trust may still erode if users cannot understand or meaningfully respond to AI-generated decisions. Cognitive trust requires not only transparency, but explanations that enable users to reason about system outputs and evaluate their implications. This challenge has driven significant research in explainable AI (XAI), particularly in counterfactual explanations (CFEs), which identify minimal feature changes that would alter a model's prediction [22]. While CFEs provide a roadmap for change, they often lack the causal constraints necessary to ensure that suggested changes are physically possible within the environment.

However, many existing CFE generation frameworks produce suggestions that are unrealistic, infeasible, or disconnected from real-world constraints [23]. Recommendations that users cannot practically implement fail to enhance understanding and instead shift the interpretive burden onto the individual. This form of explanation risks deepening skepticism rather than strengthening trust. Therefore, explainability must move beyond descriptive transparency toward actionable recourse. By embedding structural causal dependencies, feasibility constraints, and real-world cost considerations into counterfactual generation, AI systems can support not only cognitive clarity but also practical usability, thereby reinforcing both cognitive and practical trust.

1.1.4 Personalized and Preference-Aware AI Recourse

Feasibility, however, does not guarantee alignment. Even realistic and implementable counterfactual recommendations may remain misaligned with individual priorities. In real-world decision-making, users evaluate options based on personal values, financial limitations, social context, and competing obligations. Current state-of-the-art approaches approximate such variation using fixed feature-cost assumptions [24, 25] or Bayesian elicitation models [26]. While these methods acknowledge heterogeneity, they often rely on static or population-level representations that inadequately capture individualized preferences.

When counterfactual recommendations fail to reflect user-defined priorities, the cognitive and practical burden of reconciling trade-offs is transferred back to the individual. This misalignment can reduce the adoption of AI guidance and weaken long-term trust. Therefore, achieving user-aligned trust requires explicitly eliciting user preferences and integrating them directly into the counterfactual optimization process. By enforcing preference-aware constraints during recourse generation, AI-driven systems can provide recommendations that are not only feasible but also personally meaningful. Trust is ultimately realized when individuals can act upon AI guidance in ways that reflect their own goals, constraints, and lived realities. This dissertation focuses on tabular and structured decision-making contexts, where decisions are consequential, traceable, and subject to recourse.

1.2 Research Objectives

Together, these challenges motivate a trust-centric perspective that systematically aligns system integrity, cognitive transparency, and personalized actionable recourse within a unified design. Accordingly, the primary objective of this dissertation is to advance a layered trust-centric architectural perspective for AI-driven systems, instantiated through

complementary research objectives that address distinct yet interdependent dimensions of trust. This overarching objective is realized through four interdependent research objectives, each addressing a necessary layer of trustworthy AI.

1.2.1 RO I: User-Facing Motivation and Accessibility

This objective establishes the user-centric foundation upon which the framework is built. Using Major Depressive Disorder (MDD) diagnosis as a motivating case, this work characterizes the unique trust requirements that arise in sensitive, high-impact domains. In such settings, ambiguity of ground truth, limited observability, and barriers to access shape how users perceive, engage with, and rely upon AI-driven support systems.

Rather than proposing a task-specific predictive model, this contribution delineates the design constraints and evaluation gaps that emerge when AI systems operate in contexts where interpretability, accessibility, and user engagement are critical. By exposing the limitations of accuracy-centric evaluation and emphasizing scalability, transparency, and usability, this objective defines the operational problem space within which a trust-centric architecture must function. It serves as the conceptual anchor for all subsequent contributions.

1.2.2 RO II: Secure and Trustworthy Infrastructure

Building upon the user-facing foundation, this objective operationalizes trust at the system level by enforcing the integrity and reliability of AI models deployed in decentralized and resource-constrained environments. As AI systems increasingly adopt AI-as-a-Service and edge-based deployment paradigms, new attack surfaces emerge that threaten the correctness and dependability of model outputs.

This work introduces verifiable mechanisms to ensure that deployed models remain uncompromised throughout their service lifecycle. By leveraging blockchain technology for

integrity verification and fault tolerance, the framework translates trust from a perceived property into an enforceable technical guarantee. This layer establishes the required secure infrastructure necessary for higher-level forms of transparency and user interaction to be meaningful and dependable.

1.2.3 RO III: Actionable and Causal Recourse

Building on secure infrastructure, trust further depends on the system's ability to generate explanations and recommendations that are both understandable and operationally meaningful. Secure execution alone is insufficient if users cannot interpret model behavior or translate outputs into practical action. This objective advances cognitive and practical trust by transforming explanations into actionable guidance grounded in causal reasoning and real-world feasibility.

This contribution reframes counterfactual explanation (CFE) generation as a constrained, multi-objective optimization process that incorporates feasibility, cost-awareness, and structural dependencies among features. By aligning explanation mechanisms with users' practical capabilities and contextual constraints, it ensures that AI recommendations are not merely descriptive but implementable. Explainability is thus elevated from a transparency requirement to a mechanism for enabling informed and achievable user action.

1.2.4 RO IV: Personalized, Preference-Aware Recourse

Even feasible and causally grounded recommendations may fail to achieve real-world impact if they do not align with user preferences, constraints, and values. Trust, therefore, requires personalization mechanisms that adapt recommendations to individual contexts. Accordingly, this final objective completes the trust-centric architecture by embedding individual preferences directly into the generation of actionable recourse.

Table 1.1: Trust-Centric Framework Layers and Technical Implementation

Trust Layer	**Failure Condition**	**Technical Implementation**	**Objective**
User-Facing Risk	Users will not engage without trust	Accessible, Domain-Specific Design	RO I
System Integrity	Trust collapses if models are compromised	Blockchain-based Integrity Verification	RO II
Cognitive and Practical Trust	Secure models still require meaningful explanations	Causal CFEs Multi-Objective Optimization Framework	RO III
User-Aligned Trust	Explanations must align with user preferences	CFE Lexicographic Preference Enforcement	RO IV

This objective introduces explicit preference elicitation and lexicographic enforcement mechanisms that incorporate personal and contextual priorities into CFEs optimization. By reducing cognitive and practical burden and ensuring alignment with individual goals, this contribution enhances the usability of recourse and sustained adoption of AI guidance. Trust is thereby reframed as a negotiated and personalized construct, which is achieved when AI systems generate recommendations that are not only viable but aligned with users' lived realities.

Collectively, RO I–IV operationalize trust across user-facing, technical, cognitive, practical, and personalized dimensions, as summarized in Table 1.1. While developed independently, these contributions demonstrate how security, explainability, personalization, and actionable recourse can be coherently aligned within a layered design. Trust is thus formalized as a multi-layer objective embedded across deployment, explanation, and user interaction, rather than inferred from model accuracy alone.

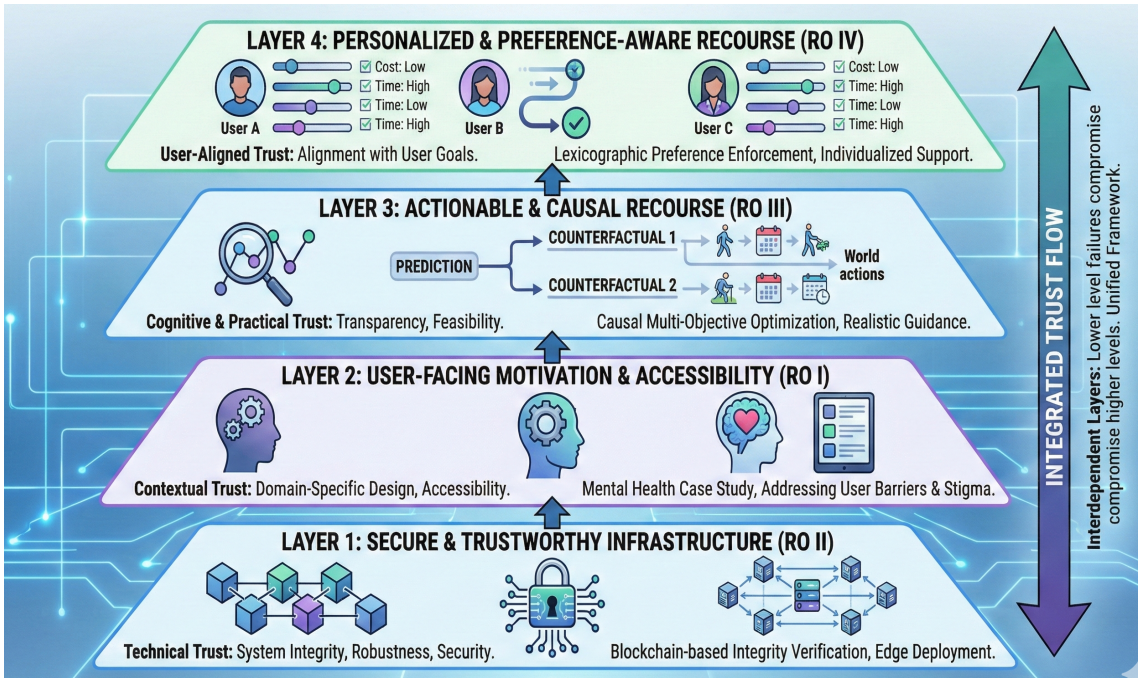


Figure 1.2: A Layered Trust-Centric Framework for Secure, User-Centric AI-driven systems (Image generated using Google Gemini)

Fig. 1.2 illustrates the proposed layered trust-centric architecture, which conceptually aligns the four research objectives within a layered trust-centric architectural model. Each contribution instantiates a distinct but interdependent layer within this architecture, with domain-grounded user motivation (RO I), secure and verifiable infrastructure (RO II), causally grounded actionable recourse (RO III), and preference-aware personalization (RO IV). These layers are interdependent; weaknesses at one level undermine guarantees at others. Compromised model integrity invalidates higher-level assurances of transparency, while inadequate personalization can negate the practical value of secure and interpretable systems. This interdependence reinforces the central thesis of this work: trustworthy AI requires coordinated consideration of technical integrity, cognitive transparency, and user-aligned actionability.

At its foundation, this dissertation establishes secure and integrity-preserving mechanisms for AI deployment. Building upon this assurance, it advances causally grounded CFE frameworks that transform predictions into actionable guidance, and further incorporates preference-aware lexicographic recourse to enable personalized decision support. Together, these contributions articulate a systems-level approach to trustworthy AI within tabular and structured decision-making environments, where outcomes are consequential and subject to recourse. While broader governance dimensions remain important, this work concentrates on aligning system robustness, explanation, and user agency within AI-driven decision support systems.

CHAPTER TWO

LITERATURE REVIEW

Trust is a critical enabler for the adoption and responsible deployment of AI in high-stakes domains, yet it is not guaranteed by accuracy alone. Prior work has shown that users and organizations are exposed to risks stemming from bias, opacity, and technical vulnerabilities, which can undermine confidence in AI-driven decisions [9, 10, 11]. To address these challenges, research has explored mechanisms for establishing technical trust through secure model deployment, verification, and integrity assurance in distributed, edge computing environments [27, 28, 29]. Complementing these efforts, the generation of cognitively and practically trustworthy explanations has been advanced through counterfactual frameworks that ensure plausibility, feasibility, and actionable guidance for users [22, 30, 31]. Finally, approaches to achieve user-aligned trust incorporate personalization and preference-aware methods, including interactive and human-in-the-loop frameworks, to ensure that recourse recommendations are meaningful and tailored to individual users [32, 33, 34].

Prior research has advanced technical security, counterfactual reasoning, actionable guidance, and personalization independently. In this dissertation, each of these dimensions is examined in relation to a specific research objective (RO II–IV), providing the foundation for a broader trust-centric perspective, as depicted in Fig. 2.1. The following sections review the literature relevant to each objective, highlighting methodological developments and gaps that motivate the contributions presented herein.

2.1 RO II: Secure and Trustworthy Infrastructure

Establishing system-level reliability in distributed intelligent environments requires mechanisms that safeguard model integrity, validate computation, and protect data flows

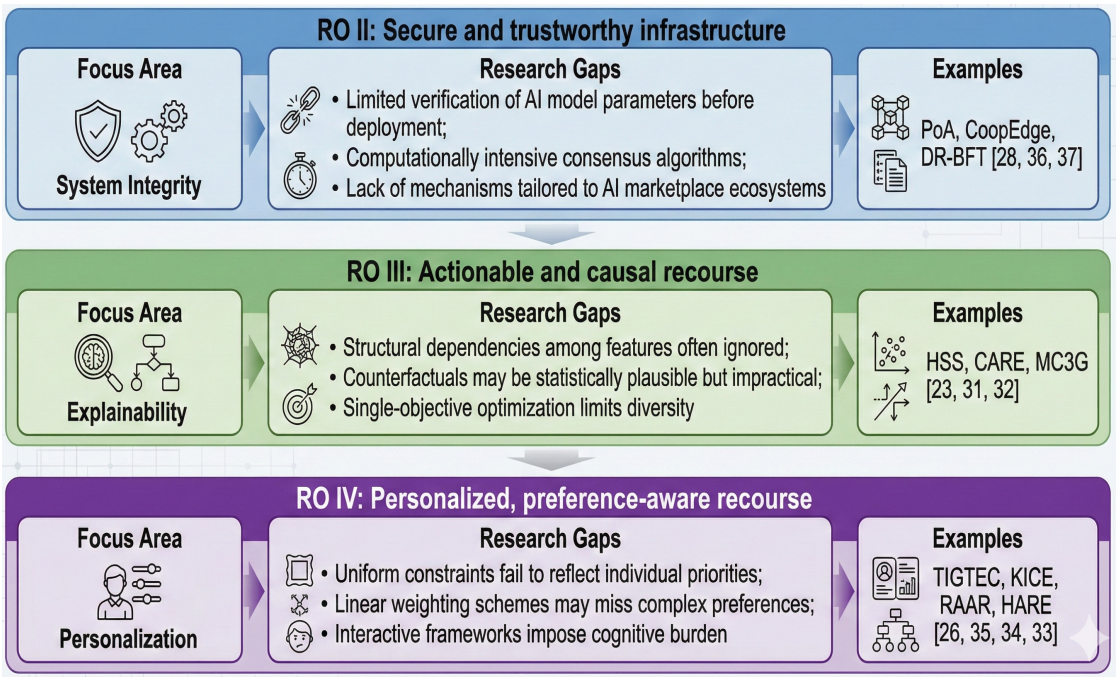


Figure 2.1: Core Limitations of Prior Work by Research Objective (Image generated using Google Gemini)

across decentralized infrastructure. In edge-enabled AI ecosystems, where learning and inference occur across heterogeneous and resource-constrained nodes, vulnerabilities in transmission, coordination, and model updates introduce risks that can compromise decision outcomes. The following works examine architectural and consensus-based approaches designed to strengthen integrity and accountability in such environments, forming the technical foundation for trust in edge-deployed AI systems.

In edge computing, Edge Intelligence (EI) is categorized into four major components: edge caching, edge training, edge inference, and edge offloading [27]. Edge caching refers to the collection and storage of data from the edge devices and the edge server. Edge training occurs at the edge server and/or edge devices. It refers to the application of intelligent algorithms to learn patterns or optimal values from the data cached at the edge. Edge inference refers to the inference made from the trained data after the edge training phase.

This inference serves as an output on the edge devices and servers. Edge offloading involves the distribution of computation processes needed for edge caching, edge training, and edge inference. All of these components are data-dependent.

However, the authors in [28] identified the following data security challenges faced in the edge AI ecosystem. The identified data security challenges are inferring users' private information during data transmission, attacks on the learning agents in the edge servers, and influencing the training dataset with malicious data, which causes the trained AI models to provide erroneous results to users. Therefore, ensuring data integrity in the edge computing architecture is very important, as the presence of compromised data would affect the overall processes present in the architecture [35]. Such aspects of the importance of data integrity for edge intelligence are further elucidated by the authors in [21, 36, 37, 38, 39, 40].

A feasible way to ensure data integrity on the Edge is through the implementation of a consensus mechanism from Blockchain technology. The consensus algorithm mechanism involves the nodes in a Blockchain network reaching a common agreement about a ledger's present state. It helps establish trust among peers and also ensures reliability in the network. The different types of consensus algorithms that exist are Proof of Work (PoW), Proof of Stake (PoS), Delegated Proof of Stake (DPoS), Proof of Authority (PoA), Practical Byzantine Fault Tolerance (PBFT), Raft, Proof of Elapsed Time, and Proof of Capacity [41] [29]. In our proposed method, we have incorporated the Proof of Authority (PoA) algorithm for validating the AI models subscribed to from the AI marketplace. Compared to other existing algorithms, the PoA algorithm has low latency and does not require intensive computational overhead, making it suitable for edge computing.

Additionally, authors in [42] proposed an architecture for data collection, storage, and analysis on the edge layer to produce a shareable outcome among the nodes in the Blockchain network. The authors built the system on four (4) different layers, namely, the sensing

layer, network layer, processing layer, and the sharing platform. The system utilized an honesty-based distributed proof of authority (HDPoA) consensus protocol to validate the data being transmitted in the Blockchain network before it is added to the blocks. The proposed architecture was implemented and tested using metrics like latency, power usage, throughput, and block size. Overall, the architecture was able to provide secure, optimized support needed for AI-enabled applications at the edge. However, the election of mining nodes is subject to their honesty level, which can be easily manipulated by attackers to promote a malicious node.

Cooperative edge computing occurs through peer-offloading, and CoopEdge is a novel blockchain-based decentralized platform that drives and supports cooperative edge computing by establishing trust through incentives among the edge servers [43]. In the CoopEdge framework, there are four roles: task publisher, task candidate, task executor, and task recorder. The task publisher publishes the task and selects the edge server that will act as the task executor. The task candidates are the multiple edge servers that respond to the published task request. The task executor is the selected edge server that will perform the requested task, and the task recorder records the performance of the task executor while executing the task. In the CoopEdge framework, edge servers are rewarded when they are assigned as task executors, completing the task in a given time, and act as task recorders, successfully committing the blocks to the blockchain network. A proposed consensus algorithm, Proof of Edge Reputation (PoEr), was implemented to execute the peer-offloaded tasks. However, the proposed framework only validates the completion time of a given task and the reward to be awarded to the task executor, not the model parameters used by the task executor to execute a peer-offloaded task.

Further, authors in [44] proposed a two-layer Blockchain-based framework to provide data integrity in a dynamic network using an edge computing architecture. The two-layer

framework comprises the end device layer and an edge server layer. The authors also proposed a consensus algorithm called Dynamic Random Byzantine Fault Tolerance (DR-BFT). DR-BFT consists of three sub-algorithms, which are string consensus, data correctness validation, and binary consensus. Depending on the number of nodes present in the network, the DR-BFT consensus algorithm reaches a consensus on a hash value that represents all the messages through the string consensus. The algorithm then invoked the data correctness validation algorithm to verify the data's validity. Once verified, the binary consensus algorithm checked and guaranteed the system's consistency. Using a simulated environment, the authors showed the effectiveness of their proposed framework. While this framework is best suited for dynamic networks, it is time-consuming to validate data integrity.

Overall, existing approaches strengthen integrity and coordination within decentralized edge environments. Yet limitations remain in verifying AI model parameters before deployment and in establishing assurance mechanisms tailored specifically to AI marketplace ecosystems. These shortcomings motivate the need for architectural designs that incorporate model-level validation and lightweight consensus mechanisms fit for resource-constrained edge environments. RO II is positioned within this context, focusing on integrity-preserving mechanisms for AI model deployment in decentralized infrastructures.

While infrastructure-level trust ensures system reliability, user-facing trust depends on transparent and actionable guidance, motivating research on counterfactual explanations and algorithmic recourse.

2.2 RO III: Actionable and Causal Recourse

Explanation mechanisms can enhance user confidence by generating recommendations that are not only outcome-changing but also coherent with real-world constraints. Existing

approaches evaluate alternative instances using multiple quality criteria and implement them through search-based, generative, or optimization-driven strategies, each balancing realism, computational cost, and interpretability in different ways. However, current methods often overlook structural dependencies among features and the practical impact of suggested changes, motivating a causally grounded, multi-criteria formulation that strengthens usability and trust.

Counterfactual Explanations (CFEs) describe minimal changes to an input instance that would alter a model's predicted outcome, while algorithmic recourse refers to the actionable steps an individual can take to achieve that outcome in practice, both requiring causal and feasibility considerations to ensure practical utility. The quality of CFEs is evaluated using several criteria [22, 45]. Validity measures whether a CFE successfully changes the predicted outcome, and diversity captures the provision of multiple alternatives. Proximity ensures the CFE remains close to the original instance, while sparsity limits changes to only a few features. Plausibility reflects the realism of the CFE relative to the data distribution, and actionability indicates whether the changes can be feasibly implemented. Effort quantifies the sequential steps required to reach the desired outcome. These criteria collectively define the objectives for a CFE framework, guiding the design and evaluation of counterfactual generation methods.

Prior work on counterfactual generation can be broadly categorized into Heuristic Search Strategies (HSS), generative models, and optimization-based approaches, each offering distinct tradeoffs [45]. In HSS, counterfactual explainer employs heuristic search methods, such as greedy search [46] and Monte Carlo simulation [47], to prioritize valid, minimal changes that guide the search towards promising candidates. The Monte Carlo Bounds for Reasonable Predictions (MC-BRP) algorithm was proposed in [47] to identify the unusual properties of a particular input data instance and generate a set of feature bounds that will

result in a desired outcome. In Counterfactual Visual Explanation (CVE) [46], CFEs were generated for image classifiers using a greedy search approach to determine the minimum change required for a specific image classification. To generate coherent CFEs, a search algorithm was proposed based on mixed integer programming with a defined mixed polytope acting as constraints [48]. CFEs generated using HSS algorithms are efficient. However, they may produce CFEs that fall outside the data distribution, leading to highly unrealistic CFEs.

The use of generative models in counterfactual generation ensures realistic CFEs that lie within the data distribution. [49] proposed a method, PlausIble Exceptionality-based Constrative Explanations (PIECE), to generate counterfactual or semi-factuals for black-box CNN image classifiers using a Generative Adversarial Network (GAN). Diffusion CFEs framework [50] integrates causal representation with the diffusion generation process to generate causal-aware CFEs. TABCF [51] leverages a transformer-based Variational AutoEncoder (VAE) model to generate CFEs for mixed tabular data. While the application of generative models creates coherent CFEs and works well with high-dimensional data, it is computationally expensive and may struggle with creating diversity due to the latent space limitations or over-regularization.

Counterfactual generation is commonly framed as an optimization problem over selected quality criteria. This allows these properties to be customized to suit domain-specific needs. These properties are encoded as objective functions, which are then optimized using existing algorithms. Distribution-Aware Counterfactual Explanation (DACE) utilizes mixed integer linear optimization to maximize plausibility [52]. The robustness of a black-box predictive model was improved using a meta-heuristic genetic algorithm in Counterfactual Explanations for Robustness, Transparency, Fairness, and Interpretability of Artificial Intelligence models (CERTIFAI) [53]. Wachter [23] implemented a gradient-based optimization algorithm

to minimize proximity and maximize plausibility. While formulating CFE generation as a single-objective optimization is computationally efficient, it often limits diversity and may overemphasize a single property due to weighting. To address these limitations, counterfactual generations are formulated as multi-objective optimization (MOO) problems.

When CFE generation is formulated as MOO, it improves counterfactual diversity and interpretability without weighting biases, but is computationally costly. Coherent Actionable Recourse (CARE), a multimodal counterfactual explainer framework, implemented the NSGA-III multi-objective optimization algorithm to produce feasible CFEs by optimizing sparsity, proximity, connectedness, coherency, actionability, and diversity [30]. While CARE generates a feasible temporal action sequence, it assumes feature independence, leading to a less efficient recourse path and causally infeasible counterfactuals. In [54], integer programming was implemented to generate the recourse cost for the generated CFEs. However, it is restricted to linear models. This hinders its applicability, as many AI systems rely on complex, non-linear models. [55] proposed a counterfactual generator to optimize validity, proximity, sparsity, and plausibility, using the NSGA-II algorithm, providing a diverse set of solutions. However, the causal dependencies between the features were not accounted for, generating counterfactuals that were plausible statistically, but not realistic. A balanced formulation that jointly considers validity, proximity, plausibility, and effort can better support realistic and practically achievable recourse. Additionally, capturing the extent to which suggested changes effectively guide users toward the desired outcome remains underexplored, highlighting the need for efficiency-aware objectives.

In terms of incorporating causality in CFE generation, ImageCFGen incorporated a causal graph in a variant of Adversarial Learned Inference (ALI) to generate realistic images as CFEs [56]. However, latent space generation can obscure precise feature interventions, limiting user interpretability and actionable implementation. The model-agnostic Causally

Constrained Counterfactual Generation (MC3G) framework integrates causal reasoning as a rule-based constraint using the FOLD-SE algorithm in the Answer Set Programming system [31]. However, it is limited to rule-based models. Recent approaches have explored integrating causal graphs with structural equation modeling (SEM) to enforce constraints in generated CFEs, improving alignment with real-world dependencies. In summary, despite advances in multi-objective optimization and generative modeling, many existing approaches do not explicitly enforce structural consistency when producing alternative outcomes. As a result, recommended changes may appear plausible statistically but remain impractical under real-world dependencies. This underscores the need for causally informed counterfactual generation mechanisms.

2.3 RO IV: Personalized, Preference-Aware Recourse

Recent work on algorithmic recourse has expanded beyond validity-focused CFE generation to incorporate user-specific feasibility, preferences, and contextual relevance. These approaches range from fixed feature costs and semantic heuristics to interactive frameworks and human-in-the-loop systems that adapt recommendations according to individual priorities. Despite these advances, challenges remain in balancing personalization, realism, and interpretability while respecting constraints that cannot be traded off, motivating methods that integrate user preferences into principled, structured optimization.

Prior work formalized CFE generation as an optimization problem, balancing validity, proximity, and sparsity to produce interpretable explanations [23, 53, 55]. While these approaches ensure valid predictions, they typically assume uniform cost structures and do not consider whether suggested changes are feasible or meaningful to users. Subsequent frameworks [57, 58, 59, 30] sought to bridge the gap between validity and actionability by enforcing feasibility and plausibility constraints, such as imposing domain-specific

rules, feature immutability, or causal relationships to prevent unrealistic recommendations. However, these constraints are generally applied uniformly across users, failing to capture individual priorities or preferences.

To address this limitation, recent research incorporates feature or user-specific priorities, either through predefined cost models or explicit preference elicitation. Fixed feature-specific costs are used to approximate user preferences in [24, 60], while Token Importance Guided TExt Counterfactuals (TIGTEC) integrates semantic cost heuristics and local feature importance to generate sparse, plausible counterfactuals [25]. Although these cost models and heuristics improve personalization beyond uniform constraints, they rely on externally specified costs that may not reflect true user priorities. In [61], user preferences are captured via soft constraints, continuous feature scoring, bounded ranges, and ranked categorical features, guiding gradient-based optimization to produce constrained, user-preferred recourse. A limitation of this framework is that the reliance on a linear weighting scheme may fail to capture complex or context-dependent preferences.

Beyond static preference modeling, a parallel line of work has explored interactive and human-in-the-loop approaches for incorporating user preferences. Guided interactive frameworks [62, 63] iteratively elicit user preferences, demonstrating improved efficiency and satisfaction over unguided exploration. However, their dependence on predefined preference dimensions may limit user autonomy. Similarly, Knowledge Integration in Counterfactual Explanation (KICE) integrates explicit user knowledge and domain expertise to tailor CFEs [34], allowing users to specify which features they understand or are willing to change. Its limitation lies in capturing knowledge as unranked constraints, offering no mechanism to prioritize features or resolve conflicting or uncertain input. Relevance-Aware Algorithmic Recourse (RAAR) introduces a relevance model, often via Bayesian optimization, to bias counterfactual generation toward personally relevant

features [33]. While this improves relevance, it increases computational cost and reduces explainability, limiting generalization. Finally, human-in-the-loop frameworks [26, 32] iteratively collect user feedback, often through pairwise comparisons or rankings of candidate explanations, to adapt the recourse to individual preferences. This produces highly personalized and actionable CFEs, but it imposes cognitive and time burdens, as multiple interaction rounds may be needed to reach satisfactory outcomes.

Collectively, these approaches reflect a progression from validity-focused uniform recourse generation towards frameworks that integrate feasibility, plausibility, and user's preferences. However, they also highlight the challenge of balancing personalization, explainability, user effort, and computational complexity. Although prior work [30, 32] utilizes weighted or iterative objective formulations, these frameworks fail to capture the non-compensatory nature of certain criteria, whereby violations of plausibility, feasibility, or user preferences cannot be offset by the gains in secondary objectives. This necessitates structured principled trade-offs that align optimization with individual constraints. To address these limitations, structured, tiered approaches have been explored that prioritize critical constraints, such as feasibility and user preferences, before optimizing secondary objectives like user effort. These hierarchical formulations respect non-compensatory criteria while integrating personalization and causal considerations, establishing a principled foundation for generating actionable, user-aligned recourse. This perspective underscores the potential for frameworks that reconcile individual priorities with realistic and interpretable counterfactuals, motivating methods that operationalize these trade-offs.

CHAPTER THREE

RO I: USER-FACING MOTIVATION AND ACCESSIBILITY

A human being's good mental health is vital to ensure emotional, psychological, and social well-being, and any condition that affects this well-being is known as a mental disorder. One such disorder is Major Depressive Disorder (MDD), or simply depression. According to the World Health Organization (WHO)[1], symptoms of depression are characterized by an individual's lack of interest in previously enjoyable activities and persistent sadness. Depression is a leading cause of disability worldwide, with almost 75% of humans suffering from MDD in developing countries who remain untreated, and approximately 1 million susceptible cases lead to suicide [64]. In the United States (U.S.) alone, MDD affects more than 16.1 million American adults each year, which constitutes about 6.7% of the U.S. population aged 18 and older [64]. Additionally, due to the global pandemic from SARS-CoV-2 (COVID-19) in the year 2020 itself, depression diagnoses were up by 873% [65]. In this regard, concerning treatment, only 36.9% of those suffering from anxiety disorder were receiving treatment [65].

The demand for mental health services is rapidly increasing, as shown in a 2018 survey by the National Council on Behavioral Health (NCHB)[2]. The survey concluded that at least 56% of people want access to mental health services but face many barriers. These barriers can be attributed to a lack of resources and/or trained healthcare providers, the social stigma associated with mental disorders, and inaccurate assessment[3]. Another survey in 2018 [18] showed that there is a shortage of mental health care professionals in every state across the U.S. This situation is further aggravated due to the high cost and insufficient insurance

[1] https://www.who.int/health-topics/depression

[2] `https://shorturl.at/s5biT`

[3] `https://rb.gy/zgznyj`

coverage for mental health conditions, leading to the difficulty of accessing mental health services by the economically challenged population. Nonetheless, most people suffering from mental illnesses have a prevalent behavior of wanting to be alone. Due to this isolation, mental disorder patients seek online venues like Twitter, Facebook, or Reddit to openly or anonymously share about discomforts and anxieties. This gives rise to data repositories comprising a variety of user-curated content on social media platforms, such as personal status, users' pictures, and geo-location, which can be mined for information. While humans have limited capacity to learn, an AI-actuated system can easily access thousands of medical information sources and help in the early detection of chronic mental health diseases in patients.

Background: The need for AI. Traditional means of providing screening (diagnosis) and monitoring, although good, are not sufficient today due to two different reasons. First is the Big Data Connection. The information generated from pervasive devices with Internet connectivity and social media platforms can aid in detecting and monitoring depressive behavior. However, analyzing Big Data generated from online platforms is not supported by traditional care and therapy. Second is the ease of accessible care and constant monitoring of MDD patients. Traditional means cannot address the accessibility issue and provide constant monitoring, in contrast to AI-enabled frameworks that address mental well-being. So it is necessary to support and extend the methods of traditional care by using AI-based frameworks to leverage the advancement in computer technology for social good. A diverse technical solution has been adopted and implemented to help solve the limited access to medical professionals. The solutions range from chatbots, IoT/Wearable devices, mobile and web applications to behavioral technologies [66, 67, 68, 69, 70, 71]. These solutions have helped to significantly increase access to the professional help needed for individuals

suffering from MDD. The nature of these solutions makes them easily accessible, thereby reducing the stigma that is associated with MDD.

Objectives and Contributions: Given the importance of AI-enabled frameworks towards diagnosing and monitoring MDD, this work surveys and discusses the different AI-enabled approaches to MDD that have been proposed in the recent literature. The contributions of this work are as follows:

- Review existing literature on AI-enabled frameworks for diagnosing mental health illnesses and summarize existing research challenges and some future directions.
- Propose a decentralized system design for an accessible, scalable AI-enabled approach for diagnosing mental health illnesses.

3.1 AI-Enabled Approaches

The **systematic search strategy**, along with some exclusion criteria were as follows. The paper identified related works based on their solutions that addressed features like reliability, accuracy, accessibility, and explainability. The works before 2015 were excluded, along with any work that had no testing or implementation, or if the data was not gathered from a valid source, like health agencies or social media. For recent literature surveyed belonging to each domain, the paper presents a discussion on their algorithmic description and the input and output types.

3.1.1 Expert System and Fuzzy Logic

In [72], the authors proposed an expert system that can be used to diagnose depression in an individual. Due to the nature of depression diagnosis, the expert system would help the psychologist to appropriately diagnose an individual. The expert system was created using Simpler Level 5 (SL5) object language, with its engine implemented in Delphi Embarcadero

RAD Studio XE6. The proposed expert system interacted with the human subjects by asking them a set of Boolean questions. Based on the answers chosen, the expert system outlined a diagnosis and recommendation to the user. The knowledge base for the expert system was created with the information documented from experienced psychologists and specialized websites for depression. The proposed expert system was thereafter evaluated by an experienced psychologist, who verified the correctness of the system output. The benefits of the expert system can be categorized by its ease of use. Further, the system can be deployed as a standalone system, without the need for an intervention from a medical professional.

The authors in [69] designed a web-based expert system using fuzzy logic to diagnose individuals suffering from depression and determine the level of severity. The system was designed to be user-centric; it could be implemented in specialized settings, such as in a psychiatric office, as well as being used by the user solely. Fuzzy logic was used to address the uncertainty or ambiguity present in human knowledge and the decision-making process. Knowledge was acquired from several resources and professionals, with the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) being the main source. The scale for the severity of depression disorder used was "very low, low, medium, high, and very high". The expert system was implemented in Jess. The system's predictive results were evaluated by psychological consultants who verified the accuracy of the results. As reported by the authors, their proposed system also achieved a 79% and 98% outcome for the metrics of sensitivity and specificity, respectively. Using Fuzzy logic gives allowance for uncertainty in decision-making, which makes it flexible and easily adaptable to different scenarios. The level of uncertainty or vagueness present in the decision-making process makes the precision and accuracy of the results obtained questionable.

3.1.2 Machine Learning Approaches

The current trends in digitization have penetrated many industries, including healthcare. The concurrent growth in medical demands and improved efficacy of machine/deep learning algorithms can help medical institutions and their professionals to detect early signs of chronic diseases, which helps to reduce the time and lower the costs of treatment. This section summarizes the various Machine Learning (ML) based frameworks that have been proposed. The literature within this category is further analyzed based on sub-categories like the type of ML algorithms ((un)supervised, semi-supervised, and reinforcement learning), along with the proposed model's respective algorithmic input and output. Although a net total of about 15 works have been reviewed in this category (section 3.3), owing to page limitations, in this section, the description of only a select few notable papers has been presented. For a complete description of the reviews, readers are kindly referred to the following document - https://rb.gy/ajeq5q.

3.1.2.1 Supervised Learning:

Machine learning approaches that belong to this algorithmic category utilize historical and classified input and output to train themselves and facilitate the intended functionalities.

Summary: To begin with, one such literature that belongs to this algorithmic category of supervised learning is [73]. It discusses the prevalence of Major Depressive Disorder (MDD) and Generalized Anxiety Disorder (GAD) in youth attending universities. The authors state that if such issues are not diagnosed at an early stage, then it leads to drinking-related harm and alcohol abuse. The paper [73] proposes a machine learning model that predicts if an individual suffers from MDD and GAD by creating a so-called Electronic Health Records (EHR) data set. The EHR comprised an undergraduate student's biometric and demographic data, with the exclusion of all psychiatric features to preserve privacy. An

ensemble algorithm consisting of Support Vector Machine (SVM), XG Boost, K-Nearest Neighbor (KNN), Random Forest, Logistic Regression, and Neural Network was deployed using Bayesian hyperparameter optimization.

Algorithmic Input: The input to the machine learning model consisted of an individual's physiological data, like blood pressure, Body Mass Index (BMI), heart rate, and demographic data like housing status, health insurance, and age. The authors used non-psychological factors such as age, housing zip code, and health insurance from the Electronic Health Record (EHR) data set to predict if a college student is suffering from Major Depressive Disorder (MDD) or General Anxiety Disorder (GAD), and explain the model's results.

Algorithmic Output: As per the author's reporting, the ensemble model achieved 70% accuracy, and some of the top features were satisfaction with living conditions, public health insurance, and parental home. An XG Boost classifier was used to predict an individual's status if they are suffering from depression or not. Overall, the Area Under Curve(AUC) scores obtained from each model in the ensemble model indicated that the model accurately predicted an individual's state, and also Shapely Additive Explanations (SHAP) was used to explain how the importance of each feature affects the overall result of a model's performance. The obtained results allowed the authors to validate that depression can be diagnosed using non-psychiatric features.

Summary: In [74], the authors have proposed the incorporation of digital intervention in predicting depression and other forms of anxiety symptoms. The authors aimed to evaluate the improvement experienced by a patient who received care via digital intervention after hospital discharge from illness around work-related stress. This research was conducted in a randomized controlled trial of 632 people who received the digital intervention. This group of users was split into two groups - a group using the digital interventions for their

follow-up and the other group receiving information from a professional. This study was done for 9 months after the hospital's discharge.

Algorithmic Input: The authors implemented an ensemble model which had 2 layers to analyze the data. The first layer, known as the base model consisted of ridge regression, random forests, general linear models, Gaussian process, support vector machines, K-nearest neighbors. The second layer was the averaging layer, which took as input the mean prediction for a given subject across the different models and validation folds.

Algorithmic Output: Based on the output of the results, the authors stated that their ML model predicted a change in the depression of a person. The authors also stated that the ML model could be capable of guiding a clinician to know the best way to allocate resources in caring for a user, either using a higher level of traditional care or a low-level resource consisting of digital intervention.

Summary: The authors in [70] aimed to evaluate chatbots utilization by users. Tess is the chatbot that was used as a case study in this paper. Tess deciphers a user's emotional needs through the content of their conversation. Tess has 12 modules, and its questionnaires or content spans across different areas of interest such as self-compassion, cognitive distortion, coping statements, and so on. These areas make Tess a robust chatbot, but also present a herculean task for the users to go through the modules in good time.

Algorithmic Input: Tess was deployed on Facebook Messenger with anonymous data collection. The authors aimed to analyze the usage of Tess, understand users' flows between the various modules, and the usage of each module.

Algorithmic Output: From the analysis carried out, the authors showed that a chatbot is an effective tool for mental health, just the major modules required to assist a user should be implemented in a chatbot, and the overall usage across the different modules was based on some factors such as the questions asked and the time required to complete a module.

Summary: In this work [71], the authors developed a depression diagnosis algorithm using an individual's sequence of responses to a virtual agent and determined their depression severity.

Algorithmic Input: This work involved diagnosing depression through the sequencing of responses obtained from a user using audio and text features. The responses were grouped as responses to specific questions asked from the Patient Health Questionnaire (PHQ) and responses that were not dependent on the questions asked. This experiment was carried out using three different scenarios to evaluate the effectiveness of their model. In the first scenario, called the context-free modeling, a regularized logistic regression was deployed, and the time a question was asked was the key factor and not the question type. In the second scenario, called weighted modeling, a regularized logistic regression was deployed, and the question type, not the timing of the question, was the key factor. In the third scenario, called sequence modeling, a bidirectional Long Short-Term Memory (LSTM) was implemented.

Algorithmic Output: The authors discovered that in the cases of context-free modeling and sequence modeling, the text features gave a better result. Whereas, in the case of weighted modeling, the audio features performed better. Overall, the best result was obtained from the weighted modeling for both the text and audio data.

Summary: It discusses the prevalence of Major Depressive Disorder (MDD) and Generalized Anxiety Disorder (GAD) in youth attending universities. The authors state that if such issues are not diagnosed at an early stage, then it leads to drinking-related harms and alcohol abuse. The paper [75] proposes a machine learning model that predicts if an individual suffers from MDD and GAD by creating a so-called Electronic Health Records (EHR) dataset. The EHR comprised an undergraduate student's biometric and demographic data, with the exclusion of all psychiatric features to preserve privacy. An ensemble algorithm consisting of Support Vector Machine (SVM), XG Boost, K-Nearest

Neighbor (KNN), Random Forest, Logistic Regression, and Neural Network was deployed using Bayesian hyperparameter optimization.

Algorithmic Input: The input to the machine learning model consisted of an individual's physiological data, like blood pressure, Body Mass Index (BMI), heart rate, and demographic data like housing status, health insurance, and age. The authors used non-psychological factors such as age, housing zip code, and health insurance from the Electronic Health Record (EHR) dataset to predict if a college student is suffering from Major Depressive Disorder (MDD) or General Anxiety Disorder (GAD), and explain the model's results.

Algorithmic Output: As per the author's reporting, the ensemble model achieved 70% accuracy, and some of the top features were satisfaction with living conditions, public health insurance, and parental home. An XG Boost classifier was used to predict an individual's status if they are suffering from depression or not. Overall, the Area Under Curve(AUC) scores obtained from each model in the ensemble model indicated that the model accurately predicted an individual's state, and also Shapley Additive Explanations (SHAP) was used to explain how the importance of each feature affects the overall result of a model's performance. The obtained results allowed the authors to validate that depression can be diagnosed using non-psychiatric features.

Summary: The authors in [76] aimed to diagnose depression in a demographically diverse population using behavioral biomarkers in audiovisual data. The authors designed a web-based evaluation called Artificial Intelligence Mental Evaluation (AiME) to identify depression in audiovisual data.

Algorithmic Input: An experiment was carried out where participants completed the evaluation using a webcam and a microphone; however, the demographic data of these participants were anonymous. Data validation was carried out to ensure the baseline data quality was met; those who did not meet the quality were discarded. A multimodal deep

learning model was used to analyze the participants' data. The prediction of the ML model involved performing a binary classification on whether the participants were clinically depressed or not, using a cutoff value from the PHQ-9 scores, and using regression to determine if a participant is treating depression as an ongoing problem based on predicting their PHQ-9 scores between 0 and 27 against the true PHQ-9 values. Data augmentation was added to reduce the potential of overfitting the model and increase the input data points.

Algorithmic Output: The mean absolute error (MAE) and root mean square error (RMSE) were used to evaluate the model. The authors discovered that when the regression models were trained on the male and female participants, the model's performance was remarkably better. The following are the performance metrics of the balanced model: Accuracy: 68.02, 95%; Precision: 68.61%; Negative Positive Value (NPV): 67.61, 95%; Sensitivity: 68.59, 95%; Specificity: 67.46, 95%; F-1 Score: 67.66, 95%.

3.1.2.2 Unsupervised Learning: is a type of machine learning technique where the label is not pre-defined but determined by the clustering of a set of data points. The labels are defined by the patterns discovered in the data set. In this sub-section, the paper presents a discussion on existing works that fall in the category of unsupervised learning to address Major Depressive Disorder.

Summary: In [77], the authors implemented an unsupervised machine learning algorithm using K-Means Clustering to predict if an individual is suffering from depression, along with predicting their depression severity. K-Means clustering uses the position of each data point instead of the summary of data points, thereby making its prediction output more holistic in nature. Moreover, the model's results were compared with the results from a traditional norm-based classification to evaluate the model's performance. Middle and High school Chinese students were the focus of this study.

Algorithmic Input: The K-Means clustering was implemented in a 13-dimensional space with 13 features obtained from the Beck Depression Inventory (BDI) questionnaire. The cluster centers were determined using the maximum and minimum points. The authors classified the severity of depression as none, mild, moderate, and severe.

Algorithmic Output: The model was able to correctly predict if an individual was suffering from depression, and when compared to the traditional norm-based classification models, the model had higher accuracy and AUC score.

3.1.2.3 Semi-supervised Learning: is a combination of supervised and unsupervised learning. In semi-supervised learning, a small amount of labeled data is used along with a large amount of unlabeled data, which can be helpful when there is an unavailability of a large amount of labeled data.

The research work presented in [78] is an example of semi-supervised learning wherein the research objective was to monitor the clinical depressive symptoms from Twitter data that imitates the PHQ-9 survey used by clinicians. The authors proposed two approaches to detect the possibility of users suffering from depression. The former was a bottom-up approach where authors used distributional semantics to unwrap the symptoms of depression. The first approach is based on Latent Dirichlet Allocation (LDA), which is specifically an unsupervised learning that views tweets as a mixture of latent topics, where a topic is a distribution of co-occurring words. However, the topics learned by LDA are not specific enough to correspond to depressive symptoms. The latter approach was based on a top-down approach where the authors added supervision to LDA by using a probabilistic topic modeling approach, named semi-supervised topic modeling over time(ssToT).

Algorithmic Input: The authors collected data of 45000 Twitter users with self-declared depression and another 2000 tweets of undeclared users. After the examination, it was

seen that most users talked about their family and companion issues and the need for their help. The authors compared the ssToT learned topics with existing semi-supervised and unsupervised learning approaches such as k-means clustering, LSA, LDA, BTM, and Partially Labeled LDA. These experiments suggested that ssToT outperformed all the state-of-the-art models, paying little heed to the corpus from which probabilities are acquired. The ssToT model was also tested as a multi-label classifier, using a 10400 tweet dataset in 192 buckets. Each bucket contains tweets that are posted by the client within a range of 14 days. The ssToT model was trained on an unlabeled data set, and the performance was estimated by utilizing the labeled data set.

Algorithmic Output: Accuracy and precision were measured for the ability to predict the presence of 9 depressive symptoms (in compliance with PHQ9), and they were found to be 0.68 and 0.72 on average. The obtained results suggest that semi-supervised topic modeling over time was successfully able to capture depression symptoms from the Twitter data-set which was competitive with a fully supervised approach.

3.1.2.4 Reinforcement Learning (RL): enables agents to learn by their interaction with the environment by trial and error, using feedback from past actions and experiences. RL is based on rewards and punishments for the agent's set of actions to perform a task.

Summary: In [79], the authors aim to understand the relations between model-derived reinforcement learning parameters and the depression symptoms and symptom change after the treatment. Studies were conducted on 101 adults, of whom 68.3% were females. The authors also assessed the changes in model-learning parameters and symptoms after some of the participants received Cognitive Behavioral Therapy (CBT).

Algorithmic Input: The participants responded to the Mood and Anxiety Symptom Questionnaire (MASQ), a validated self-report measure of the symptoms of anhedonia,

negative affect, and arousal, as well as the Beck Depression Inventory (BDI) to assess overall depression severity, the Wechsler Test of Adult Reading to estimate verbal IQ, and a demographic questionnaire. Out of 101 participants in the study, a total of 69 participants suffered from MDD, and 32 participants had no history of MDD. The computational model analysis of behavior choices and neural data identified contemporary learning with symptoms during reward and loss learning.

Algorithmic Output: The results showed that during reward learning, the reward values increased slowly with increased anhedonia, as it was associated with derived model units such as learning rate, outcome sensitivity, and neural signals. The results during the loss function manifested that the learning parameter associated with negative affect was found to be outcome shift, and the model-learning parameter associated with disrupted neural encoding of learning signals. After mapping the reinforcement learning model parameters with the symptoms of MDD, it was observed that after CBT, the features associated with the symptoms had shown possible learning-based therapeutic processes.

3.2 Explainable AI (XAI)

The recent advances in the utilization of AI for the medical field have been restricted due to the absence of interpretation for complicated models. This decreases the trust of clients when it comes to the output generated by AI-enabled models. The AI models that explain their outcomes are better known as explainable AI (XAI) and can be classified into two major categories [80]. First, as Ante-hoc, where the framework incorporates logic in the model. Second, as Post-hoc, the frameworks utilize a simpler model to clarify a black-box model. In the following section, the paper discusses some of the notable literature from the mental health domain that proposes the usage of XAI.

Summary: In [81], authors propose to detect the early signs of depression from users' social media posts. They combined XAI with natural language processing (NLP). The XAI model used Local Interpretable Model-Agnostic Explanations (LIME) [82] for interpretation. LIME is a local model interpretation method utilizing local surrogate models to surmise forecasts of black-box models. The authors observed that LIME interpretation was sensitive to pre-processing of the data set, especially the choice of whether to remove stop-words from the data set. The authors removed stop-words from the data set to study the behavior of occurrence of personal pronouns in the depression data set.

Algorithm Input: consisted of Urdu and English text data from the social media platform Reddit and applied NLP algorithms to predict if a user is suffering from depression. The author used bag-of-Words (BoW) and term-frequency times inverse document-frequency (TF-IDF) features and used them in machine learning classifiers like Logistic Regression and Random Forest classifiers.

Algorithm Output: for the author's experimental setup, the Logistic regression classifiers performed better with an F1 score of 0.89 for TF-IDF features as compared to the Random Forest classifier (F1 score of 0.84).

Summary: In [83], the framework proposed by the authors incorporated the SHapely Additive exPlanations (SHAP) [84] for the purpose of model interpretation. SHAP uses values that are an average of the marginal contributions of a single feature value across all possible partnerships of the features. It helps to represent the impact certain features have on the model outcome with which these shapely values are associated. The authors in [83] used these shapely values to assess and foster a novel AI approach for predicting mental health risk in individuals with diabetes mellitus.

Algorithm Input: Data was collected from 142,432 people suffering from diabetes, and their mental health status was verified using two sources - claims data and medication

prescription data. The participant's behavior was classified into four categories, including demographics and glucometer, coaching, and event data. Data sets were then collected to make member period occurrences, and descriptive analyses were performed to comprehend the connection between psychological well-being status and passive sensing signals. The model used for training the data set was an ensemble of ten LightGBM models, and it was evaluated using sensitivity, specificity, precision, the area under the curve, F1 score, accuracy, and confusion matrix.

Algorithm Output: The outcome as reported by the authors showed that the proposed model performed with a score of more than 0.5 for sensitivity, specificity, AUC, and accuracy. The SHAP values determined the elements that offer more towards the classification output, such as demographics, participants' emotional state during blood glucose checks, time of day of blood glucose checks, and blood glucose values. The authors were able to effectively anticipate the mental health risk in users experiencing diabetes and show the significance of different elements that contributed most towards the classification output of the model.

Summary: The authors in [85] introduced a framework (What-If tool or WIT) that can visually analyze the ML systems. The WIT tool supports both local and global interpretation; that is, it can analyze a local data point as well as analyze model behavior across the whole data set. This framework allows users to find the nearest counterfactual to better understand the model's behavior. The tool also enables users to analyze the relationship between the features and the predicted output using partial dependence plots, which helps to better understand what features contribute more to the prediction outcome.

Algorithm Input: The authors gave three contextual analyses to show the utilization of WIT to analyze the model performance. The first study, directed by ML specialists, utilized the WIT to analyze regression models. A sample of 2000 data was stacked in the regression model. In the second study, a programmer of a large technology company used WIT to

predict the health metric for clinical patients. WIT, coupled with two regression models, showed that the inference results and the data-point visualization results were bigger for the first regression model than the second. More interestingly, it was observed that there was a bug in the calculation that caused the first and the last value of the input feature to trade. The error matrix was sensible for the model, but the model was tackling an alternate issue because of this bug. Finally, the third study was led by computer science students from M.I.T examining the legitimacy of the stop-and-frisk practices of the Boston police department. They used a data set of 150,000 records and prepared a linear classifier model. The dataset had features like age, race, sex, and criminal record of an individual being stopped by police, alongside the date and location of the offender.

Algorithm Output: The first study observed that the partial dependence plot was flat for some features of every data point. In the second case study, the WIT tool found a bug in the software that prevented a certain feature from being fed to the model, which resulted in a wrong decision. The output of the classifier in the third study was either positive or negative, i.e., whether the offender was searched or not searched during the confrontation. This outcome was coupled with WIT that identified that features like age, gender, and race played a key role in determining whether the offender was frisked or not.

3.3 Discussions

Fig. 3.1 and 3.2 summarize the literature classified using different categories like machine learning algorithm types, input data type, and so forth. The numerical values in these illustrations represent the number of reviewed papers that belong to a respective category. Further discussions presented in this section illustrate some of the design challenges identified in this field, the scope of future directions, and the tentative design of an AI

system to address mental health illnesses, keeping in mind computational design parameters like accessibility, efficiency, and effectiveness.

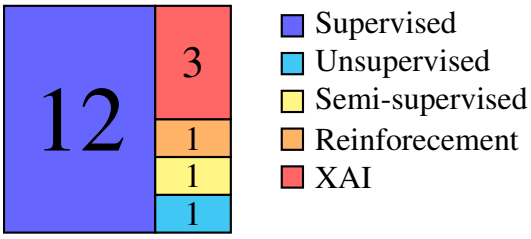


Figure 3.1: Literature Classification: Machine Learning Algorithm Types

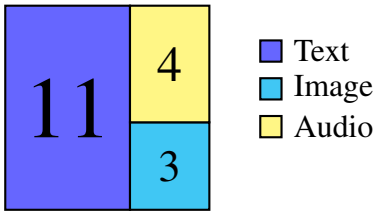


Figure 3.2: Literature Classification - Algorithm Input Types

3.3.1 Lessons Learned: Current Challenges and Future Directions

An important finding of this survey is that the majority of the AI-enabled approaches belong to the category of Machine or Deep Learning frameworks (ML). This is owing to the dynamic nature of the information associated with depression and the growing number of patients, which makes fuzzy and expert systems an ill-fit in terms of computational design criteria. The process of developing expert systems or fuzzy logic-based systems requires extensive information gathering, which can inherently be tedious. Thereafter, these systems require a knowledge base to be developed from the gathered information. These knowledge bases can quickly become outdated and, in addition, are challenged with issues of scalability and accuracy. Therefore, the majority of the existing works use ML frameworks since an ML model can be built using a lot of data, making it robust and can be retrained with new data, making it accurate, scalable, and up-to-date. A model can be easily trained and deployed quickly, as the professional input needed can be obtained from a variety of sources in a short time.

The second finding from the survey is that, within the domain of ML frameworks, the majority of the existing works incorporate the supervised learning methodology compared to unsupervised, semi-supervised, or reinforcement learning methods. The reason behind the popularity of supervised learning can be attributed to a few cases. First, in the mental health research domain, the majority of the AI-enabled frameworks consider the problem (e.g., MDD prediction) as a binary classification - depressed or not depressed. This makes labeling easier when compared to, say, unsupervised learning, which tries to infer its clusters based on the data set, which can be challenging at times. Hence, there is a preference for supervised learning frameworks, where the classes are already defined, and the models know what they are looking for. Secondly, in these domains, the availability of a reliable, accurate, and substantial amount of data is a bottleneck. This bottleneck proportionately impacts the performance of any ML-based AI framework. Supervised learning, when compared to unsupervised learning, does not require too much data. Third, given the current state of the art of AI systems, in the medical field, it is still favorable to have a human expert intervention. Supervised learning allows this process at an early stage, i.e., when the models are being trained, and thereby allows for more accurate models when compared to unsupervised or semi-supervised learning methods.

Additionally, the reinforcement learning framework is not best suited for this type of problem, i.e., depression diagnosis. In reinforcement learning, the agent goes through various states, and depending on the outcome, they either get rewarded or punished. Due to the dynamic nature of how depression can be diagnosed in an individual, there might be a chance of the agent not prioritizing a particular state or discounting a state because it does not seem like a 'norm' when it can be useful in diagnosing depression in the individual. Also, there is a question of how long the agent runs before it arrives at a decision and can prove the result obtained is valid. Furthermore, the functional benefits of supervised learning

algorithms over their counterparts currently make it a more preferable choice. However, looking ahead, researchers in this domain need to assess an important question: if there is a paradigm shift towards unsupervised (or semi-supervised) learning algorithms, what kind of benefits and challenges will be encountered? To begin with, the benefits of such a paradigm shift will be in the ability to identify patterns that exist in diagnosing depressive disorders that are currently unknown, even among medical professionals. With this in mind, the outcome of the AI models no longer needs to be a simple yes or no classification. The ranking of depression severity would become more flexible as these algorithms will be able to analyze patterns, relationships, and causality that exist in the data points. This will help to identify features that are easily susceptible to a depression level in an individual.

However, the primary challenge will be in terms of data gathering. The unsupervised and semi-supervised models require a lot of data. Thereafter, functionally designing these frameworks will also be challenging. An imperative question in this regard is, as the patterns are being established across the data points, if the first data set classifies individuals suffering from depression, would the second iteration from the result obtained from the first iteration be used to evaluate depression severity? Researchers will also need to address the issues of a model's bias. In case a data set is obtained from a set of the population (say, individuals attending college), it is possible that the features used to predict depression would be different from another set of the population (say, middle-aged career individuals). As such, one will not be able to design a one-size-fits-all framework.

In terms of data collection, currently, the common means through which most AI-enabled technologies gather data is through text. This text could be through social media sites or users answering questionnaires delivered through chatbots. According to the investigations done in this field, a person's social media activity via content posted, such as the kind of words used and the frequency of the words used, can serve as an indicator for recognizing if

a person is suffering from depression. Additionally, the use of biological data is an effective way of diagnosing individuals who are suffering from depression due to the information that can be easily seen when the data is being read. Moreover, using non-psychiatric features to predict depression is also a helpful way of predicting if a user is suffering, as it helps to minimize the stigma that is attached to a depression diagnosis. It also helps in estimating or predicting the likelihood of a certain class of individuals suffering from depression if certain features are identified.

Finally, the researchers in this domain need to keep in mind that the acceptance of their proposed framework by the end-users can only come through model outcomes that are trustworthy and reliable. This can be achieved by making the proposed frameworks more transparent to end-users. These requirements necessitate the integration of explainable AI (XAI) frameworks to some degree. The incorporation of XAI frameworks will allow end-users with or without prior experience in the domain to be able to accept (trust) or reject a prediction if they understand the reasoning behind it. For example, if a system predicts whether a person suffers from a disease, the expert (doctor) can see the symptoms that contribute to the prediction to either accept or reject a prediction.

3.3.2 Proposed AI System Design

In this section, based on the challenges discussed in the previous section, the paper presents a system architecture of an AI-enabled approach for the diagnosis of mental health disorders like MDD. The goal of this work in progress is to design a decentralized system that will be capable of leveraging different AI-enabled approaches and work with different input data types with an explanation of the generated output. The motivation behind such a computational design is driven by the need for increased scalability and accessibility by

improving framework flexibility, which has not been addressed in the existing frameworks. The system design for the proposed AI framework is illustrated in Fig. 3.3.

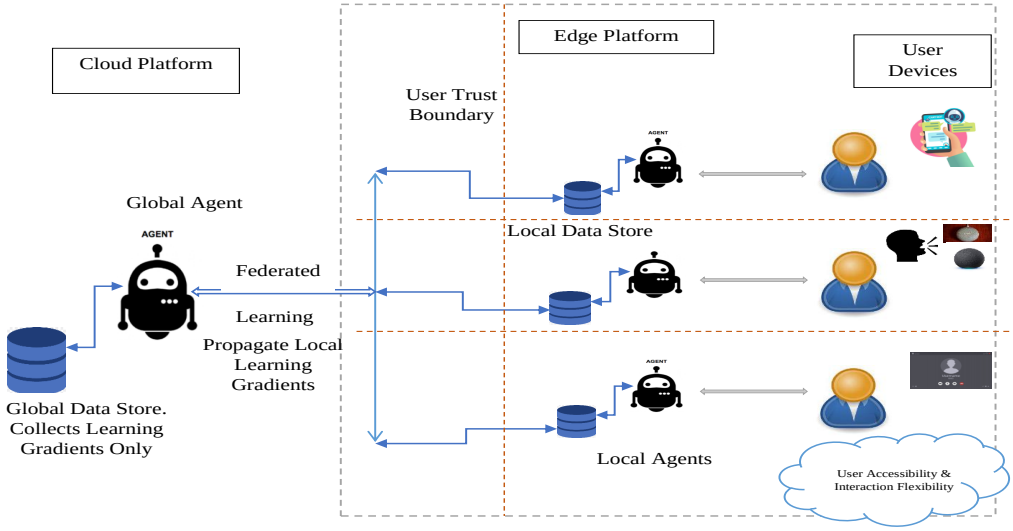


Figure 3.3: Proposed System Architecture for AI Framework

As shown in Fig. 3.3, the system is divided into three distinct platforms. First, the user-centric platform consists of various user devices like cellphones, tablets, or desktops. Second, the Edge platform will host a local agent (or LAI), which will improve the quality of service and preserve the privacy of user data. These local agents will learn from the data collected through the interactions with the users. A user will have the flexibility of interacting in any way that they want - either through text chats, audio interactions, or audio-video interactions. Depending on the user's choice, a suitable local agent will be deployed ad hoc to their Edge platform. The Cloud platform will host the global agent (or GAI), whose objective will be to synchronize the activities between different local AI agents hosted on various Edge Platforms. The learning on the Cloud platform will be done by receiving the transmitted learning gradients from the local agents instead of actual user

data, thereby improving user privacy and incorporating the concepts of federated machine learning. Given the decentralized and distributed nature of the proposed framework, it can then be dispensed as-a-service on an ad-hoc basis with applications similar to IBM Watson's Personality Insights [86] service but more geared toward diagnosing depression.

CHAPTER FOUR

RO II: SECURE AND TRUSTWORTHY INFRASTRUCTURE

The public Cloud computing paradigm is characterized by pooled hardware and software resources that are offered as a service to its users. The users upload their data to the data centers and develop their applications on the Cloud platform. However, as technology progresses, a majority of the next-gen applications are being built using the Internet of Things (IoT) infrastructure. These IoT applications generate huge volumes of data streams and constantly transmit them for processing to the Cloud platform, which can penalize the latency and bandwidth requirements. This, in turn, will affect the real-time decision-making capabilities of critical IoT-based applications like connected and autonomous vehicles.

Therefore, to address these requirements for IoT-based applications - lower bandwidth, lower latency, and faster processing - the Edge computing paradigm has been proposed by the research community. The edge computing paradigm is inherently decentralized and distributed in nature. The principal notion of this paradigm is to bring the computational processes closer to the source of the data [21]. In edge computing, there is an increased volume of generated data from the edge devices (devices that are used in edge computing such as IoT sensors and wearables), and Artificial Intelligence (AI) applications are leveraged to process this big data. AI applications deployed on the edge computing paradigm are referred to as Edge Intelligence (EI) applications [87, 88]. These applications provide real-time data processing, analysis, and decision-making without the need to rely heavily on the centralized Cloud computing paradigm. Currently, numerous application domains like smart grids [89], smart health care [90, 91], internet of vehicles [92], precision agriculture [93], and AR/VR applications [94] are incorporating EI in their system architecture.

The increasing demand for AI integration in edge computing shows the need for an accessible and dynamic approach to acquiring and deploying AI models. To meet this demand, this work envisions an *AI marketplace* for edge computing, a paradigm motivated by the likes of platforms such as the AWS marketplace, GenesisAI, and SingularityNET [95], where AI models (EI applications) are published for edge devices by third-party developers. The envisioned AI marketplace will have a decentralized peer-to-peer architecture and will allow various users to publish and subscribe to various AI models. This decentralized approach will ensure a more efficient distribution and access to AI models. However, though the envisioned notion of a decentralized AI marketplace for the users, by the users, enhances user-centric functionality, it also introduces security concerns, such as compromised AI models [87], collusion attacks, and data poisoning.

Furthermore, edge devices lack the necessary processing power to carry out traditional security measures as-is on the subscribed AI models before use, making them vulnerable to threats such as service manipulation through compromised AI models [21] and data inference attacks with a user's private information [28]. For example, in a scenario where a 3rd party developer trains an AI model with poisoned data, and a user downloads this published malicious AI model for service recommendations. Upon interaction with the AI model on their edge device, the user discovers that the AI model provides incorrect service suggestions, malfunctions frequently leading to service disruption, and exfiltrates sensitive data for other malicious purposes. This exposes the user to increased vulnerability to various security threats and makes the user lose trust in the AI model. Therefore, there is a need to validate the published AI models as trustworthy services before they interact with the user's data [88]. To address this challenge, we propose an integrity verification framework that establishes trustworthiness in the subscribed AI models before being deployed to the user's edge devices.

Existing literature has proposed different schemes to address the challenge of integrity verification in Edge Intelligence (EI) applications. The authors in [20] proposed a data security-enhanced Fine-Grained Access Control mechanism (FGAC) to ensure there is a secure transmission of data during data access in Mobile Edge Computing (MEC). The authors [96] proposed two integrity-checking protocols called ICE-basic and ICE-batch to allow a third-party verifier to validate data integrity without violating users' data privacy. However, in the current state of the art, data needs to be cached before integrity verification can be carried out, and it cannot provide (near) real-time integrity verification. Furthermore, a user's information is not preserved during the integrity verification process, causing data leakage that can be exploited for other malicious attacks, like identity theft. Therefore, there is a need for an integrity verification framework that verifies data integrity in Edge Intelligence (EI) applications before interacting with the user's data to ensure data privacy is not compromised.

Furthermore, the adoption of a centralized trust authority in a distributed and decentralized environment presents significant challenges, including a single point of failure, increased latency, and limited scalability [97]. This impacts the performance of Edge Intelligence (EI) applications, making them unsuitable for an edge computing environment. Therefore, there is a need for a decentralized trust authority. For our proposed integrity verification framework, we leverage Blockchain technology to validate the integrity of the subscribed AI models from a decentralized AI marketplace. The current state of the art incorporates Blockchain technology for various security use-cases on the edge platform such as a secure therapy application [98], an edge-based distributed control system [99], for authentication system [100], and a two-layer blockchain-based framework for novel consensus algorithm called Dynamic Random Byzantine Fault Tolerance (DR-BFT) to provide data integrity in a dynamic network in the edge computing architecture [44]. Additionally, authors in

[101] proposed a cache data integrity approach called EDI-V to verify the edge data on apps deployed on the edge. In EDI-V, the authors proposed a novel Merkle hash tree called Variable Merkle Hash Tree (VMHT) for auditing the data integrity to solve the problem of Edge Data Integrity (EDI). However, the integrity of the data is verified using a sampling strategy, and the lack of data integrity verification before the AI models interact with users' data is still unresolved. To address these highlighted shortcomings, the major contribution of this work is as follows:

- Introduce an Edge computing architecture to deploy AI models as a service to edge platforms, envisioned as a decentralized AI marketplace
- Propose a validation framework that validates the integrity of a subscribed AI model from the AI marketplace.
- Evaluate the effectiveness of the proposed framework by analyzing the feasibility of data poisoning attacks by varying percentages of compromised edge servers and on the correlation of the predictor values along with their input features.

4.1 Network Scenario and Threat Model

This section presents some of the preliminary concepts relevant to the proposed framework, such that it can achieve the following defined objectives:

- Deploy on-demand AI models as a service to users
- Validate AI model parameters
- Establish trustworthiness among edge servers

The proposed framework is built on three distinct layers for the overall edge architecture, as shown in Fig. 4.1. These layers are *edge device, edge server,* and *an AI marketplace* layer.

A three-layer architecture was adopted because it is considered to be safer when compared to a four-layer architecture [102].

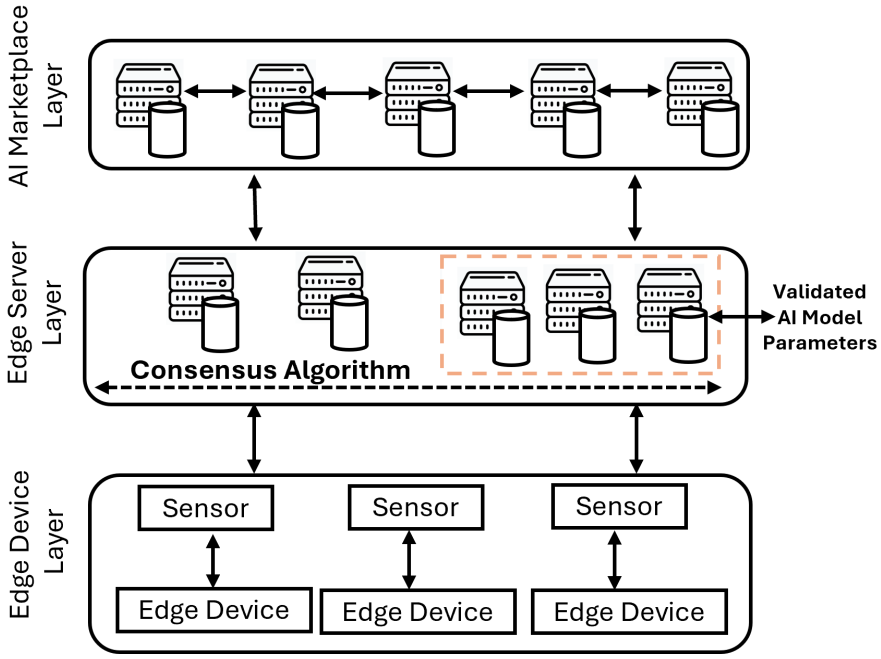


Figure 4.1: Conceptual overview of the Edge architecture and proposed framework deployment

4.1.1 Edge Architecture

4.1.1.1 AI Marketplace Layer: is similar to a traditional Cloud platform. This layer can support bulk and archival storage and batch data processing capabilities compared to the edge servers. The AI models subscribed by the users and deployed to the edge server(s) will be stored in this layer. Examples of AI models that can be hosted in this layer are Vosk [103], OpenPose [104], and DeepSpeech [105]. This layer will also be decentralized and distributed in nature.

4.1.1.2 Edge Server Layer: consists of servers located at the edge of the network. Edge servers provide real-time data processing and decision-making to the edge devices. Also, they can provide local storage to temporarily store data required for data processing. The edge servers provide computation and other resources needed by the edge devices. However, compared to traditional data centers, these servers have limited processing, memory, and storage capabilities. In our proposed framework, this layer stores the AI models deployed to the edge device upon a user's request, and the model updates are derived after interaction(s) with the edge device. The model updates received from the edge servers are transmitted to the edge devices. The servers in this layer are also decentralized and distributed, similar to the edge devices.

4.1.1.3 Edge Device Layer consists of intelligent edge devices, such as Internet of Things (IoT) sensors, wearables, and smart appliances, that connect to the system's overall onboard processing capabilities. The sensors in (embedded) or paired with (added) these devices introduce intelligence to the edge devices. This layer functionally supports the processes of data acquisition and transmission to the edge server layer. Due to the proliferation of edge devices, they are decentralized and distributed.

Fig. 4.2 shows the different steps of the data flow in our proposed framework. In our proposed framework, the data received from the edge devices is transmitted to the edge server layer. At the edge device layer, the user's request for data processing from a requested AI model triggers the deployment of the requested AI model from the AI marketplace. The edge server layer has the edge servers involved in the blockchain network. To validate the subscribed AI model parameters and establish trustworthiness among the edge servers, a consensus algorithm is used, as shown in Fig. 4.1. The consensus algorithm used is the Proof of Authority (PoA).

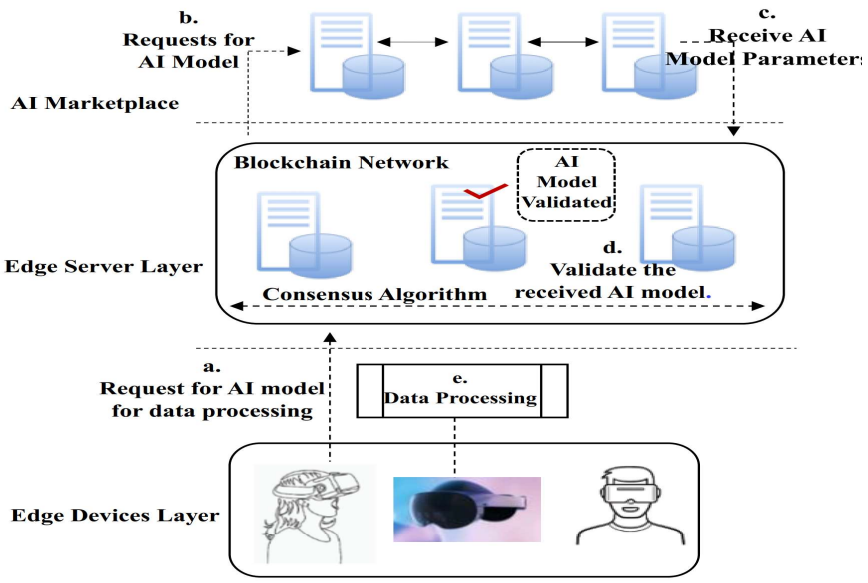


Figure 4.2: Proposed Framework Data Flow

Proof of Authority (PoA) is a reputation-based consensus protocol where trusted nodes are responsible for transaction validation and new block creation[106]. The trusted nodes are referred to as validators. Validators are chosen based on their identity, reputation, and trustworthiness. These validators exist as peer-to-peer nodes to validate transactions and play the role of a miner to create new blocks. Validators are identified by their unique ID or public key within the network. However, validators must act honestly to maintain their reputation, as malicious behavior can lead to their removal by the system administrator. PoA is a permissioned blockchain network because only the validators have the responsibility to validate transactions and create new blocks. PoA algorithms are less resource-intensive because no mining (computationally expensive work) is carried out, as it only requires a sign-off from trusted nodes to create a new block, rather than solving complex cryptographic puzzles. This makes the PoA a lightweight consensus algorithm. In PoA, there are no incentives.

PoA enables faster block generation time and requires low energy consumption. It promotes transparency and accountability as the block creators are easily identified. However, if 51% of the validators are malicious, the entire network is susceptible to malicious attacks. In our proposed framework, the PoA algorithm is deployed at the edge server layer, and the edge servers act as the nodes in the blockchain network.

Once the AI model parameters have been validated, the AI models interact with the data received from the edge device layer. The AI models hosted on the AI marketplace layer consist of different applications that render services to a specific group of end users. The AI models have the third-party developers' information and the application. In our proposed framework, we *assume* the hosting platform for the AI marketplace is secured, but the AI models need to be validated before they can be used by the users. In the AI marketplace layer, the status of the AI model parameter validation is unknown; therefore, it needs to be validated before interacting with the user's device.

4.1.2 Security Assumptions and Threat Model

The following assumptions were made in the proposed framework:

1. The network communication is encrypted such that there is no MITM attack.
2. Majority of the edge servers deployed in a region are non-malicious.
3. The datasets that are used by the different uncompromised edge servers are not poisoned and have highly correlated values.

Based on the aforementioned assumptions, the outline of the proposed framework's threat model is as follows.

- Data Poisoning Attack. The data used for training the AI models published in the AI marketplace can be compromised.

- Collusion Attack. A non-majority number of the edge servers can be compromised, and the compromised servers can also collude with the malicious AI model developer.

4.2 Proposed Framework

This section presents the methodology used by the proposed framework to validate the integrity of subscribed AI model parameters before interacting with a user. Table 4.1 summarizes the primary mathematical notation used throughout this chapter.

Table 4.1: Mathematical Notations Used in the Proposed Framework

Notation	**Description**
δ	A transaction representing an edge server's request to validate the parameters of a subscribed AI model.
ν_{v}	Unique identifier for the AI model vendor and the application name.
ρ_{r}	Unique identifier for the user and their edge device requesting the AI model service.
σ_{s}	Unique identifier for the edge server fulfilling the user's request.
μ_{m}	AI model data received from the vendor; assumed to be a data frame where rows represent individual data points and columns represent dataset features.
τ_{t}	Timestamp indicating when the transaction occurs in the blockchain network.
I_v	Data consistency metric used to verify consistency across the columns of the model dataset.
D_v	Duplicate value metric used to detect duplicate entries within the AI model dataset.
T_v	Data type verification metric used to ensure that each feature has the correct data type and lies within the expected range.
N_v	Missing value metric used to determine the number of missing values present in the dataset.
O_v	Outlier detection metric used to identify abnormal or significantly different values in the dataset.

4.2.1 AI Model Validation

The AI marketplace allows third-party developers to develop AI models and make them available for general use. However, the AI models' parameters cannot be vetted to be accurate until they have been validated using a consensus mechanism. An AI model's parameter validation ensures that the AI model was trained with the correct data, which affects the overall model's accuracy performance. It also ensures that the AI models are working in tandem with the other edge servers in the layer. Finally, it establishes trustworthiness among all the edge servers.

4.2.1.1 The Blockchain Network: In our proposed framework, the blockchain network exists only on the edge server layer. A transaction, δ, represents an edge server's request to validate the parameters of a subscribed AI model. A transaction, δ, from an edge server is depicted in eq. (4.1).

$$\text{Transaction}(\delta) = (\nu_{\text{v}}, \rho_{\text{r}}, \sigma_{\text{s}}, \mu_{\text{m}}, \tau_{\text{t}}) \tag{4.1}$$

Where ν_{v} represents the unique identifier for the AI model vendor and the application name, ρ_{r} represents the unique identifier for the user and their edge device, σ_{s} represents the unique identifier for the edge server fulfilling the user's request, μ_{m} represents to the AI model data received from the vendor and τ_{t} indicates the timestamp of the transaction. The transaction, δ, serves as input to the PoA consensus algorithm.

4.2.1.2 Proof of Authority (PoA): Edge computing architecture, a decentralized, distributed architecture, requires a consensus algorithm to validate the transactions without the presence of a central authority. Consensus algorithms help prevent fraud and double-spending. The consensus algorithm used in our framework is the Proof of Authority (PoA) technique. The PoA consensus establishes trustworthiness through the reputation and integrity of its

validators. The edge servers in the edge server layer are the participants in the PoA consensus mechanism. In our proposed framework, we assigned the nodes (edge servers) to three roles: Validator Nodes (VNs), Edge Nodes (ENs), and Server Nodes (SNs). The Validator Nodes (VNs) act as the validators and do not fulfill a user's request. Only the Validator Nodes (VNs) act as mining nodes. Server Nodes (SNs) are nodes with validated transactions included in the blockchain that fulfill a user's request, and Edge Nodes (ENs) are nodes with transactions yet to be validated and cannot interact with the Server Nodes (SNs). Upon transaction validation, an EN is promoted to be an SN.

In every transaction request in the PoA algorithm, the received AI model parameters are to be validated by the Validator Nodes (VNs). Model parameters represent the model data (μ_{m}) that a user requires for data processing. Upon receipt of the AI model parameters in the PoA algorithm, the integrity of μ_{m} is evaluated before the transaction can be validated. In our proposed framework μ_{m} is assumed to be a data frame made up of rows and columns, where rows represent the individual data points, and columns represent the features of the dataset. To validate the model's parameters, μ_{m} passed through an integrity check. The model data (μ_{m}) is only validated when the dataset satisfies the integrity check constraints. Once the model parameters are validated, the transaction request is fulfilled, a consensus is carried out, and the Edge Node (EN) is promoted to a Server Node (SN). An AI model parameter is deemed authentic if it passes the data integrity check carried out during the transaction validation.

4.2.1.3 Parameters Validation: The parameters of the subscribed AI model from the AI marketplace are validated for authenticity by the deployed edge servers. To validate the authenticity of the AI model parameters, a data integrity check was conducted. Conducting

an integrity check of the AI model parameters ensures there is consistency across the data, detects corrupted data present in the AI model parameters, and improves data trustworthiness.

To validate the model's parameters, μ_{m} was assessed for data consistency (I_v), duplicate values (D_v), data type verification (T_v), number of missing values (N_v), and number of outliers (O_v). For I_v, μ_{m} is analyzed for consistency across the columns, and T_v verifies that the right data type is present, alongside the data type, within the expected range. The O_v examines the number of outliers present in the dataset. This helps to identify values that could be erroneous or significantly different from the existing data in the AI model parameters. Additionally, N_v and D_v inspect for the number of missing values and duplicate values present in μ_{m}. These metrics were chosen because the results of these checks can be used as a metric to determine if μ_{m} has poisoned data before interacting with the actual user's data. The information submitted by the vendor in the AI marketplace will be used to inform the range of I_v and T_v. Overall, conducting an integrity evaluation on the data ensures the data is accurate and consistent and has no outliers. This ensures the AI model parameters are non-malicious and secure for the user to interact with.

4.2.1.4 Consensus Algorithm Comparison: In this subsection, a comparative study on Proof of Work (PoW), Proof of Stake (PoS), Delegated Proof of Stake (DPoS), and Proof of Authority (PoA) consensus algorithms is discussed to further inform our choice of PoA over the other available blockchain algorithms. The following metrics were used to carry out the comparison - *security, energy consumption, and adaptability to the edge architecture.*

In the Proof of Work (PoW) consensus algorithm, the next block in the blockchain network is created when a mathematical puzzle is solved. This puzzle is done based on a hash function, and solving the hash function serves as a confirmation of the transaction in the stored blocks in the blockchain network. A nonce value is used when a device is trying

to solve the hash. This value can be equal to or smaller than a certain given value, making the PoW algorithm easy to scale and flexible in any condition. PoW is a highly secure and decentralized algorithm. However, a lot of computational resources are wasted when mining and validating the blocks. In addition, solving the puzzle takes time, making it have less throughput; hence, it is not best suited for real-time applications.

In Proof of Stake (PoS), the next block creator is determined based on a combination of random selection, the creator's stake supply, and age. Since PoS uses the stake associated with the randomly selected node, it is more computationally efficient compared to the PoW algorithm. The PoS algorithm has a faster creation time, high throughput, and energy efficiency; however, it is not as secure as the nodes in the PoW algorithm because there is no validation to show that a node is genuine and can be trusted. The stake could be created using malicious transactions that cannot be verified. Additionally, the centralization of PoS would not work in a distributed, decentralized environment like the edge environment.

Delegated PoS (DPoS) is an improved form of PoS. In DPoS, the nodes in the blockchain network select representatives who vote to validate the blocks before adding them to the blockchain network. The number of nodes implemented is limited, and each block can determine the amount of time needed to publish each block. This algorithm is scalable, energy-efficient, and has low-cost transactions. However, it is a semi-centralized algorithm that is not best suited for the edge environment.

In the Proof of Authority (PoA) consensus algorithm, the next block in the blockchain network is created when the pre-approved validators validate the transaction. This algorithm is based on trust and relies on the integrity and reputation of the validators. It does not require a large amount of resources; resource-intensive computations are not carried out. This makes it suitable for edge computing environments. It is easily scalable and provides

decentralized control. However, the limited number of validators can affect the time required for a block to be generated.

4.3 Simulation and Results

This section presents the simulations that were carried out to validate the efficacy of the proposed framework. The simulations were done in two distinct steps. First, the application of the PoA consensus algorithm was simulated to validate model parameters. Then, to further supplement and evaluate the performance of our proposed framework, a feature correlation analysis was performed, which will show the accuracy in data consistency across the model parameters in the AI marketplace and the edge servers after the integrity verification has been conducted by the PoA. The following metrics were used to evaluate the performance of these simulations: transaction throughput, block creation time, and fault tolerance to evaluate the PoA consensus mechanism; and model precision score for the feature correlation.

4.3.1 Simulation Setup

This simulation was carried out on a platform with 16GB RAM; a 13th Gen Intel(R) Core(TM) i9-13900H 2.60 GHz computer with Windows 11 as the operating system. The details of the code base of this implementation can be found in the [107].

Datasets overview: The datasets used for the simulation were mostly sensor reading datasets. Datasets that had sensor reading data were selected because it is very similar to the type of data that would be generated in an edge architecture as the AI models interact with the edge devices. The datasets were further categorized into categorical datasets or continuous datasets. The occupancy dataset [1], activity sensor data [108] dataset, electrical grid [109], and banknote authentication [110] dataset were categorical datasets, while power consumption [111] is the only continuous dataset that was used.

Data poisoning setup: To evaluate the performance of the proposed framework, two datasets, the occupancy dataset and [1] electrical grid dataset [109], were poisoned to simulate the scenario of: (a) attacker inserting incorrect values (changing the expected data type in terms of the expected range and adding randomly generated outlier values), and (b) attacker deleting authentic values. The first scenario was applied to the occupancy dataset, whereas the second scenario was incorporated in the electrical grid dataset.

Setup for PoA algorithm: To validate the Proof of Authority (PoA) consensus algorithm, the proof-of-concept blockchain network was implemented using Jupyter notebooks. In the PoA simulation, there were 30 edge servers. 5 edge servers were assigned to the Validator Nodes (VNs), and 25 edge servers were assigned to be Edge Nodes (ENs), pre-AI model parameter validation. When a transaction is successfully validated, ENs are promoted to Server Nodes (SNs). The consensus threshold value was set to 3 since we considered 5 VNs in our simulation. This means a new block (promoting an EN to an SN) is created once the consensus threshold is satisfied.

Feature Correlation setup: To verify data consistency between the AI model parameters in the AI marketplace and the Server Nodes (SNs), the feature correlation in the datasets was evaluated. Four servers were simulated, where three servers act as edge servers and the fourth server hosts the training data, representing the AI Marketplace. Machine Learning models were implemented as AI models. For this simulation, 3 edge servers were simulated to evaluate the feature correlation because the data will be consistent irrespective of the number of servers used, as it is dependent on the output of PoA. The machine learning models implemented were Logistic regression, Naive Bayes classifier, Random Forest Regressor, Support Vector Machine (SVM), K-nearest neighbors (KNN), and Linear regression. The objective of this work is not to improve the accuracy of any model but to highlight the

effectiveness of the proposed work, which is to develop a framework that can validate the AI model obtained from the marketplace.

To simulate the AI model parameters for evaluating the feature correlation, each of the datasets was divided into 4 parts, where 55% of the data was for the marketplace AI model (that resembled a downloaded pre-trained model from the marketplace). In contrast, the remaining 45% of the data was equally distributed among the remaining 3 edge servers. The dataset was split using the 55:45 ratio to ensure each edge server model had sufficient data and was not overfitted or underfitted. Nonetheless, the *scope* of this simulation is to verify data consistency across the edge servers and the AI marketplace model, and not to improve the latent accuracy and/or precision of the underlying models subscribed from the AI marketplace.

4.3.2 Results and Discussion

This section presents the simulation results for the PoA consensus algorithm and the correlation between model data parameters. Based on the PoA simulation setup described in the previous section, Fig. 4.3 represents the log of the node activity in the blockchain network during the process of validating model data parameters, μ_{m}.

```
Node 1 added to the blockchain.
Node 2 added to the blockchain.
Node 3 added to the blockchain.
Node 4 failed the data integrity check.
Node 5 added to the blockchain.
Node 6 failed the data integrity check.
Node 7 failed the data integrity check.

Total number of nodes in the chain:  5
```

Figure 4.3: Log of Node Activity in the Blockchain

The performance of PoA was evaluated using the following performance metrics: fault tolerance, block creation time, and transaction throughput per validator. Fault tolerance is the ability of the PoA to continue functioning correctly and maintaining its integrity even when some Validator Nodes (VN) fail. In our simulation, the PoA achieved an 84.7% success rate and a 15.3% failure rate. This shows that the proposed PoA consensus mechanism is fault-tolerant even when faced with malicious nodes. Block creation time, as shown in Fig. 4.4, is the total time taken for a node to be added or discarded from the blockchain network. This shows that the proposed PoA consensus mechanism has a fast processing time, which reduces latency and improves system responsiveness, making it suitable for adoption in edge computing. Additionally, in the simulated PoA blockchain network, the average time taken to validate a transaction and create a new Server Node (SN) was between 0.0051s and 0.0137s. The values for the CPU usage for the entire PoA simulation were between 5.6% to 10%. The transaction throughput per validator refers to the number of transactions processed in a given time by the validators, as shown in Fig. 4.5.

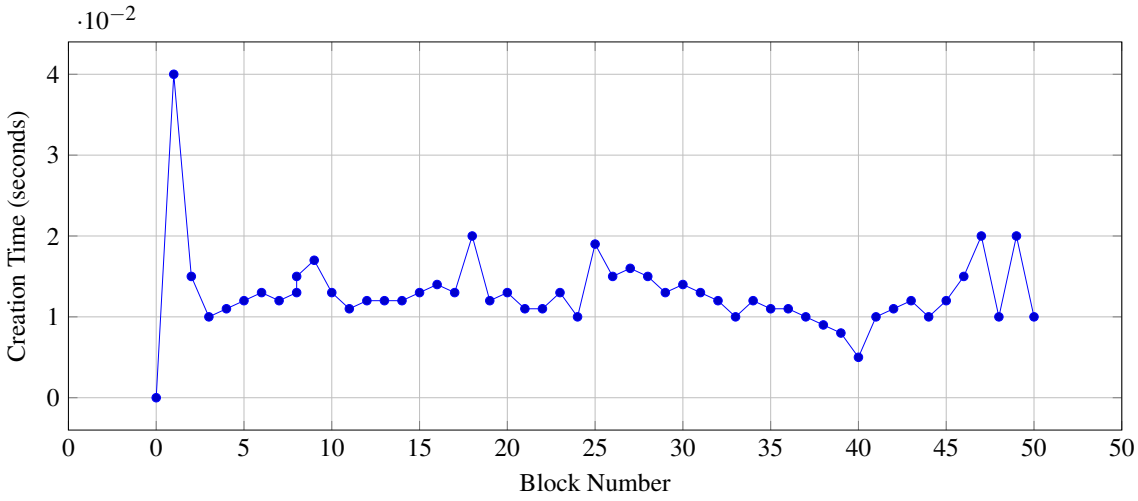


Figure 4.4: Time taken to validate and create new blocks after verifying AI model parameters (LR, GNB, SVM for the occupancy dataset[1])

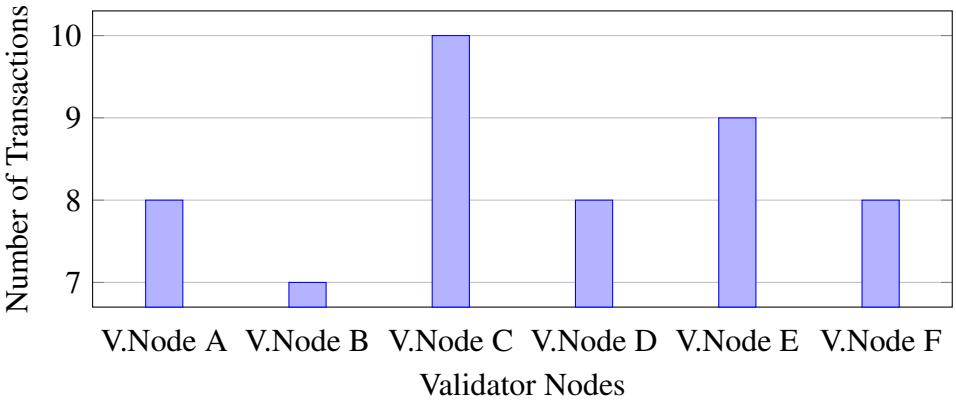


Figure 4.5: Transaction Throughput per Validator

The precision scores generated by the application of the proposed framework in the different datasets can be seen in Table 4.2 and Table 4.3. In Table 4.2, LR denotes Logistic Regression, GNB denotes Gaussian Naive Bayes, and SVM denotes Support Vector Machine. In Table 4.3, LR denotes Linear Regression, KNN denotes K Nearest Neighbor, and RF denotes Random Forest Regressor. In both Table 4.2 and Table 4.3, the Marketplace Model column depicts the precision on the dataset (which can be considered as training accuracy), which is used to train the simulated model that is downloaded from the marketplace by the different edge servers. The columns of edge servers 1, 2, and 3 represent the precision score of the dataset that is used to validate the pre-trained model that is downloaded from the marketplace.

From Table 4.2, it can be seen that the precision score of the various validating edge servers closely resembles the marketplace precision score. This shows the consistency in data integrity across all the edge servers (SNs) after the AI model parameters have been validated. To demonstrate that the properties of the data remain the same, the feature importance information of each of the datasets was gathered mentioned in Table 4.2. However, it is to be noted that in Table 4.3, the precision score of the marketplace model is much less than the

precision scores of edge servers 1, 2, and 3. This translates to the fact that the marketplace model might be facing an over-fitting or under-fitting challenge. This is essentially outside of the scope of the proposed framework and can be solved by selecting an AI model that could be more appropriate for the given dataset under consideration. The correlation of the different features on the predictor values can be seen in Fig. 4.6 to Fig. 4.7. In Fig. 4.6 to Fig. 4.9, we can see the correlation of the features on the predictor value closely resembles the AI model and the edge devices. However, the correlation does not match in Fig. 4.10 and 4.7. This is because the data used for validation in the edge devices does not closely resemble the data used for training the AI model of the marketplace. It should be noted that the correlation pattern is consistent irrespective of the number of edge servers deployed in a region, provided the transaction is validated.

Table 4.2: Precision score comparison for categorical datasets

Datasets	AI Model	Marketplace Model	Edge Server 1	Edge Server 2	Edge Server 3
Occupancy[1]	LR	97.5%	98.3%	97.9%	98.3%
	GNB	93.7%	94.1%	93.4%	94%
	SVM	97.9%	98.3%	97.9%	97.8%
Bank Note Auth.[110]	LR	98.9%	99.2%	99.2%	99.2%
	GNB	84.8%	84.6%	81.7%	84.6%
	SVM	98.6%	98.4%	99.4%	98.4%
Activity Sensor Data [108]	LR	99.8%	32.69%	66.3%	32.69%
	GNB	100%	32.6%	55.4%	32.6%
	SVM	41.4%	50.0%	41.4%	50.0%
Electrical Grid[109]	LR	100.0%	100.0%	100.0%	100.0%
	GNB	100.0%	100.0%	100.0%	100.0%
	SVM	100.0%	100.0%	100.0%	100.0%

Table 4.3: Precision score comparison for continuous datasets

Datasets	AI Model Classifiers	Marketplace Model	Edge Server 1	Edge Server 2	Edge Server 3
Power Consumption [111]	LR	81.0%	80.0%	81.0%	80%
	KNN	32.0%	85.0%	81.0%	85.0%
	RF Regressor	40.0%	83.0%	77.0%	83%

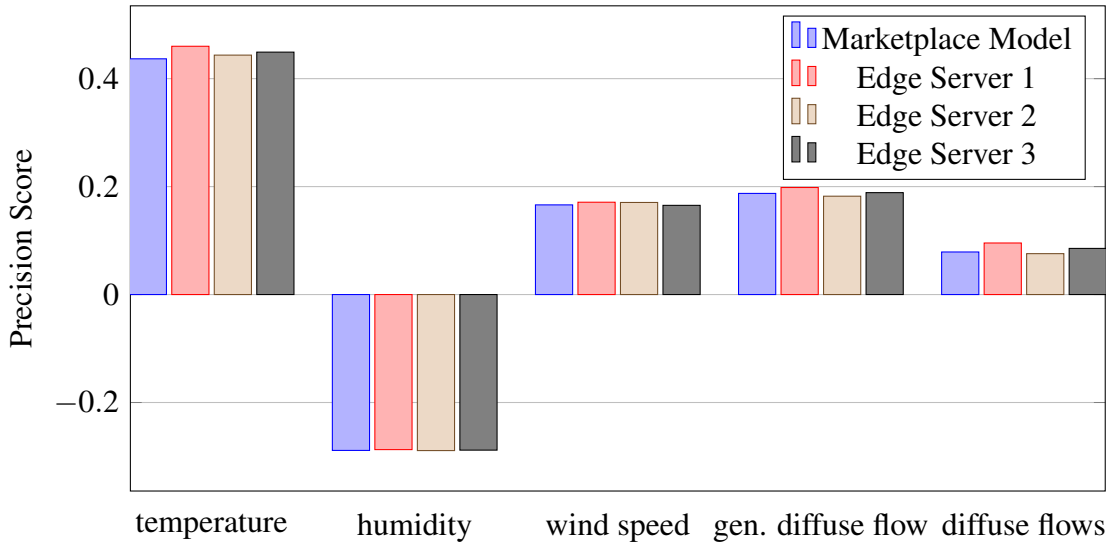


Figure 4.6: Feature Correlation of the Power Consumption Dataset

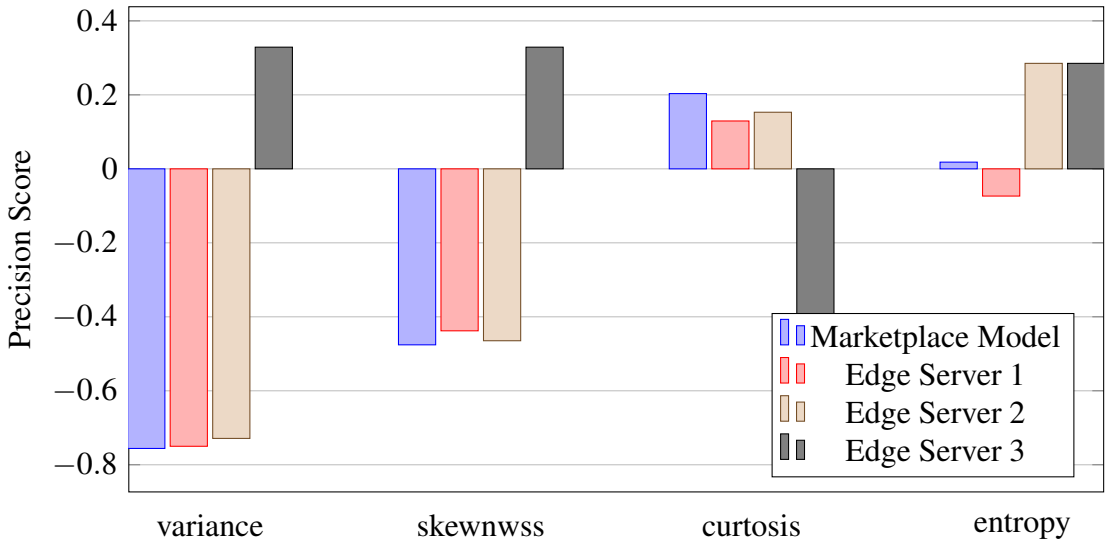


Figure 4.7: Feature Correlation of Bank Note Authentication Dataset

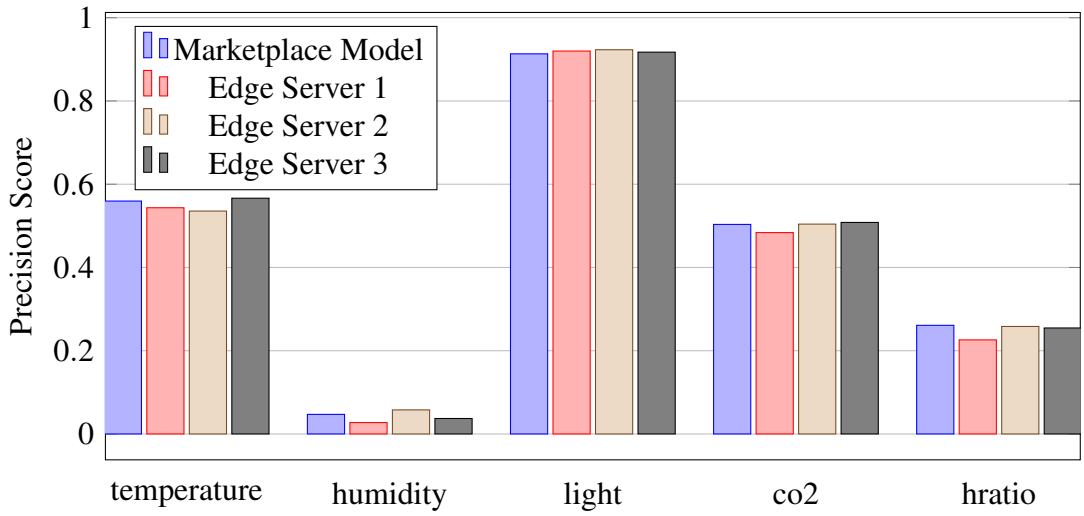


Figure 4.8: Feature Correlation of the Occupancy Dataset

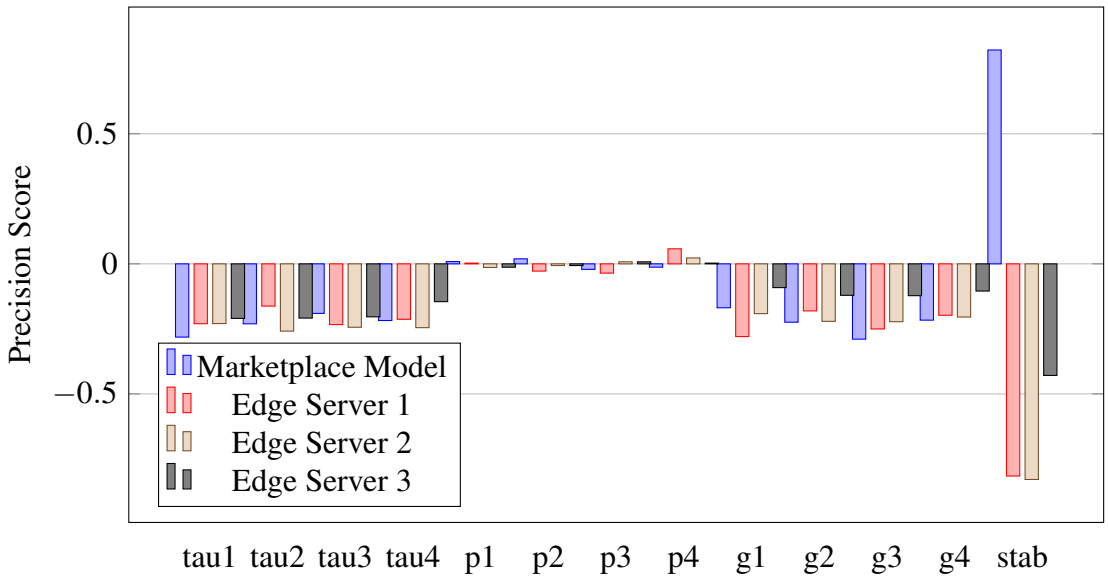


Figure 4.9: Feature Correlation of the Electrical Grid Dataset

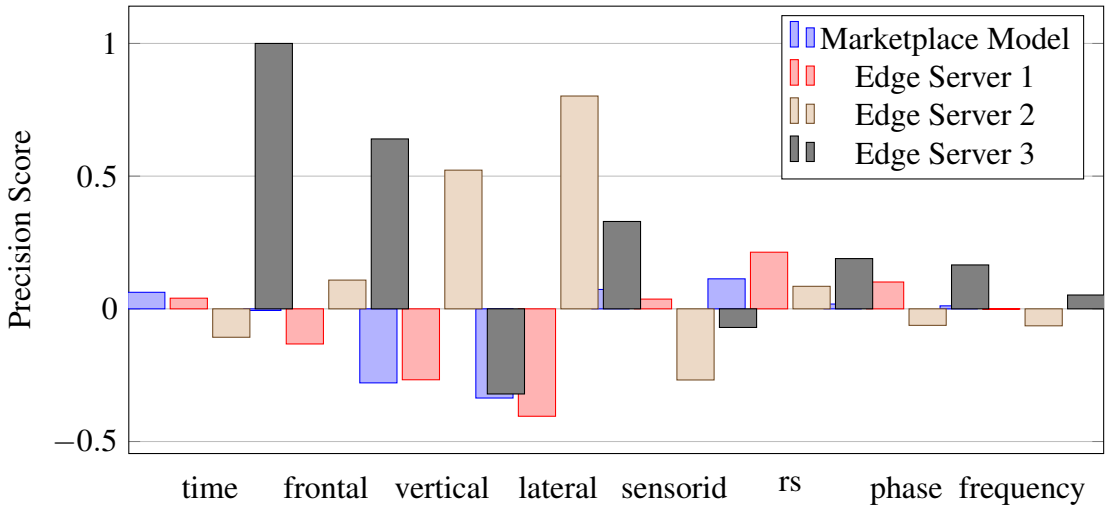


Figure 4.10: Feature Correlation of Activity Sensor Dataset

Table 4.4: Attack Possibility under different percentage of compromised edges:- ✗: framework cannot be compromised, ✓: framework can be compromised

% of Validator Nodes compromised	Without Collusion	With Collusion
$<= 10$	✗	✗
$<= 20$	✗	✗
$<= 30$	✗	✗
$<= 40$	✗	✗
< 50	✗	✗
> 50	✓	✓

In Table 4.4, the efficacy of the proposed framework was evaluated by applying different percentages of compromised Validator Nodes (VN) with and without collusion. In this scenario, collusion means that the marketplace model developer who used data poisoning while generating the pre-trained model is colluding with some of the edge servers. It can be seen that if the percentage of compromised or malicious VNs is less than 50%, then the efficacy of the proposed system cannot be compromised even under the impact of a compromised pre-trained model. This is as a result of the fact that with the majority of the VNs being uncompromised, the model will be deemed as compromised using the consensus algorithm after the uncompromised VNs have validated the transaction request of the edge servers. However, if the number of compromised edge servers is greater than 50% (though it is more difficult to achieve in PoA because the VNs are known and trusted entities), then it resembles the 51% attack scenario, where the majority of the nodes in the blockchain (in this case, the Server Nodes (SNs)) are compromised, and in such a scenario it is practically infeasible to execute a consensus algorithm.

CHAPTER FIVE

RO III: ACTIONABLE AND CAUSAL RECOURSE

Black-box machine learning (ML) algorithms in predictive applications can hinder users' trust in these applications. Explainable AI (XAI) seeks to address this challenge and provide transparency by generating human-understandable explanations of ML output [112]. Feature attribution XAI approaches, such as SHAP [113] and LIME [114], provide post-hoc explanations that highlight which features contributed most to the model's predicted outcome, clarifying *why* a specific outcome occurred. However, they do not capture the underlying cause-and-effect mechanism that determines how changes in the feature influence the outcomes. This may lead to misleading or infeasible explanations. Counterfactual Explanations (CFEs) address this challenge by highlighting influential features and indicating how deliberate changes to input variables could alter the model's prediction [22]. CFEs are post-hoc local explanations that provide why a decision occurred and how it could have been different by directly answering "what minimal change(s) would alter the decision?" [45]. They aim to embody causal reasoning by depicting how an alternative 'what-if' scenario would have altered the predicted outcome. This provides users with actionable insights that improve understanding of the model's outcome. A good counterfactual exhibits the characteristics of *validity, proximity, plausibility, effort, and diversity* [22]. Although these characteristics define what makes a good counterfactual, the practical implementations often face challenges that can limit their effectiveness.

Traditional methods for CFE generations [23, 46] often produce counterfactuals that are mathematically plausible but unrealistic in practice. For instance, a generated CFE can inform a user to reduce their age to qualify for loan approval, which is not realistic because a person's age cannot be reduced. Therefore, existing literature has proposed

different CFE generation frameworks to address this limitation. A plausibility-feasibility constraint language, PLAF, was introduced to encode domain-specific constraints, ensuring plausibility and feasibility of CFEs in real-time [59]. However, reliance on predefined domain-specific constraints limits its applicability, as the constraints may not capture the full complexity of real-world causal dynamics. Case-Based Reasoning was implemented to generate realistic CFEs based on explanatory coverage [115], but incomplete or biased case bases can produce unrealistic CFEs. Moreover, there is a lack of context-sensitive reasoning, as what was feasible for a past case may not be feasible for a current case. Similarly, [116] relied on native data points to ensure realistic feature combinations while enforcing sparsity and diversity. However, the use of generated datapoints does not guarantee realistic counterfactuals, as they capture only correlation instead of causal relationships. These shortcomings highlight the need for counterfactual generation frameworks to incorporate causal relationships, since only causally consistent changes can ensure that counterfactuals are both realistic and actionable in practice.

In this direction, [58] integrated causal relationships among the features to generate CFEs, but did not generate diverse counterfactuals simultaneously. This restricts the utility of counterfactuals, as it limits the feasible pathways a user can use to achieve a desired outcome. Further, the CoGS framework accounts for causal dependencies among the features using the FOLD-SE algorithm in a goal-directed Answer Set Programming System (CASP) to generate counterfactuals [117]. However, the reliance on CASP with the FOLD-SE algorithm makes the framework computationally intensive and limits its scalability to high-dimensional datasets like [118] and [119]. Likewise, [56] generates realistic counterfactual images using causal graphs in a variant of Adversarial Learned Inference (ALI). However, it is focused on bias mitigation in image classifiers, limiting its generation of CFEs for any arbitrary

predictive model. These frameworks, while integrating causality, are limited in their fusion of diversity, generalizability, and scalability.

To address these challenges, a model-agnostic counterfactual generation framework leveraging causal relationships was proposed to generate realistic, diverse, and actionable CFEs. Generating realistic and actionable CFEs requires enforcing sparsity and proximity to balance interpretability with feasibility, while also ensuring diversity to provide multiple actionable pathways to achieve the desired outcome. This requires balancing in tandem to identify the optimal solutions, which cannot be done manually as the search space for possible counterfactuals grows exponentially with the number of features. Therefore, the framework is formulated as a multi-objective optimization problem utilizing four existing objectives (validity, proximity, plausibility, and effort) and proposing a *novel* objective called recourse efficiency. Recourse efficiency helps capture the required cost and input to implement the actionable changes in a real-world setting. It assesses how achievable the desired outcome is for a user based on the suggested counterfactual changes. This allows users to focus on actionable, high-impact interventions. The main contributions of this research are:

- Design a model-agnostic CFE generation framework leveraging feature attribution to identify influential features and generate valid, diverse, and actionable counterfactuals, while enforcing causal validity through causal graphs and Structural Equation Models (SEM), using the NSGA-III optimization algorithm.

- Introduce a novel metric, recourse efficiency, which quantifies counterfactuals based on minimal, causally feasible interventions, ensuring users can achieve their desired outcomes using the most efficient actionable guidance.

- Validating the proposed framework's outcome using five tabular datasets (MovieLens [120], Adult Income[118], Default Credit Card [119], German Credit [121], and HELOC [122]), and comparing against existing frameworks of DiCE [123], MOC [55], and CARE [30] using performance metrics such as validity, plausibility, proximity, effort, recourse efficiency, and diversity.

5.1 Proposed Framework

The proposed model-agnostic counterfactual explanations (CFEs) generation framework comprises interconnected modules that cohesively work together to ensure the generation of valid, diverse, and causally plausible CFEs, as shown in Fig. 5.1. The Data and ML Model Module establishes the foundation of reliability for CFE generation by ensuring the model's predictions are well-calibrated and unbiased, thereby preventing the generation of invalid or biased CFEs. The Features Module reduces the search space through feature splitting and SHAP-based selection, enforcing the ethical and interpretive constraints. The Causal Inference Module introduces causal constraints through Directed Acyclic Graph (DAG) and Structural Equation Model (SEM) equations, ensuring feature perturbation occurs in a causally consistent manner. The Optimization Module balances the multiple objectives using the NSGA-III algorithm. A detailed description of these modules are as follows.

Table 5.1 summarizes the primary mathematical notation used throughout this chapter.

5.1.1 Data and ML Model Module

The quality of the input data directly influences the validity and plausibility of a generated CFE. All data preprocessing tasks for the input instance(s) are managed in this module. First, all null values are removed from the data instance to prevent the generation of CFEs based on incomplete information. Secondly, to preserve the semantic meaning of the features, the

Table 5.1: Mathematical Notation Used in RO III

Symbol	Description
$\mathscr{F}$	Complete feature space of the dataset used by the predictive model.
$\mathscr{A}$	Set of actionable features that can be modified by the user.
$\mathscr{I}$	Set of immutable features that cannot be modified due to ethical or legal constraints.
x	Original input instance from the dataset.
x'	Counterfactual instance generated to achieve a desired outcome.
G	Directed Acyclic Graph (DAG) representing causal relationships among features.
$f_j(\cdot)$	Structural function defining the causal mechanism for feature j.
η_j	Noise term capturing unobserved variability in the structural equation model (SEM).
σ_j^2	Variance of the Gaussian noise term associated with feature j.
$\mathscr{N}$	Set of non-actionable, non-immutable features updated through SEM propagation.
$m(\cdot)$	Black-box predictive model used to determine outcomes.
y^*	Desired outcome for the counterfactual explanation.
$\theta_v(x')$	Validity objective measuring whether the counterfactual achieves the desired outcome.
$\theta_p(x')$	Proximity objective measuring the distance between x' and the original instance x.
$\theta_c(x')$	Causal plausibility objective measuring adherence to the SEM.
$\theta_e(x')$	Effort objective measuring the number of sequential causal changes required.
$\theta_r(x')$	Recourse efficiency metric quantifying the efficiency of achieving the desired outcome relative to causal impact.
$C(x,x')$	Ordered set of causal interventions applied to transform x into x'.
$\text{Desc}(\mathscr{A})$	Set of descendant nodes of actionable features in the causal graph.
$\nu_c(x')$	Validity constraint ensuring the counterfactual exceeds the model decision boundary.
τ	Decision boundary threshold for predicted probability.

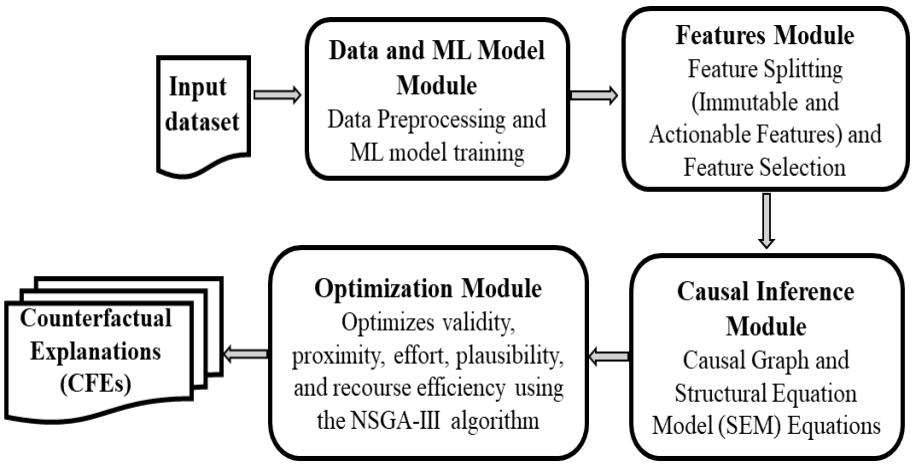


Figure 5.1: Proposed Counterfactual Generation Framework

categorical and ordinal features are encoded using label and ordinal encoding techniques, respectively. This ensures consistent data types that ML algorithms can effectively interpret. Thereafter, the encoded data is split into training and testing sets, and the standard scaler is used to prevent bias from features with larger numeric ranges. The training set is passed to the ML algorithm for the model to be trained. Our proposed framework is model-agnostic because it operates using the model's input-output behavior, without considering its internal architecture.

Additional data preprocessing is required when the minority class is underrepresented in imbalanced datasets. A dataset is considered imbalanced when target classes occur with substantially unequal frequencies, such as when at least one class represents less than 30% of the observations. Such an imbalance can cause the generated CFEs to be biased toward the majority class, undermining their fairness, reliability, and practical usefulness. It may also lead to inaccessible recourse, reducing the actionable guidance provided to a user. Therefore, to mitigate these challenges, we apply the Synthetic Minority Over-sampling Technique (SMOTE) and class weights techniques. SMOTE generates synthetic minority

samples by selecting a minority instance, identifying its nearest minority neighbors, and generating new samples along the lines connecting them. This increases the representation of the minority, while preserving the underlying data structure, and class weights further reinforce the importance of correctly classifying minority examples.

Further, to enhance the validity of CFEs by aligning predicted probabilities with true likelihoods, the trained models were calibrated using the isotonic regression technique. This entails fitting a non-parametric, monotonic function that maps a model's predicted probability to actual observed outcomes, to correct overconfident or underconfident predictions. This transforms unreliable models' output into a trustworthy probability estimate, improving CFE's plausibility. The calibrated models were cross-validated to ensure it is not overfit. The output of this module are the trained models with their initial predicted outcome, which will serve as a basis for the counterfactual generation. This serves as input to the features module.

5.1.2 Features Module

The features module improves the efficiency of our proposed framework by reducing its search space and ensuring the generated CFEs are ethical, realistic, and actionable. It does so by performing two distinct activities: *feature splitting* to distinguish actionable from immutable features, and *feature selection* to identify influential features. In feature splitting, the feature space, $\mathscr{F}$ (eq. (5.1)), is split into two sets: actionable ($\mathscr{A}$) and immutable ($\mathscr{I}$) features.

$$\mathscr{F} = \{c_1, c_2, \ldots, c_n\} = \mathscr{A} \cup \mathscr{I} : \mathscr{A}, \mathscr{I} \subseteq \mathscr{F} \text{ and } \mathscr{A} \cap \mathscr{I} = \emptyset \tag{5.1}$$

A user can feasibly change an actionable feature to achieve a desired outcome. Examples include income, level of education, and weight. Immutable features cannot be changed

due to biological, ethical, and legal limitations, such as race or birthdate. We considered *age* as an immutable feature within a temporal scope to provide the users with actionable CFEs. This ensures the user can implement the suggested changes immediately to achieve the desired outcome. Additionally, age is considered a protected attribute under fairness regulations in domains such as financial services and healthcare. Therefore, treating age as a mutable feature in CFEs can lead to biased suggestions, undermining fairness, as considered by [124].

Furthermore, we must ensure that counterfactuals target changes that actually alter the outcome and enhance interpretability. This requires highlighting the impact of each feature on the target variable and identifying the most influential features that affect the predicted outcome. To do so, we employ SHapley Additive exPlanations (SHAP) [84]. The integration of SHAP in our proposed framework ensures that the feature perturbation is close to the feature space by identifying the smallest set of impactful features. This makes the search space for the CFE generation smaller and relevant, as it avoids the exploration of irrelevant or weakly contributing features. Additionally, the variation of SHAP values across the input instances highlights different influential features for different cases, which aids in the creation of multiple actionable pathways in CFEs. Furthermore, feature selection aids in the generation of sparse CFEs, as only the most important features are perturbed in the feature space. This provides the user with actionable CFEs that are easier to understand. While SHAP excels at identifying the correlations that exist among the features to identify impactful features, it does not understand the causal relationships between the features, making it susceptible to generating unrealistic CFEs. To address this, we identify and rank the important features and feed them to the causal inference module.

5.1.3 Causal Inference Module

Enforcing causal validity in the feature interventions is imperative to the generation of realistic and actionable CFEs. We employ Directed Acyclic Graph (DAG) and Structural Equation Model (SEM) equations to construct causal graphs and establish causal structure among the features of the input instance.

The DAG is a causal graph that encodes the direction and dependency of cause-and-effect relationships among the features. This depicts how the changes in one variable influence other variables, facilitating feasible interventions to ensure causal consistency in CFE. To create causal graphs, we first use Fast Causal Inference (FCI) [125], a constraint-based search method that infers causal structure from conditional independence tests. However, FCI is limited in its functional assumptions, leading to reduced precision of inferred causal directions. Furthermore, it is sensitive to statistical errors, creating missed connections. Therefore, to improve the causal edge identification and have a fully DAG, we incorporate the DirectLiNGAM method [126], a functional causal discovery method, that applies linear non-Gaussian relationships to infer causal directions. The merge of FCI and DirectLiNGAM ensures that latent cofounders are detected, promoting robustness and accuracy of the inferred causal graph. This guarantees that the counterfactual generation is grounded in a thorough understanding of the underlying causal structure. Furthermore, integrating causal graphs with the ranked features highlights the nodes with the greatest influence on both the outcome and downstream causal pathways.

The inferred causal graph defines the permissible interventions, preventing unnecessary or redundant changes to multiple features by allowing only parent nodes to be directly perturbed. We then use SEM to propagate these interventions downstream to the child nodes, ensuring that CFEs remain consistent with the data manifold. In SEM, each feature is modeled as a function of its parent features in the causal graph, ensuring that changes to

actionable features propagate consistently through dependent features:

$$SEM(x'_j) = f_j\left(x'_{\mathrm{pa}(X_j)}\right) + \eta_j, \quad j \in \mathscr{N}, \quad \eta_j \sim \mathscr{N}(0, \sigma_j^2) \tag{5.2}$$

where x'_j is the counterfactual value of the j-th feature, $f_j(\cdot)$ is the structural function representing the causal mechanism for feature j, η_j is the stochastic noise that captures unobserved factors or randomness, and $x'_{pa(X_j)}$ are the counterfactual values of the parent features of j. $\mathscr{N}$ is the set of non-actionable, non-immutable features to which the SEM propagation is applied, and σ_j^2 controls the variance of the Gaussian noise, regulating the stochasticity in the SEM. The SEM encodes functional relationships, thereby preserving the causal consistency of the features. The integration of SEM with causal graphs also provides traceability, as users can see how changes in one feature propagate to others in the generated CFEs. This improves interpretability and trustworthiness.

Overall, this module applies two levels of validation for realistic CFE generation: internal consistency checks through causal graphs, which provide structural fidelity, and external validation through SEM, which ensures that the intervention effect is reflected in the generated CFEs, providing statistical fidelity. The generated causal graphs and SEM equations serve as constraints in counterfactual generation (eq. (5.3)), guiding the search space and maintaining causal plausibility.

$$x' \in \mathscr{S}(G, f) \iff \begin{cases} x'_i = x_i, & \forall i \in \mathscr{I}, \\ x'_a = u_a, & \forall a \in \mathscr{A}, \\ x'_j = SEM(x'_j) \end{cases} \tag{5.3}$$

where x' is a counterfactual and $S(G, f)$ is the set of valid CFEs generated by the causal graph, G, and the structural function, f, ensuring causal consistency. x'_i is the value of

an immutable feature, x'_a is the value of an actionable feature, directly set to the assigned intervention value u_a, and x'_j is the value of non-actionable, non-immutable features $j \in \mathscr{N}$, updated using $SEM(x'_j)$ to propagate the causal effects of the interventions.

5.1.4 Optimization Module

The generation of CFEs that are feasible for users to implement (x' in eq. (5.3)) requires the optimization of various objectives such that one can balance intelligibility with feasibility. We formalize these objectives as follows.

Validity ($\theta_v(x')$): ensures that the counterfactual achieves the desired outcome, as shown in eq. (5.4). Let $m(\cdot)$ represent the black-box predictive model, and the predicted outcome is $m(x')$.

$$\theta_v(x') = \begin{cases} 0, & m(x') = y^* \\ 1, & m(x') \neq y^* \end{cases} \tag{5.4}$$

where y^* represents the desired outcome. A CFE is valid when $\theta_v(x')$ is 0. This indicates that the counterfactual successfully achieves the desired outcome. Likewise, the CFE fails to flip the predicted outcome when $\theta_v(x')$ is 1. In our proposed framework, we enforced a validity constraint, ν_c, to ensure that the objectives are optimized within the feasible search space of only valid CFEs, as shown in eq. (5.5).

$$\nu_c(x') = \max\{0, \tau - P(y^* \mid x')\} \leq 0 \tag{5.5}$$

where $P(y^* \mid x')$ is the model's predicted probability for the desired class, τ is a confidence threshold corresponding to the decision boundary, and x' is a counterfactual. The value range of τ was set to $0 < \tau \leq 0.50$ to ensure a feasible yet inclusive ν_c, such that CFEs must exceed the model's decision boundary without being overly restrictive, hence avoiding ambiguous or overconfident CFEs.

Proximity ($\theta_p(x')$): measures how close the generated counterfactual x' is to the original data instance, x, reflecting the minimal perturbation required to achieve a CFE. We calculate this using the Euclidean distance (L2 norm) [127] in the feature space (eq. (5.6)). The L2 norm balances contributions across multiple features and ensures that $\theta_p(x')$ is consistently measured, independent of the feature rotations, and provides a smooth, differentiable metric suitable for optimization.

$$\theta_p(x') = \|x' - x\|_2 \tag{5.6}$$

Where $\|\cdot\|_2$ is the L2 norm. Lower values of $\theta_p(x')$ indicate that x' is closer to x in feature space. Since this optimization module is designed as a minimization problem, the best numerical value for all these metrics is as close to zero for any chosen numerical scale ([0,1], [0,10], or [0,100], etc.). For our framework, we select the [0,10] scale and define values $0 \leq \theta_p(x') \leq 1$ to denote minimal or no required change, values $1 < \theta_c(x') < 5$ reflect noticeable but achievable modifications, and values $\theta_p(x') \geq 5$ indicate substantial, less practical changes.

Causal Plausibility ($\theta_c(x')$): evaluates the feasibility of x' by measuring its adherence to the SEM. We calculate this value using the negative log-likelihood of observing the x' under the Gaussian distribution predicted by the SEM, as shown in eq. (5.7).

$$\theta_c(x') = \sum_{j \in \mathscr{N}} \left[\frac{1}{2}\log(2\pi\sigma_j^2) + \frac{(x'_j - \mu_j(x'_{\text{pa}(j)}))^2}{2\sigma_j^2} \right] \tag{5.7}$$

where $j \in \mathscr{N}$ is described in eq. 5.2. The mean and variance of the conditional Gaussian distribution for feature j are given by $\mu_j(x'_{\text{pa}(j)})$ and σ_j^2, respectively. The application of negative log-likelihood quantifies how the feature perturbations respect the causal structure, through a smooth, differentiable function with respect to the CFE. Lower values of $\theta_c(x')$ indicate higher plausibility, meaning the CFE is more realistic. We define values $0 \leq$

$\theta_c(x') \leq 1$ as very high plausibility with strong adherence to causal relationships, values $1 < \theta_c(x') < 3$ indicate moderate plausibility with minor causal violations, and values $\theta_c(x') \geq 3$ reflect low plausibility with substantial causal violations, reducing the realism of the counterfactual.

Effort ($\theta_e(x')$): measures the total number of sequential changes required to transform x into x' (eq. (5.8)). This captures causally consistent action paths towards the CFEs and reflects the practical cost for a user to achieve the desired outcome.

$$\theta_e(x') = |C(x,x')| \tag{5.8}$$

where $C(x,x')$ denotes the ordered set of changes applied under the causal graph G. The value $\theta_e(x')$ accounts for both direct changes to actionable features, $\mathscr{A}$, and indirect causal propagation along the SEM and DAG, forming the recourse chain. Unlike sparsity, which only counts the number of perturbed features, effort also captures the order of changes and causal dependencies. Lower values of $\theta_e(x')$ indicate that x' requires less effort to achieve. We define values $0 \leq \theta_e(x') \leq 5$ as very low effort, requiring minimal changes, values $5 < \theta_e(x') < 10$ reflect moderate effort with multiple but feasible changes, and values $\theta_e(x') \geq 10$ indicates high effort, with substantial sequential modifications required, making the CFE less practical.

Recourse Efficiency ($\theta_r(x')$): We propose this new metric to quantify how effectively a x' achieves the desired outcome relative to the causal impact of actionable changes on their descendants. It is defined as the ratio of the total effort required to reach x' to the total magnitude of causal changes in downstream features, as shown in eq. (5.9). This encourages the generation of CFEs that are both achievable and minimally disruptive, penalizing those

that induce unnecessarily large changes.

$$\theta_r(x') = \frac{|C(x,x')|}{\sum_{j \in \text{Desc}(\mathscr{A})} |x'_j - x_j| + \varepsilon} \tag{5.9}$$

Where $|C(x,x')|$ denotes effort ($\theta_e(x')$), $\text{Desc}(\mathscr{A})$ is the set of descendants of actionable features, and ε is a small constant ($0.0 < \varepsilon \leq 0.5$) to prevent division by zero. Lower values of $\theta_r(x')$ indicate that the CFE achieves the desired outcome efficiently, while minimizing unnecessary disruption. We define values $0 \leq \theta_r(x') \leq 0.50$ as very high efficiency, where the CFE achieves the desired outcome with minimal $\theta_e(x')$ relative to its causal impact, values $0.50 < \theta_r(x') < 1$ reflects moderate efficiency, with some unnecessary or indirect changes and values $\theta_r(x') \geq 1$ indicates low efficiency, where the counterfactual is less practical due to disproportionate effort or excessive downstream impact.

We incorporate all these objectives as a multi-objective minimization problem as shown in eq. 5.10 and solve it using the NSGA-III algorithm [128].

$$\begin{aligned} \textit{minimize}: \quad & \theta_v(x'), \theta_p(x'), \theta_c(x'), \theta_e(x'), \theta_r(x') \\ \text{s.t.} \quad & \nu(x') \leq 0, \text{ and } x' \in \mathscr{S}(G,f) \end{aligned} \tag{5.10}$$

where $\nu(x')$ is the validity constraint described in eq. (5.5) and $x' \in \mathscr{S}(G,f)$ is the SEM constraints described in eq. (5.2). NSGA-III uses a reference-point-based selection process and is well-suited for scalable optimization in high-dimensional, constrained, and hybrid (discrete and continuous) spaces. The predefined set of reference directions guarantees interpretability by generating CFEs that reflect different priorities. Although our proposed framework is capable of generating multiple CFEs for each input data instance, to preserve practical utility, we limit the number to four. This provides a minimal and diverse set of representative solutions to the objective tradeoffs, without risking biased representation if

fewer than four are considered. Four options strike a balance between diversity and cognitive tractability, as established in [129].

5.2 Simulation and Results

The objective of our simulation is to demonstrate the effectiveness of our proposed framework for generating realistic, actionable CFEs. Simulations were performed on a single workstation with 16GB of RAM and a 13th Gen Intel(R) Core(TM) i9-13900H 2.60 GHz CPU, using Python 3.13. The details of the codebase for this simulation can be found in [130].

Datasets: Five tabular datasets were selected (Table 5.2) for their diverse characteristics. Adult Income [118] predicts whether an individual's income exceeds $50K/year based on

Table 5.2: Dataset used in our simulations

Dataset	**Features**	**Samples**	**Decision Variable**
Adult Income [118]	14	30162	($\leq$ 50K, >50K)
HELOC [122]	24	10459	(Bad, Good)
MovieLens [120]	24	50000	(Like, Dislike)
German Credit[121]	10	1000	(Bad, Good)
Default Credit Card [119]	25	30000	(No, Yes)

demographic and employment attributes. Home Equity Line of Credit (HELOC) [122] is used to predict whether an applicant is likely to be late on payment, based on anonymized credit history and financial behavior. MovieLens [120] predicts individual user ratings for movies. In our scenario, we introduced a new binary target variable, *Like*, which reflects the popularity of a movie based on the total number of received user ratings. German Credit [121] classifies individuals as good or bad credit risks based on financial creditworthiness.

Default Credit Card [119] contains demographic and financial data of credit card clients from Taiwan and is used to predict whether a client will default on their next month's payment. Data preprocessing tasks were applied (Section 5.1.1) and features were categorized into actionable and immutable sets (Section 5.1.2). The immutable features set contains age, sex, gender, race, marital status, ethnicity, birth date, ID, and native country. All other features were considered actionable.

Model Training: The datasets were split into training and test sets using a 70:30 ratio. Features in the training sets were standardized using the StandardScaler method. Training data was used to train the black box models of Logistic Regression, Support Vector Model (SVM), and Random Forest classifiers. Classifier algorithms have defined decision boundaries and target variables, making them suitable for CFEs. Classifier algorithms have a defined decision boundary and a target variable, making them suitable for CFEs. Whereas clustering algorithms lack target variables and decision boundaries, as their primary function is to group similar data points, making them unsuitable for our context. These models were calibrated using isotonic regression with 5-fold stratified cross-validation to improve the alignment between predicted and actual outcomes. Table 5.3 shows the accuracy score of the implemented models.

Table 5.3: Accuracy of trained black box ML models

Dataset	**Logistic Regression**	**Random Forest**	**SVM**
Adult Income [118]	0.765	0.834	0.784
HELOC [122]	0.721	0.723	0.720
MovieLens [120]	0.573	0.567	0.598
German Credit [121]	0.732	0.643	0.567
Default Credit Card [119]	0.679	0.803	0.784

Algorithmic Parameters and Performance Metrics: The algorithmic parameters of the NSGA-III, such as population size, number of generations, mutation rate, and crossover probability, were empirically evaluated to predetermined values of 100, 126, 0.9, and 0.5, respectively. The optimization objectives from Section 5.1.4 were used as performance metrics to quantitatively evaluate the quality of the generated CFEs. The defined value ranges that are used to assess the quality of CFEs across these objectives are summarized in Table 5.4.

Table 5.4: Conceptual value ranges for CFE objectives

Objectives	**Low (Very Good)**	**Moderate**	**High (Poor)**
$\theta_p(x')$	$0 \leq \theta_p \leq 1$	$1 < \theta_p < 5$	$\theta_p \geq 5$
$\theta_c(x')$	$0 \leq \theta_c \leq 1$	$1 < \theta_c < 3$	$\theta_c \geq 3$
$\theta_e(x')$	$0 \leq \theta_e \leq 5$	$5 < \theta_e < 10$	$\theta_e \geq 10$
$\theta_r(x')$	$0 \leq \theta_r \leq 0.50$	$0.50 < \theta_r < 1$	$\theta_r \geq 1$

5.2.1 Quantitative Evaluation Using CFE Objectives

Table 5.5 shows the objective values obtained for multiple CFEs generated for an original data instance, x. Each row represents a distinct CFE, allowing comparison of $\theta_p(x')$, $\theta_c(x')$, $\theta_e(x')$, and $\theta_r(x')$ across the different datasets.

First, the framework generates valid CFEs ($\theta_v(x') = 0$). The CFEs also maintain strong faithfulness to their causal models, with $\theta_c(x')$ values ranging from very good to moderate in all datasets except MovieLens, where weak and noisy feature correlations limit feasible causal changes. This variation reflects feature mutability and correlations, as datasets with more actionable features (Default Credit Card, HELOC) require smaller changes, while those with immutable or highly correlated features (Adult Income, German Credit,

MovieLens) require larger adjustments to maintain causal faithfulness. Consequently, the changes needed to obtain these CFEs range from minimal for Default Credit Card, to moderate ($1.0 < \theta_p(x') < 5$) for HELOC, MovieLens, and German Credit, and substantial ($\theta_p(x') \geq 5$) for Adult Income, because prioritizing $\theta_v(x')$ over minimality ($\theta_p(x')$) leads to larger adjustments for the user to achieve the desired outcome.

Table 5.5: Objective-Based Evaluation of CFEs for Proposed Framework

Dataset	$\theta_v(x')$	$\theta_p(x')$	$\theta_c(x')$	$\theta_e(x')$	$\theta_r(x')$
Adult Income [118]	0.00	5.67	2.16	5	1.88
	0.00	5.17	2.03	5	1.89
	0.00	5.17	1.82	5	2.06
	0.00	5.17	1.96	5	2.08
HELOC [122]	0.00	1.93	0.0	22	0.29
	0.00	1.93	0.0	23	0.28
	0.00	1.93	0.0	23	0.28
	0.00	1.93	0.0	22	0.29
MovieLens [120]	0.00	2.78	4.15	18	0.46
	0.00	2.79	3.77	18	0.50
	0.00	2.77	4.23	18	0.53
	0.00	2.81	3.63	18	0.50
German Credit [121]	0.00	1.05	2.49	5	0.33
	0.00	1.06	2.30	5	0.36
	0.00	1.07	1.83	5	0.53
	0.00	1.08	2.02	5	0.41
Default Credit Card [119]	0.00	0.85	1.02	18	0.14
	0.00	0.83	0.85	18	0.15
	0.00	0.82	2.22	18	0.17
	0.00	0.84	1.25	18	0.15

$\theta_v(x')$, $\theta_p(x')$, $\theta_c(x')$, $\theta_e(x')$, and $\theta_r(x')$ are validity, proximity, plausibility, effort, and recourse efficiency, respectively. A lower value is better.

The effort ($\theta_e(x')$) to achieve these CFEs is very low for Adult Income and German Credit datasets ($0 \leq \theta_e(x') \leq 5$), but substantial for HELOC, MovieLens, and Default Credit Card ($\theta_e(x') \geq 10$), because these datasets require broader and more distributed feature changes. This arises from the need to respect causal dependencies under the constraint in eq. 5.2, which necessitates larger, coordinated changes to maintain plausible CFEs. The MovieLens dataset is an exception, as its weak, densely correlated features leave no clear causal pathways, forcing many coordinated adjustments that yield high effort ($\theta_e(x')$) without ensuring high plausibility ($\theta_c(x')$). Finally, for the newly proposed recourse efficiency metric ($\theta_r(x')$), HELOC, German Credit, and Default Credit Card perform very well ($0 \leq \theta_r(x') \leq 0.50$), indicating that user interventions produce substantial downstream causal effects. This reflects the cost–benefit dynamics of user actions, where even with similar $\theta_e(x')$, CFEs differ in $\theta_r(x')$ due to differences in causal structure and how strongly interventions propagate. Default Credit Card achieves substantial causal impact per action, whereas MovieLens yields a lower $\theta_r(x')$ despite comparable effort. Adult Income yields poor $\theta_r(x')$ results ($\theta_r(x') \geq 1.0$) because interventions on actionable features have weak causal influence and propagate minimally through the causal graph. Overall, the proposed framework naturally penalizes datasets with weak or unstable causal structure, demonstrating its ability to identify scenarios where CFEs are minimally actionable yet maximally effective.

Discussions on Recourse Efficiency (θ_r): With θ_r, users can gauge how strongly their interventions propagate through the causal structure, an insight that the existing $\theta_e(x')$ metric alone cannot provide. For example, CFEs from the MovieLens and Default Credit Card datasets require similar effort yet produce markedly different downstream effects, a contrast visible only through θ_r. A CFE may seem efficient because it requires little effort (e.g., Adult Income) or small changes (e.g., German Credit), but if its interventions have weak causal influence, it is less actionable. By capturing the ratio of effort to causal impact,

θ_r helps users determine whether additional actions meaningfully increase their leverage. Without this metric, users would be unable to assess whether the required effort is justified. Thus, θ_r enables more informed, cost-aware selection of the most effective pathway to achieve the desired outcome. For further validation, all observations from the instance-based evaluations are reinforced using parallel coordinates plots (PCP), which capture the tradeoffs among the objectives. These plots illustrate the variety of Pareto-optimal counterfactuals and confirm the relationship among the objectives for each dataset, and are described in the next sub-subsection.

5.2.2 Performance of CFEs as Pareto Front Solutions

Fig. 5.2 presents the interactions between the objectives, where the tradeoffs occur, and which CFE offers the best overall balance. Since the value of validity ($\theta_v(x')$) is constant across all CFEs (Table 5.5), they are not considered in the plots. The output scores are normalized ([0,1]) for ease of visualization, and lower values correspond to better CFE quality. The number of solutions presented in the PCP represents the number of generated Pareto-optimal solutions.

In Figs. 5.2a, 5.2d, and 5.2e, CFEs with higher $\theta_p(x')$ achieve higher θ_c, while low $\theta_p(x')$ often results in reduced $\theta_c(x')$, making them less practical. This occurs because the datasets' actionable features exert strong causal influence, so larger coordinated adjustments yield more effective downstream effects, highlighting the tradeoff between minimality and plausibility. In contrast, Fig. 5.2b shows a more extreme version: despite a small number of solutions, CFEs are perfectly plausible ($\theta_c(x') = 0$) because the dataset contains no immutable features. Regarding recourse efficiency, larger $\theta_p(x')$ generally leads to higher $\theta_r(x')$, except in Fig. 5.2a, where $\theta_r(x')$ is poor (scores ≥ 1.0) because adjustments to actionable features have minimal effect on downstream outcomes. All other datasets

(Figs. 5.2b, 5.2d, 5.2e) exhibit higher $\theta_r(x')$, reflecting that changes to actionable features propagate effectively and influence the predicted outcome. Overall, the Pareto set highlights clear tradeoffs: minimal changes may be inefficient or implausible, whereas CFEs with higher plausibility produce stronger causal impact and better recourse efficiency, reinforcing the tradeoffs identified in Section 5.2.1.

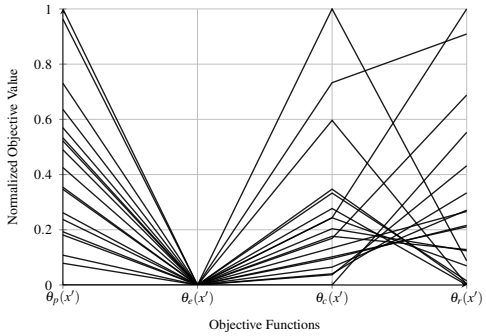
5.2.3 CFE Distribution Score ($\delta_s(x')$)

To identify solutions that are feasible and distinct, a cluster-based CFE distribution score, $\delta_s(x')$, was introduced. This quantifies the variation among valid counterfactuals for a given instance, x, by capturing both the spread of CFEs across the feature space and their balanced distribution. It adaptively determines the number of clusters using silhouette analysis and quantifies diversity as the product of cluster coverage and normalized entropy, reflecting multiple, distinct, and actionable pathways a user can take to achieve the desired outcome.

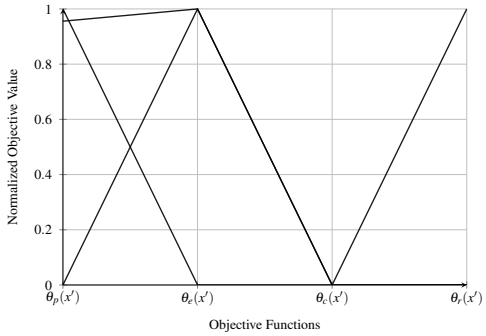
$$\delta_s(x') = \underbrace{\frac{\sum_{k=1}^{K} \mathbf{1}_{[n_k>0]}}{K}}_{\text{Coverage}} \times \underbrace{\frac{-\sum_{k=1}^{K} P_k \log P_k}{\log K}}_{\text{Normalized Entropy}} \tag{5.11}$$

where K is the total number of clusters determined via the adaptive K-Means clustering based on the silhouette score. n_k refers to the number of CFEs in cluster k, and $\mathbf{1}_{[n_k>0]}$ is an indicator function that returns 1 if cluster k contains at least one counterfactual, otherwise 0. $P_k = \frac{n_k}{\sum_{j=1}^{K} n_j}$ is the probability of CFEs in cluster k, and it is normalized by $\log K$ to ensure entropy ranges from 0 to 1. Unlike traditional distance-based metrics used in [123] and [131], $\delta_s(x')$ provides a more informative measure of variety by accounting for both the balance and distributions of CFEs across the feature space.

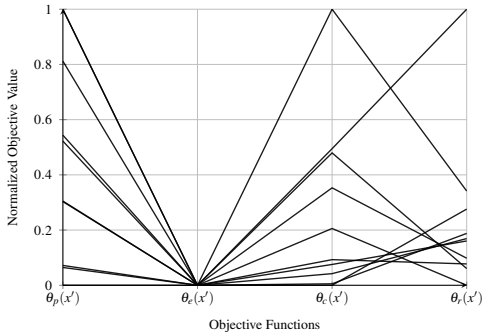
A high $\delta_s(x')$ value indicates that CFEs are widely dispersed and evenly distributed across clusters. Table 5.6 reports the $\delta_s(x')$ values for all valid CFEs for an original data instance, x. HELOC achieves the highest $\delta_s(x')$ (0.96) despite producing only five CFEs.



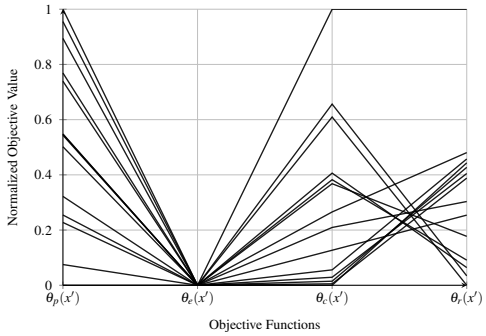
(a) Adult Income



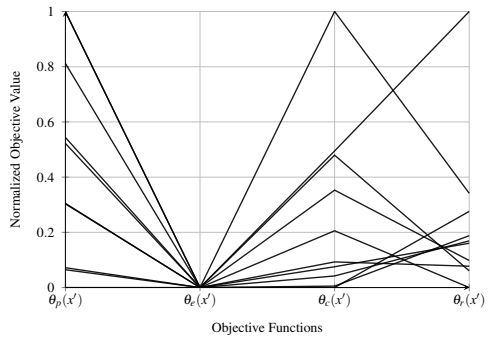
(b) HELOC



(c) MovieLens



(d) German Credit



(e) Default Credit Card

Figure 5.2: Parallel Coordinates Plots for various datasets showing Pareto-optimal performance.

This results from the strict causal constraints in eq. 5.2, reflected in its low $\theta_c(x')$ values, which permit only a few feasible modifications, but those solutions fall in distinct regions of the feature space. Default Credit Card also shows a high $\delta_s(x')$ (0.93) across ten CFEs. The more flexible feature space supports broader exploration of tradeoffs, as evidenced by variation in $\theta_c(x')$, leading to more heterogeneous recourse options. Adult Income and MovieLens exhibit moderate heterogeneity (0.86 and 0.90), reflecting more constrained but still balanced coverage. These patterns arise from how $\theta_e(x')$ and $\theta_c(x')$ interact with each dataset's structure. Adult Income's mix of categorical and continuous features limits smooth exploration, while MovieLens's highly correlated behavioral attributes restrict movement across clusters.

Table 5.6: CFE Distribution Score and Cluster Composition

Dataset	$\delta_s(x')$	**#CFEs**	**#Clusters**	**Cluster Composition**
German Credit	0.65	18	2	0^{15}, 1^{3}
Adult Income	0.86	14	2	0^{4}, 1^{10}
MovieLens	0.90	16	2	0^{5}, 1^{11}
Default Credit Card	0.93	10	5	0^{1}, 1^{3}, 2^{3}, 3^{1}, 4^{2}
HELOC	0.96	5	3	0^{2}, 1^{2}, 2^{1}

Cluster composition is given as clusterID$^{\#CFEs}$, indicating #CFEs per cluster.

In contrast, the German Credit exhibits limited distribution ($\delta_s(x') = 0.65$), with CFEs concentrated in a small set of clusters. This arises from its constrained feature space, as indicated by its $\theta_p(x')$ and $\theta_c(x')$ values. Although 18 CFEs were produced, the optimization consistently returns to the same region, where only a few features *job, duration, credit amount, and purpose* change while others remain static (Table 5.7). Thus, generating

feasible CFEs depends on similar adjustments, leading to a single dominant cluster and limited distribution.

Overall, the results reveal that datasets with richer feature structure and looser causal dependencies, such as HELOC and Default Credit, promote higher $\delta_s(x')$, which enhances users' autonomy. Conversely, lower $\delta_s(x')$ in the MovieLens, German Credit, and Adult Income corresponds to stronger causal constraints and higher effort for change, limiting the variety of feasible CFEs. This further supports the interdependence of objectives observed in Section 5.2.1 and Fig. 5.2.

Table 5.7: Generated CFEs from the German Credit Dataset

Age	**Sex**	**Job**	**Housing**	**Saving Account**	**Checking Account**	**Credit Amount**	**Duration**	**Purpose**	**Risk**
35	Male	2.0000	Free	Little	Little	3386.00	12	Car	Bad
35	Male	1.8634	Free	Little	Little	3759.86	21.50	Business	**Good**
35	Male	1.5721	Free	Little	Little	3759.86	21.15	Car	**Good**
35	Male	2.6625	Free	Little	Little	3759.86	21.28	Domestic Appliances	**Good**

All columns represent features of the German Credit Dataset [121]. The values represent the original data and the corresponding CFEs generated.

5.2.4 Quantitative comparison with existing works

To evaluate the efficacy of the framework, we compared its performance with three existing frameworks - DiCE [123], MOC [55], and CARE [30]. These frameworks were selected because they align with our framework's central objective of producing actionable CFEs. DiCE employs a custom genetic algorithm to generate diverse CFEs by optimizing proximity, diversity, and validity through a weighted loss function. For our comparison,

we additionally integrated plausibility into DiCE. MOC generates diverse CFEs by jointly optimizing validity, plausibility, proximity, and sparsity using the NSGA-II optimization algorithm. CARE produces coherent and actionable CFEs by optimizing sparsity, proximity, connectedness, coherency, actionability, and diversity using the NSGA-III. For consistency with our framework, CARE's outcome, proximity, and actionability terms were aligned with validity, proximity, and plausibility. Across all frameworks, we incorporated our effort and recourse-efficiency objectives to enable uniform comparison. Effort was included because it directly captures sparsity (the number of feature changes), aligning with objectives already considered in these frameworks.

The results of the comparative analysis (Table 5.8) demonstrate that all frameworks, including our proposed framework, successfully achieve $\theta_v(x') = 0$, confirming that the optimization consistently finds perturbations that alter the model's prediction. The CARE framework yields high effort values ($\theta_e(x')$ = 30 for Adult Income and 20 for German Credit) that arise when a substantial portion of features (at least 25%) are immutable, limiting the actionable space and forcing the framework to rely on more extensive changes to achieve the desired outcome. Although these CFEs achieve mathematically efficient recourse ($\theta_r(x')$ = 0.139 and 0.007), the required adjustments often have limited downstream impact and violate causal constraints, as reflected by high $\theta_c(x')$ values (5.207 and 4.227). This indicates that the CARE framework, while efficient in a numeric sense, may generate CFEs that are unrealistic or difficult to implement in practice. Conversely, for the MovieLens dataset, where fewer immutable constraints exist, the framework produces highly plausible CFEs ($\theta_c(x')$ = 1.149) with low $\theta_r(x')$ (0.044), demonstrating that when causal flexibility exists, the framework can generate feasible and effective CFEs. However, these CFEs still require substantial overall changes (as seen by $\theta_p(x')$ = 3.914 and $\theta_e(x')$ = 2), which could make them aspirational rather than fully actionable for a user. Overall, these results highlight

a critical trade-off in CARE, that while it can optimize for mathematical efficiency and plausibility, practical actionability may be compromised when many constraints exist. This underscores the importance of balancing efficiency with causal realism and user feasibility.

Table 5.8: Quantitative Comparative Analysis with Existing Works

Dataset	**Methods**	$\theta_v(x')$	$\theta_p(x')$	$\theta_c(x')$	$\theta_e(x')$	$\theta_r(x')$
Adult Income [118]	DiCE	0.000	3.057	2.062	7.000	0.423
	MOC	0.000	3.038	3.216	**4.000**	1.086
	CARE	0.000	4.277	5.207	30.00	**0.139**
	Our Work	**0.000**	**2.997**	**1.818**	5.000	1.229
HELOC [122]	DiCE	0.000	2.642	4.029	22.00	**0.105**
	MOC	0.000	3.593	1.063	**21.00**	0.555
	CARE	0.000	**0.976**	2.249	23.00	0.196
	Our Work	**0.000**	1.849	**0.000**	23.00	0.259
MovieLens [120]	DiCE	0.000	3.745	**1.091**	22.00	0.155
	MOC	0.000	**2.473**	1.964	21.00	0.365
	CARE	0.000	3.914	1.149	22.00	**0.044**
	Our Work	**0.000**	2.537	4.260	**18.00**	0.363
German Credit [121]	DiCE	0.000	2.320	**1.120**	7.000	0.239
	MOC	0.000	3.300	2.316	**3.000**	1.472
	CARE	0.000	2.430	4.227	20.00	**0.007**
	Our Work	**0.000**	**2.310**	1.953	5.000	0.664
Default Credit Card [119]	DiCE	0.000	2.885	3.917	19.00	0.151
	MOC	0.000	5.829	2.448	20.00	0.988
	CARE	0.000	1.101	5.816	19.00	0.970
	Our Work	**0.000**	**1.007**	**2.130**	**18.00**	**0.151**

$\theta_v(x')$, $\theta_p(x')$, $\theta_c(x')$, $\theta_e(x')$, $\theta_r(x')$ are validity, proximity, plausibility, effort, and recourse efficiency.

Additionally, the values of $\theta_p(x')$ are often moderate to high (Table 5.4), reflecting noticeable deviations from the original instance, x, as DiCE does not explicitly minimize perturbations. $\theta_c(x')$ is also moderate, indicating that while some causal relationships are preserved, CFEs may not fully align with the data manifold. $\theta_e(x')$ values are feasible but not optimized, resulting in adjustments that are actionable but not always minimal. Consequently, $\theta_r(x')$ is low to moderate (0.105 - 0.423), showing that the ratio of user $\theta_e(x')$ to causal impact is not always ideal. Overall, DiCE produces valid CFEs that are moderately plausible and practical, but its lack of explicit multi-objective optimization leads to trade-offs between efficiency, realism, and proximity, making the CFEs less balanced compared to our proposed framework.

Furthermore, the MOC framework generally performs well in minimizing $\theta_e(x')$ across datasets, except for MovieLens and Default Credit Card, reflecting its design to reduce the magnitude of individual feature changes. However, this low per-feature effort often comes at the cost of higher $\theta_p(x')$ and, in some cases, higher $\theta_c(x')$, because the cumulative effect of multiple small changes can result in CFEs that substantially deviate from the original instance or violate causal constraints. As a result, although the required effort per feature is small, the generated CFEs may appear unrealistic or infeasible, reducing their practical actionability. For example, in the Default Credit Card, both our proposed framework and DiCE achieve a high $\theta_r(x')$ (0.151), yet plausibility is compromised ($\theta_c(x')$ = 3.917), and more adjustments are required ($\theta_p(x')$ = 2.885). This demonstrates that high $\theta_r(x')$ can be misleading if evaluated in isolation, as the recommended paths may not be actionable in practice. Therefore, evaluating counterfactuals jointly on $\theta_e(x')$, $\theta_p(x')$, and $\theta_c(x')$ is crucial to ensure that suggested changes are not only mathematically efficient but also realistic and implementable by users.

In the Adult Income, HELOC, and Default Credit Card, our proposed framework achieves a high $\theta_c(x')$, due to patterns described in Section 5.2.1. Low-to-moderate $\theta_p(x')$ in Adult Income, German Credit, and Default Credit Card datasets arises from sufficient feature-space flexibility, allowing minimal yet impactful changes. The $\theta_e(x')$ values further clarify these patterns, as seen in HELOC and MovieLens having higher $\theta_e(x')$ values because realistic causal changes in these domains require modifying multiple interdependent features. In contrast, Adult Income and Default Credit Card exhibit lower $\theta_e(x')$, as plausible recourse can be achieved with a few targeted changes. Overall, the results show that our proposed framework balances causal consistency, minimal changes, and actionable recourse, therefore, generating feasible CFEs across diverse datasets.

Across datasets, our proposed framework consistently balances trade-offs among objectives, generating more stable and practical CFEs compared to DiCE, MOC, and CARE. It achieves an average normalized score (AvgNorm) of 0.279, computed by first min-max normalizing the four minimized objectives ($\theta_p(x')$, $\theta_c(x')$, $\theta_e(x')$, and $\theta_r(x')$) per dataset (excluding the constant $\theta_v(x')$), averaging these normalized values to obtain a per-dataset AvgNorm, and finally averaging across all five datasets. This corresponds to an overall baseline improvement of approximately 38.8%, reflecting more efficient and realistic CFEs. These gains are primarily driven by consistent improvements in $\theta_p(x')$ and $\theta_c(x')$, ensuring that suggested changes are both minimal and causally aligned. Overall, these results show that our framework produces actionable CFEs that align with realistic expectations, enhancing user trust and practical applicability.

5.2.5 Qualitative Comparison with Existing Works

Quantitative evaluations and numerical metrics alone cannot show *how* CFEs behave or whether their suggested changes are realistic. Therefore, qualitative comparisons are

required to evaluate the practicality of CFEs in terms of interpretability and feature-level realism that numerical metrics may miss. In each framework, we qualitatively evaluate the CFEs by analyzing the kind of features being modified (actionable vs. immutable), realism of the suggested changes, alignment to causal relationships (e.g., whether interventions propagate logically), practicality of implementing the changes from a user's perspective, and the extent to which the CFEs reflect the defined objectives.

We used the German dataset for this evaluation because its features correspond to realistic factors considered for credit approval. It provides a diverse set of continuous and categorical features that allows for evaluating the robustness of our proposed framework against the baseline frameworks. The original instance, x, represents a 35-year-old male with free housing, low savings and checking accounts, a credit amount of 3,386 over a 12-month duration, employed at job level 2, and applying for a car loan, labeled as a high-risk applicant. From Table 5.9, we observe that the changes suggested by DiCE (x'_{DiCE}) violate semantic plausibility, as it suggests a decrease in age to get the loan approval ($x'_{\text{DiCE}} = 31.0$). This is invalid since age cannot be reduced, making the generated CFE unrealistic. Likewise, CARE (x'_{CARE}) suggests an increase in age ($x'_{\text{CARE}} = 36.6$), making the CFE realistic (as a user's age can increase), but less actionable for a user. While the progression of age is feasible, it is not a controllable variable, as user cannot intentionally modify their age to achieve their desired outcome. Instead, the user would merely age into that state over time, with no guarantee that the surrounding conditions or model inputs would remain unchanged enough to produce the desired credit outcome. This shows that though the CFEs from CARE [30] achieved a high recourse efficiency as indicated by $\theta_r(x')$ (0.007) in Table 5.8, it does not necessarily translate to actionable, feasible CFEs.

Additionally, the suggested changes from x'_{DiCE} are substantial and demonstrate limited causal awareness. For instance, the CFE suggests reducing the credit amount and duration,

Table 5.9: Qualitative Comparative Analysis of CFES with Existing Works

Data	Age	Sex	Job	Housing	Saving Account	Checking Account	Credit Amount	Duration	Purpose	Risk
x	35.0	Male	2.00	Free	Little	Little	3386.00	12.00	Car	Bad
x'_{DiCE}	31.0	Male	2.00	Own	Little	Little	276.00	6.00	Car	**Good**
x'_{MOC}	34.9	Male	2.00	Free	Little	Little	3399.70	11.99	Car	**Good**
x'_{CARE}	36.6	Male	1.87	Free	Little	Little	276.00	26.75	Radio/TV	**Good**
$x'_{\text{Our Work}}$	35.0	Male	2.66	Free	Little	Little	3759.86	21.28	Business	**Good**

All columns represent features of the German Credit Dataset [121]. x is the orignal data, x'_{DiCE}, x'_{MOC}, x'_{CARE}, and $x'_{\text{Our Work}}$ represents the counterfactuals generated from the DiCE, MOC, CARE and proposed framework, respectively.

while increasing the housing status and maintaining the job level unchanged. This implies an extreme intervention in features to achieve validity, but it ignores contextual constraints in the CFE generation process, leading to misleading suggestions. MOC (x'_{MOC}) and our proposed framework ($x'_{\text{Our Work}}$), in contrast, attain validity with the smallest and most causally coherent perturbations by minimally adjusting the age ($x'_{\text{MOC}} = 34.9$), credit amount ($x'_{\text{MOC}} = 3399.70$, $x'_{\text{Our Work}} = 3759.86$) and duration ($x'_{\text{MOC}} = 11.99$, $x'_{\text{Our Work}} = 21.28$). This implies that the CFEs respect real-world causal dependencies, as an increase in credit amount and loan duration is a plausible way to improve trustworthiness for a borrower, thus providing the user with actionable guidance.

Furthermore, $x'_{\text{Our Work}}$ retains the same age, sex, and housing as x, but increases job level to 2.66, extends loan duration to 21.28 months, and slightly increases credit amount to 3759.86, with the purpose changing to business. While some of these adjustments, such as increasing job level, may require time, they remain behaviorally actionable because users can deliberately work toward them through employment choices or skill development. This stands in contrast to age-based changes, as seen in x'_{CARE}, which occurs passively and cannot be directly controlled. Even if the time horizon is similar, age progression

cannot be intentionally directed, nor does it ensure that future circumstances will align to yield the desired outcome. Thus, the changes suggested by $x'_{\text{Our Work}}$ reflect realistically attainable user actions, resulting in a causally plausible and genuinely actionable form of recourse. It also emphasizes recourse efficiency by suggesting a cost-effective yet semantically valuable adjustment path. Overall, the x'_{MOC} and $x'_{\text{Our Work}}$ CFEs exemplify high-quality, interpretable recourse with causal and behavioral fidelity, while x'_{DiCE} and x'_{CARE} illustrate contrasting extremes of minimal constraint and structured causality.

CHAPTER SIX

RO IV: PERSONALIZED, PREFERENCE-AWARE RECOURSE

Explainable AI seeks to make ML decisions understandable to humans by providing explanations for why a model produced a particular outcome [132]. However, many widely used approaches, such as SHAP [113] and LIME [114], primarily describe model behavior rather than help users to understand how they might influence or respond to outcomes. To address this limitation, Counterfactual Explanations (CFEs) present hypothetical *what-if* scenarios that show users how minimal, feasible changes to input features could alter a predicted outcome [133]. As model-agnostic, post-hoc explanations, CFEs provide users with actionable guidance that enhances their ability to understand and act upon decisions [22]. An effective counterfactual demonstrates the properties of validity, proximity, plausibility, and sparsity [22]. Hence, they are widely applied in domains such as healthcare [134] and finance [135].

Most counterfactual generation frameworks [55, 115, 23, 53] often assume all suggested feature changes are equally feasible and acceptable to users. They produce technically valid counterfactuals misaligned with users' abilities, intentions, or preferences. However, a change that is feasible for one user may be challenging for another depending on their preferences and constraints. For example, counterfactuals generated on the German credit dataset [121] in [136] suggest reducing the credit amount from 3386 to 276 or altering the loan purpose from "car" to another category to obtain a favorable outcome. While these changes alter the model's prediction, they are largely infeasible in practice, as a loan of 276 may not satisfy the needs motivating the original request, and modifying the loan purpose assumes a degree of flexibility users typically do not possess. Such counterfactuals can diminish trust, signal indifference to individual circumstances, and

reinforce perceptions of algorithmic unfairness. Fundamentally, this reveals a misalignment between counterfactual optimization objectives and users' real-world costs, underscoring the need for user-preference-aware CFE generation.

Existing literature, such as Adaptive Feature Weight Genetic Explanations (AFWGE), adjusts feature weights dynamically during CFEs generation to prioritize features [137]. While adaptive, the weights reflect the model's priorities rather than the user's real-world priorities, and can therefore mislead users into thinking the suggested changes are feasible or desirable. The User Feedback-Based Counterfactual Explanations (UFCE) framework incorporates user-specified constraints (feature ranges) into counterfactual generation to refine the search on actionable feature subsets and improve the feasibility of the explanations [138]. However, UFCE relies on static, preset constraints and cannot infer or prioritize features based on individual user intent, limiting its adaptability and personalization. In [139], users' preferences are inferred from pairwise comparisons between candidate CFEs and are used to iteratively guide their generation. However, the need for repeated user interaction increases both time costs and cognitive load. The Personalized Algorithmic Recourse (PEAR) framework uses Bayesian Preference Elicitation to infer individualized feature cost preferences, which are then integrated into a reinforcement learning framework with Monte Carlo Tree Search to efficiently generate personalized recourse plans [26]. However, the effectiveness of the Bayesian model depends on the prior distribution over user costs, and if the prior is poorly specified, the resulting recourse plans may still misalign with true user preferences.

To address these shortcomings, we propose a lexicographic counterfactual generation framework that integrates user-elicited preferences and causality in relationships to generate realistic, personalized, actionable CFEs *without* requiring repeated user interaction. Users' interaction with CFEs is based on perceived effort and personal constraints, not abstract

mathematical distances. Therefore, changes that appear minor to a model may obscure psychological, social, or practical costs, misaligning counterfactuals with real user effort. The proposed framework explicitly elicits feature-level preferences (importance rankings, constraint types) from users, ensuring that these preferences are incorporated as structured constraints during counterfactual generation. Valid CFEs do not guarantee that the proposed changes are feasible or actionable for users. As a result, generating actionable CFEs requires balancing user effort through proximity and sparsity, enforcing real-world and causal barriers via plausibility, and aligning recommendations with individual preference satisfaction. Also, the relative importance of these requirements varies across users and contexts, necessitating adaptive prioritization rather than uniform treatment. This motivates a lexicographic structure in which the objective ordering reflects user-specific preferences and constraints, so that lower-priority gains do not compensate for higher-priority violations. Therefore, we design our framework as a tiered recursive multi-objective optimization problem over proximity, sparsity, plausibility, and preference adherence using the NSGA-III algorithm. The main contributions of this research are:

- Design a lexicographic objective hierarchy that encodes user-specified priority tiers over counterfactual objectives, enabling sequential optimization that ensures feasible, actionable CFEs aligned with user preferences

- Validate the proposed framework's outcome using three tabular datasets (Adult Income [118], German Credit [121], and Student Performance [140]), and compare against existing frameworks of UP-AR [61], PEAR [26] and HARE [32] using performance metrics such as validity, plausibility, proximity, sparsity and preference satisfaction rate.

6.1 Proposed Framework

The proposed user-centric counterfactual generation framework is a system pipeline of interconnected modules that convert user-specified preferences into feasible, personalized actions, as shown in Fig 6.1. Elicited preferences are first validated against a causal model to eliminate infeasible interventions. Counterfactuals are then generated through lexicographic optimization, with validity and proximity enforced as primary constraints while optimizing user cost, plausibility, and sparsity, simultaneously. The application of instance-level adaptive proximity thresholds ensures that recommendations remain realistic and tailored to individual effort. Table 6.1 and Table 6.2 summarize the primary mathematical notation

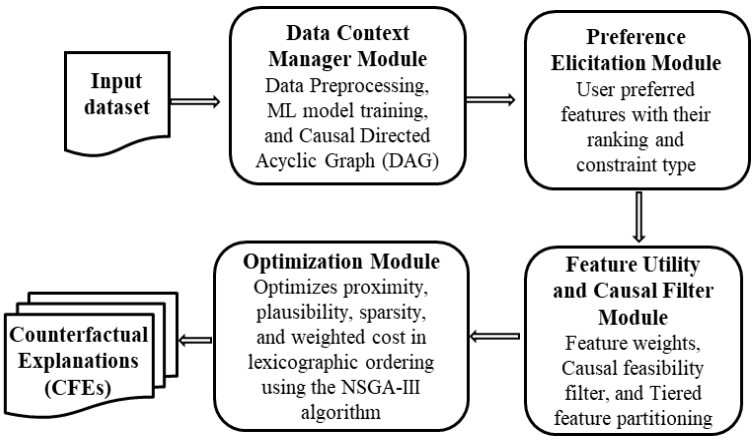


Figure 6.1: Proposed User-Centric Counterfactual Generation Framework

used throughout this chapter.

6.1.1 Data Context Manager Module

The quality of the input data context directly determines the feasibility of generated CFEs, as users' preferences and constraints are defined over the semantic structure of the

Table 6.1: Data and Feature Notation Used in RO IV

Symbol	Description
$\mathscr{F}$	Complete feature set describing the dataset.
$\mathscr{F}_A$	Set of actionable features that users can modify.
$\mathscr{F}_I$	Set of immutable features that cannot be modified.
$\mathscr{G}$	Directed Acyclic Graph (DAG) capturing causal dependencies among features.
x	Original input instance representing the user's current state.
x'	Generated counterfactual instance representing recommended changes.

dataset. This module serves as a domain-aware abstraction layer that transforms raw feature values into semantically meaningful representations, ensuring that data cleaning, encoding, and normalization preserve the real-world interpretation of features over which users express preferences.The processed data are then used to train the predictive model, providing a consistent and unbiased foundation for downstream counterfactual generation.

Additionally, features $\mathscr{F}$ are categorized into actionable ($\mathscr{F}_A$) and immutable ($\mathscr{F}_I$) sets to ensure counterfactual recommendations remain ethically and practically feasible. $\mathscr{F}_A$ includes features a user can realistically change to improve their outcome, while $\mathscr{F}_I$ consists of features that cannot be altered due to biological, legal, or ethical constraints. To further restrict the search space to realistic interventions, a causal Directed Acyclic Graph (DAG) $\mathscr{G}$ is incorporated to encode structural dependencies among variables. This bounds CFE generation to respect causal relationships while staying aligned with user-specified choices. By enforcing feasibility and realism prior to preference elicitation, this module ensures users interact with a stable and interpretable decision space, serving as a single source of truth for feature semantics, constraints, and causal structure throughout the pipeline. Feature constraints and $\mathscr{G}$ thus define hard structural boundaries that delimit the feasible

Table 6.2: User Preferences and Optimization Objectives Notation (RO IV)

Symbol	Description
$\mathscr{U}_{\text{pref}}$	Set of user-specified feature preferences guiding counterfactual generation.
τ	Actionable feature selected by the user for potential modification.
ρ	Ordinal rank representing the relative importance of modifying feature τ.
ψ	Boolean variable indicating whether a preference is a hard constraint (ψ_h) or soft constraint (ψ_s).
$\mathscr{P}_u$	Set of actionable features selected by the user during preference elicitation.
$\mathscr{P}_{\text{valid}}$	Subset of user preferences validated as causally feasible.
α_τ	Preference-derived penalty weight assigned to feature τ.
$\rho_{\max}$	Maximum rank value among user preferences.
λ	Global scaling factor controlling preference enforcement strength.
$\mathscr{T}_1$	Tier 1 feature set containing validated user-preferred features.
$\mathscr{T}_2$	Tier 2 feature set containing causally downstream actionable features.
$\theta_p(x')$	Proximity objective measuring distance between the counterfactual and original instance.
$\theta_r(x')$	Plausibility objective measuring causal realism of the counterfactual.
$\theta_s(x')$	Sparsity objective measuring the number of modified features.
$\theta_{wc}(x')$	Weighted user cost objective incorporating preference ranks.
Γ_1	Primary feasibility constraint set enforcing validity and mandatory preferences.
Γ_2	Secondary objective set optimized after feasibility constraints are satisfied.
ε	Proximity threshold limiting the allowable perturbation from the original instance.
δ	Stability parameter controlling threshold selection in dynamic proximity tuning.
N_{valid}	Number of feasible counterfactuals satisfying primary constraints.

decision space, ensuring users can express preferences only over actionable and causally valid dimensions

6.1.2 Preference Elicitation Module

This module serves as an interactive interface between the user and the counterfactual generation pipeline for eliciting user-specified priorities to enable transparent and traceable personalization. It allows users to express their intent and priorities, which directly guide the counterfactual generation process, while ensuring preferences are stated over a stable, interpretable, and feasible decision space. Unlike traditional methods that assume uniform feature costs, this module enforces a hierarchical preference structure, which is essential for user-centric optimization.

User preferences ($\mathscr{U}_{\text{pref}}$) are represented through a structured, feature-level schema that encodes both preference expression and its influence on downstream CFE generation, as shown in eq. (6.1).

$$\mathscr{U}_{\text{pref}} = \{(\tau, \rho, \psi) \mid \tau \in \mathscr{F}_A, \rho \in \mathbb{Z}^+, \psi \in \{0,1\}\} \tag{6.1}$$

where τ denotes an actionable feature a user wishes to modify. $\tau \in \mathscr{F}_A$ ensures users express preference over mutable variables, preventing infeasible recommendations. ρ is an ordinal importance rank that indicates the relative priority of a τ, with lower ranks indicating higher importance. This induces a lexicographic hierarchy that enables the optimization module to distinguish primary objectives from secondary trade-offs. ψ is a boolean flag specifying whether the preference is a hard (ψ_h, 1 = mandatory change) or a soft constraint (ψ_s, 0 = desirable but negotiable). We denote ψ_h as the subset of user preferences that are actively enforced, induced by the binary activation variables, ψ_τ ($\psi_h = \{\tau \in \mathscr{U}_{\text{pref}} \mid \psi_\tau = 1\}$). This distinction allows the framework to model real-world user constraints, where certain

requirements must be strictly satisfied while others may be relaxed in favor of higher-priority objectives. This explicit, structured representation ensures that preferences are interpretable and auditable for the optimization module.

User interaction during elicitation is actively guided by a formal protocol ($\mathscr{P}_u$) in eq. (6.2).

$$\mathscr{P}_u \subseteq \mathscr{F}_A \text{ s.t. } \mathbf{x} = \eta,\ k_{\min} \leq |\mathscr{P}_u| \leq k_{\max} \tag{6.2}$$

To ensure realistic recommendations, user preferences are elicited based on their current instance ($\mathbf{x} = \eta$), using their actual feature values as a reference. The set of actionable features identified by the data context manager explicitly restricts the preference selection to features that can be feasibly modified. To balance expressiveness with cognitive effort, the $\mathscr{P}_u$ constrains users to specify a bounded number of preferred features (where $k_{\min}$ = three and $k_{\max}$ = five). The range of three to five features was selected based on evidence that users can reliably consider and prioritize a small set of options without cognitive overload [141]. This controlled elicitation strategy encourages users to meaningfully prioritize features rather than indiscriminately expressing preferences, while still allowing sufficient flexibility to capture individual contexts. This also prevents the optimization problem from becoming over-determined (too many constraints) or under-specified (too vague). The preferences are expressed over original feature semantics to preserve alignment with how users conceptualize their own decision-making context. The module outputs a ranked, machine-readable representation of $\mathscr{U}_{\text{pref}}$ that cleanly separates user intent from optimization-relevant constraints. This ensures a principled and reproducible optimization process while preserving interpretability. This ranked list of desired changes is then passed to the feature utility module for transformation and causal consistency.

6.1.3 Feature Utility and Causal Filter Module

This module bridges user intent and optimization by transforming the elicited $\mathscr{U}_{\text{pref}}$ into a quantitative, feature-conditioned utility function, organizing features into prioritized tiers while enforcing causal feasibility, thereby enabling direct optimization that remains aligned with user intent. For each preferred feature $\tau \in \mathscr{U}_{\text{pref}}$, a feature-specific penalty weight α_τ is derived from its user-assigned rank ρ via inverse normalization (eq. (6.3)), ensuring that higher-priority features incur lower penalties during counterfactual optimization.

$$\hat{\rho}_\tau = 1 - \frac{\rho_\tau - 1}{\rho_{\max}}, \qquad \alpha_\tau = \lambda \cdot (1 - \hat{\rho}_\tau) \tag{6.3}$$

where ρ_τ denotes the user-assigned rank of feature τ, $\rho_{\max}$ is the maximum rank across all features in $\mathscr{U}_{\text{pref}}$, and λ is a global scaling factor controlling the overall strength of preference enforcement. This formulation allows the framework to balance user preference satisfaction against other optimization objectives. By assigning a lower α_τ to higher-ranked features, the framework prioritizes changes that matter most to the user while penalizing modifications to less important features. $\mathscr{U}_{\text{pref}}$ is then encoded in an optimization-ready form that will be validated for causal feasibility before the counterfactual optimization step.

6.1.3.1 Causal Feasibility Filter: This filter validates the elicited $\tau \in \mathscr{U}_{\text{pref}}$ against the causal graph, $\mathscr{G}$, ensuring all requested changes are causally feasible. It also avoids unnecessary computation on infeasible 'dead-end' solutions, where requested changes to $\mathscr{F}_A$ cannot occur without modifying their causal parents. A $\mathscr{U}_{\text{pref}}$ is validated against the structural dependencies encoded in $\mathscr{G}$ by enforcing a direct parental consistency check, defined in eq. (6.4).

$$\mathscr{P}_{valid} = \{(\tau, \rho, \psi) \in \mathscr{U}_{\text{pref}} \mid \forall \pi \in \text{Pa}(\tau)_{\mathscr{G}}, \text{Consistent}(\tau, \pi)\} \tag{6.4}$$

Where π is a direct causal parent (($\text{Pa}(\tau)_{\mathscr{G}}$)) of feature τ, and $Consistent(\tau, \pi)$ indicates the value of τ respects its causal relationship with π. The causal filter systematically examines the immediate causal neighborhood of each τ to enforce two logical constraints: immutable parent constraint and interventional independence. Under the immutable parent constraint, if a requested change to τ would require modifying an immutable parent ($\tau \in \mathscr{F}_I$), the request is rejected as physically infeasible. Under interventional independence, actionable features ($\tau \in \mathscr{F}_A$) are modeled as external interventions, allowing them to vary independently of immutable parents. This ensures that the proposed framework distinguishes between interventions (which override causal parents) and dependent changes (which must respect parental constraints). For example, while a user cannot directly change a dependent variable like risk score if it strictly relies on the immutable parent age, they can intervene on an actionable variable like income, effectively breaking the standard causal dependency. By isolating physically realizable interventions from structural violations, the module ensures the optimization process focuses exclusively on actionable counterfactual scenarios. The filter outputs a refined set of validated preferences, $\mathscr{P}_{\text{valid}}$, which is an input to the tiered feature partitioning module, ensuring that generated CFEs remain both desirable and causally feasible.

6.1.3.2 Tiered Feature Partitioning: To prioritize user-desired changes and organize the remaining actionable features ($\mathscr{F}_A$) according to causal dependencies in $\mathscr{G}$, this module partitions features into two tiers. Tier 1 ($\mathscr{T}_1$), enforcing primary user goals, comprises user-preferred features from $\mathscr{P}_{\text{valid}}$ that are both actionable and causally feasible, depicting the highest-priority changes to satisfy user intent. Tier 2 ($\mathscr{T}_2$) includes $\mathscr{F}_A$ that are causally downstream of $\mathscr{T}_1$ features in $\mathscr{G}$ but were not explicitly selected by the user (eq. 6.5). These features provide indirect interventions that expand the solution space while maintaining

causal consistency.

$$\mathscr{T}_2 = \{f \in \mathscr{F}_A \mid \exists \tau \in \mathscr{T}_1, f \in \text{Descendants}(\tau, \mathscr{G})\} \setminus \mathscr{T}_1 \tag{6.5}$$

$\mathscr{T}_2$ features allow secondary interventions and are filtered to preserve structural constraints. Hard constraints (ψ_h) enforce mandatory $\mathscr{T}_1$ features, while soft constraints (ψ_s) provide weighted guidance for flexible, preferred changes (eq. 6.1). By organizing features into tiers, the module balances user intent with causal feasibility, capturing both direct and indirect pathways through a set of validated $\mathscr{T}_1$ and $\mathscr{T}_2$ that inform the lexicographic optimization hierarchy in the optimization module.

6.1.4 Optimization Module

CFEs' generation requires optimizing multiple objectives to balance feasibility, realism, user effort, and preference satisfaction. We formalize these objectives as follows.

Proximity ($\theta_p(x')$): captures the effort required from a user to achieve the desired outcome by measuring how close the CFE, x', is to the original instance, x, using the Euclidean distance (L2 norm) [127], as shown in eq.(6.6).

$$\theta_p(x') = \|x' - x\|_2 \tag{6.6}$$

Lower values indicate smaller, easier changes, making the counterfactual more practical for the user.

Plausibility ($\theta_r(x')$): evaluates a CFE's realistic implementation by comparing it against the conditional Gaussian distributions defined by the causal model, $\mathscr{G}$ (eq.(6.7))

$$\theta_r(x') = \sum_{\tau \in \mathscr{F}_\text{A}} \frac{\left(x'_\tau - \mu_\tau(\text{Pa}(\tau))\right)^2}{2\sigma_\tau^2} \tag{6.7}$$

Where μ_f and σ represent the mean and variance of the conditional Gaussian distribution for feature τ. Lower values indicate higher causal plausibility, meaning the counterfactual aligns with feasible, realistic changes.

Sparsity ($\theta_s(x')$): counts the number of features changed that a user can choose to modify to achieve the target, reducing cognitive or practical effort, as shown in eq. (6.8).

$$\theta_s(x') = \sum_{\tau \in \mathscr{F}_{\mathrm{A}}} \mathbb{I}\left(x'_\tau \neq x_\tau\right) \tag{6.8}$$

Where $\mathbb{I}(\cdot)$ is the indicator function that evaluates to 1 if feature j has been modified and 0 otherwise.

Weighted User Cost ($\theta_{wc}(x')$): incorporates user preference ranks (ρ) to assign penalty weights, w_τ, incentivizing desired changes while discouraging unnecessary modifications (eq. 6.9).

$$\begin{aligned} \theta_{wc}(x') &= \sum_{\tau \in \mathscr{F}_{\mathrm{A}}} w_\tau \mathbb{I}\left(x'_\tau \neq x_\tau\right), \\ w_\tau &= \begin{cases} \alpha_\tau + \varepsilon, & \tau \in \mathscr{P}_{\text{valid}} \\ \lambda_{\max}, & \tau \notin \mathscr{P}_{\text{valid}} \end{cases} \end{aligned} \tag{6.9}$$

where ε is the proximity threshold. Preferred features, τ, incur lower penalties (w_τ) to encourage their change, while unranked features in $\mathscr{F}_A$ receive maximal weights, inhibiting modifications unless necessary. Although these objectives collectively define desirable CFEs, users often prioritize certain requirements as non-negotiable. To respect these hierarchies and prevent undesirable trade-offs, we enforce primary objectives and constraints first, followed by secondary user-preference criteria using a lexicographic ordering.

6.1.4.1 Lexicographic Ordering The counterfactual search is formulated as a hierarchically constrained optimization problem, with objectives (θ) partitioned into two priority levels, Γ_1

and Γ_2, such that $\Gamma_1 \prec_{lex} \Gamma_2$. Any solution with a lower violation in Γ_1 is strictly preferred, irrespective of its performance in Γ_2.

In Γ_1, a lexicographically dominant feasibility constraint ensures that a counterfactual, x', achieves the target outcome, respects ε, and satisfies mandatory feature modifications (ψ_h), thereby acting as a gatekeeper for the search space (eq. 6.10). This ensures users are only given realistic, attainable recommendations.

$$\begin{aligned}\Gamma_1(x') = & \underbrace{\mathbb{I}\big(f(x') \neq y_{\text{target}}\big)}_{\text{Validity}} \\ & + \underbrace{\max\big(0, \theta_p(x') - \varepsilon\big)}_{\text{Proximity Threshold}} \\ & + \underbrace{\sum_{\tau \in \psi_h} \mathbb{I}\left(x'_\tau \notin \mathscr{V}_\tau\right)}_{\text{Hard Constraints}}\end{aligned} \tag{6.10}$$

Here, ε bounds the allowable proximity $\theta_p(x')$, while $\mathscr{V}_j$ defines the allowed values for mandatory features. $\theta_p(x')$ is both minimized as an objective and capped by ε, enabling fine-grained optimization within a feasible neighborhood. Infeasible solutions incur a large penalty (λ), ensuring lexicographic priority over secondary objectives: weighted user cost ($\theta_{wc}(x')$), plausibility ($\theta_r(x')$), and sparsity ($\theta_s(x')$).

However, a single global ε is insufficient because feasibility and effort vary across users. We therefore adopt dynamic, instance-level thresholds (eq.(6.11)), which adaptively identify feasible proximity regimes per instance using local sensitivity analysis. Given a grid of candidate proximity thresholds $\{\varepsilon_k\}_{k=1}^{K}$, stable lexicographic structures are identified by selecting thresholds that yield both a sufficient number of feasible counterfactuals and a

stable secondary objective behavior. A threshold ε_k is considered stable if:

$$\varepsilon_k \in \mathscr{E}_{\text{stable}} \iff \begin{cases} N_{\text{valid}}(\varepsilon_k) \geq N_{\min}, \\ |\theta_r(\varepsilon_k) - \theta_r(\varepsilon_{k-1})| < \delta, \end{cases} \tag{6.11}$$

where $N_{\text{valid}}(\varepsilon_k)$ denotes the number of counterfactuals satisfying the Tier-1 (Γ_1) constraints at threshold ε_k, $\theta_r(\varepsilon_k)$ is the average plausibility score over valid solutions, $N_{\min}$ is minimum feasibility requirement, and δ controls stability across successive thresholds. The stability heuristic, δ, ensures that ε is relaxed just enough to allow the optimizer to converge on a diverse Pareto front, but no further, thereby preventing the generation of unnecessarily distant or unrealistic counterfactuals. Overall, ε prevents premature infeasibility while preserving strict priority over secondary objectives. We solve the resulting lexicographically constrained multi-objective problem (eq. 6.12) using NSGA-III [128]. Feasible solutions satisfying Γ_1 are first selected, after which NSGA-III jointly optimizes the secondary objectives in Γ_2 using reference directions. We adopt NSGA-III for its reference-point-based selection, which naturally produces diverse and interpretable counterfactual recommendations that align with different user priorities in high-dimensional constrained spaces.

$$\begin{aligned} &\text{Stage 1:} \quad \min\ \Gamma_1(x') \\ &\text{Stage 2:} \quad \min\ \Gamma_2(x') = \big(\theta_p(x'),\, \theta_r(x'),\, \theta_s(x'),\, \theta_{wc}(x')\big) \\ &\qquad\qquad \text{s.t. } x' \in \Gamma_1(x') \end{aligned} \tag{6.12}$$

6.2 Simulation and Results

The objective of our simulation is to demonstrate the effectiveness of the proposed framework for generating personalized, actionable CFEs. Simulations were done on a single workstation with 16GB RAM, with a 13th Gen Intel(R) i9 2.60 GHz CPU, implemented using Python 3.13. The details of the codebase for this simulation can be found in [107].

Dataset and Model Training: Three tabular datasets were selected because they model individual-level outcomes, such as credit risk, academic performance, and income, using rich demographic, social, and contextual attributes (Table 6.3). German Credit [121] classifies individuals as good or bad credit risks based on financial creditworthiness. Adult Income [118] predicts whether an individual's income exceeds $50K/year based on demographic and employment attributes. Student Performance [140] predicts students' final grades and academic outcomes based on demographic, social, and educational attributes. This structure supports personalized, user-centric counterfactual evaluation, with features preprocessed and split into immutable (e.g., age, sex, race) and actionable sets. Three black box models, namely, Logistic Regression, SVM, and Random Forest, were chosen for their clear decision boundaries, making them suitable for counterfactual explanations, and were trained using a standardized 70:30 split with isotonic calibration. Evaluating these models is not the main scope of this work.

Table 6.3: Datasets used in simulations

Dataset	**Feat.**	**N**	**Label**
German Credit	10	1000	(Bad, Good)
Student Performance	33	395	(Pass, Fail)
Adult Income	14	30162	($\leq$ 50K, $\geq$50K)

Causal Graph Integration: The integration of causality in a multi-objective framework was extensively discussed in [136]. We leveraged the Directed Acyclic Graph (DAG) with the Structural Equation Model (SEM) utilized in the proposed model-agnostic counterfactual generation framework.

User Proxy: We leveraged a GPT-based LLM agent (GPT-4o-mini model) to act as a context-aware reasoning agent for eliciting user preferences. Acting as a "user proxy", the LLM simulates human decision-making by identifying the features it would prefer to change to achieve a desired outcome within the proposed framework. To evaluate robustness and adaptability, we simulate diverse user profiles representing heterogeneous preference structures (Table 6.4), preventing the optimization from overfitting to a single user type and enabling systematic assessment across varying intents. The personas are as follows: stability-oriented (favor low-disruption changes), income-maximizing (accept higher effort for financial gains), and effort-minimizing (prioritize minimal, incremental adjustments). In each simulation run, the agent adopts a distinct persona to guide preference elicitation. It ingests the user's factual profile, causal constraints, and persona to determine which features are logically and psychologically mutable. Finally, it outputs a ranked list of preferred feature changes with accompanying rationales, which are quantitatively mapped to optimization penalty weights (α) to steer the evolutionary search (Section 6.1.3).

Table 6.4: Simulated User Personas and Optimization Preferences

Persona	Characteristics	Impact on Optimization	Temp.
Stability Oriented	Favors preserving the employment and demographic status quo.	Conservatively constrains demographic and employment features, reducing the feasible CFE space.	0.5
Income Maximizing	Seeks financial maximization and accepts high effort when beneficial.	Reduces penalties on high-effort features, permitting higher-cost, higher-utility CFEs.	0.7
Effort Minimizing	Prioritizes low-resistance changes and avoids structural shifts.	Favors small, incremental changes to continuous features over categorical changes.	0.9

Algorithmic Parameters and Performance Metrics: The NSGA-III parameters, such as population size, number of generations, mutation rate, crossover probability, and reference partitions, were empirically set to 150, 100, 10, 0.9, and 6, respectively. Proposed framework performance is evaluated by measuring how well the generated CFEs satisfy the optimization objectives defined in Section 6.1.4, which directly capture user-relevant quality criteria. To quantify alignment between a generated CFE and a user's stated preferences, we further introduce a rank-aware evaluation metric, the Preference Satisfaction Ratio (PSR), shown in eq.(6.13). Unlike sparsity, which assigns equal weight to all feature changes, PSR weights each change by its user-specified importance, thereby avoiding false successes in which low-priority preferences are satisfied while higher-priority ones are ignored. By explicitly encoding preference ordering and non-linear importance, PSR provides a more faithful metric of user-centric utility.

$$\text{PSR} = \frac{\sum_{\tau \in M} \hat{\rho}_\tau}{\sum_{\tau \in \mathscr{P}_{\text{valid}}} \hat{\rho}_\tau} \times 100\% \tag{6.13}$$

where $\mathscr{P}_{\text{valid}}$ denotes the set of validated user-preferred features, $M \subseteq \mathscr{P}_{\text{valid}}$ is the subset of preferred features modified in the generated CFE, and $\hat{\rho}_\tau$ is the normalized importance score of feature τ, defined in eq.(6.3).

6.2.1 Quantitative Evaluation Using CFE Objectives

To assess the impact of user preferences on the quality of generated CFEs, we compare our proposed user-aware framework with the baseline approach in [136], which does not incorporate user preferences, thereby isolating the effect of user-aware constraints.

Table 6.5 demonstrates that incorporating user constraints substantially reduces the effort (θ_s) required to achieve desired outcomes. By allowing users to select and rank actionable features, the proposed framework prioritizes changes that are both influential and feasible,

resulting in targeted and practical CFEs. The dramatic reduction in proximity for German Credit ($\theta_p = 0.004$ vs 1.06) highlights that user-aware constraints guide the optimization to remain close to the original instance, minimizing disruptive changes. Moreover, setting proximity as a threshold ensures that the optimizer consistently favors feasible, low-effort interventions before addressing secondary objectives, reflecting a deliberate user-centric prioritization in the search process. Overall, these results confirm that integrating user preferences not only improves alignment of CFEs with individual intent but also enhances efficiency, realism, and actionable relevance.

The quantitative evaluation across datasets illustrates how integrating user preferences impacts CFEs generation under different complexity scenarios. For the Adult Income, the proposed framework achieves perfect preference efficiency by modifying a single rank-1 feature ($\rho_\tau = 1; \theta_s = 1$, $\theta_{wc} = 1$), yielding a PSR of 40% that is optimal in this setting, as the prediction is flipped through one high-impact, user-preferred change with no penalty from lower-ranked features. In German Credit, the proposed framework demonstrates a high preference alignment by producing a CFE with $\theta_s = 2.00$ and $\theta_{wc} = 2.25$, where the small excess cost reflects inclusion of the second-ranked feature while fully preserving the top priority, resulting in strong alignment (PSR = 70%) with minimal user effort. In contrast, the Student Performance dataset reveals an unavoidable trade-off: although the framework attains high preference satisfaction (PSR = 73%), doing so requires substantial additional changes ($\theta_s = 10.0$, $\theta_{wc} = 23.4$), reflecting the incurred user cost needed to cross the decision boundary and demonstrating that the framework prioritizes what users care about most, even when preference-aligned recourse is mathematically costly. The weighted cost (θ_{wc}) provides a transparent indicator that validates the lexicographic optimization scheme (eq.(6.12)), showing that the proposed framework prioritizes top-ranked features while incurring higher penalties only when necessary to satisfy the decision boundary. The

baseline model, $x'_{\text{No User}}$, lacks θ_{wc} and PSR since it ignores user preferences. Integrating user preferences produces sparse, plausible, and cost-aware counterfactuals, while exhibiting trade-offs when the model's decision boundary limits full preference satisfaction.

Table 6.5: Quantitative Evaluation Using CFE Objectives

Dataset	Framework	θ_s	θ_p	θ_r	θ_{wc}	PSR
Adult Income	$x'_{\text{No User}}$	7.00	5.17	2.03	-	-
	$x'_{\text{Our Work}}$	**1.00**	**0.20**	**0.63**	**1.00**	40%
German Credit	$x'_{\text{No User}}$	4.00	1.06	2.50	-	-
	$x'_{\text{Our Work}}$	**2.00**	**0.004**	**0.78**	**2.25**	70%
Student Performance	$x'_{\text{No User}}$	24.0	21.10	6.26	-	-
	$x'_{\text{Our Work}}$	**10.00**	**2.34**	**2.62**	**23.4**	73%

$\theta_{(s,p,r,wc)}$ = sparsity, proximity, plausibility and weighted cost. Lower values are better.

6.2.2 Qualitative Analysis: German Credit Case Study

The qualitative evaluation of the proposed framework is conducted on a German Credit instance to show how it operationalizes high-level user preferences, leveraging the dataset's mix of realistic categorical and continuous features to evaluate its adaptability. The user profile, U_x, represents a 29-year-old male who owns his housing, has moderate checking and low savings accounts, a credit amount of 11,328 over a 24-month duration, employed at job level 3, and applying for a vacation/others loan, labeled by the predictive model as a high-risk applicant. The LLM user proxy, adopting a risk-aware persona, generated a hierarchical preference structure prioritizing credit amount (Rank 1) and duration (Rank 2) as hard constraints, followed by checking account (Rank 3) and savings account (Rank 4). By strictly enforcing this ranking and constraint type, the proposed framework assigned

Table 6.6: Comparison of Original Instance (x) and Generated CFE (x') for User (U_x)

Feature	**Rank**	**Constraint**	**Original Value** (x)	**CFE Value** (x')
Credit Amount	1	Hard	11,328 DM	*12,822 DM*
Duration	2	Hard	24 Months	*26 Months*
Checking Account	3	Soft	Moderate	Moderate
Savings Account	4	Soft	Little	Little

minimal optimization penalties to the top priorities ($\alpha = 0.0$ for Rank 1 and $\alpha = 0.25$ for Rank 2), ensuring that these critical 'hard' features were the primary levers for recourse. In response, the optimization module successfully identified a valid counterfactual that strictly adhered to this priority list.

As shown in the results in Table 6.6, the algorithm generated a solution that modified only the user's top two preferred features while leaving lower-priority attributes like checking account and saving account unchanged. The proposed framework suggested adjusting the Credit Amount from 11,328 DM to 12,822 DM and extending the Loan Duration from 24 months to 26 months. While the user's initial rationale hypothesized that reducing these values might be preferable, the predictive model's decision boundary required a calculated increase to achieve validity. By exactly targeting the Rank 1 and Rank 2 features, the framework achieved a high degree of preference satisfaction, handling the trade-off between the user's desired *strategic levers* and the model's mathematical requirements for recourse.

6.2.3 Comparative Analysis with Existing Works

To evaluate the efficacy of our framework, we compared its performance with three existing frameworks: UP-AR [61], PEAR [26], and HARE [32]. These frameworks were selected because they align with our framework's central objective of generating personalized CFEs. UP-AR encodes user preferences as soft constraints, ranked features, and bounded

ranges to guide gradient-based recourse. PEAR models individualized feature costs via Bayesian preference elicitation. HARE iteratively refines personalized CFEs by comparing pairwise explanations. Unlike these methods, our framework captures structured user intent in a single elicitation step while enforcing causal feasibility and lexicographic optimization.

The results of the comparative analysis (Table 6.7) demonstrate that HARE and PEAR systematically enforce multiple feature changes ($\theta_s \geq 2$) across all datasets because their uniform feature weighting allows the optimizer to modify whichever attributes are most convenient mathematically, regardless of their relevance to the user. Our framework instead prioritizes sparsity by explicitly aligning changes with user-ranked preferences, resulting in fewer but more meaningful interventions. On the Student Performance dataset, the high PSR (73%) confirms that these limited changes directly reflect user-specified priorities rather than incidental model artifacts, yielding counterfactuals that are both technically valid and practically actionable. Additionally, the HARE framework achieves the lowest plausibility scores (θ_r), producing CFEs that closely match statistically typical profiles, but the absence of causal modeling raises concerns about their practical feasibility. This is evident in the Adult Income dataset, where a very low θ_r (0.007) coincides with poor user alignment (PSR = 25%), indicating that statistically plausible solutions can conflict with individual preferences. However, our framework allows principled deviations from the statistical norm when supported by causal feasibility and user preferences, yielding more actionable and personalized recommendations.

Our framework yields a lower PSR (40%) by enforcing minimal user effort, requiring only one feature change ($\theta_s = 1$), whereas UP-AR attains a higher PSR (0.51) by inducing more changes ($\theta_s = 4$). This highlights a key trade-off between cognitive efficiency and aggregate preference satisfaction, as our work prioritizes satisfying the user's top-ranked preference over multiple lower-priority ones. This shows that changing a single high-priority

Table 6.7: Quantitative Comparative Analysis with Existing Works

Dataset	Framework	θ_s	θ_p	θ_r	PSR
Adult Income	Our Work	**1.00**	1.06	0.06	40%
	HARE	4.00	0.70	**0.007**	25%
	PEAR	4.00	1.56	7.22	29%
	UP-AR	4.00	**0.08**	1.80	**51%**
German Credit	Our Work	**2.00**	**0.004**	0.78	**70%**
	HARE	2.00	0.74	**0.19**	50%
	PEAR	4.00	1.79	3.23	61%
	UP-AR	3.00	0.22	1.49	68%
Student Performance	Our Work	10.0	2.34	2.62	**73%**
	HARE	13.0	2.73	**0.25**	25%
	PEAR	17.0	3.58	4.90	46%
	UP-AR	**2.00**	**0.06**	4.00	40%

$\theta_(s, p, r)$ = sparsity, proximity, plausibility. Lower value is better, except for PSR.

feature yields more actionable and practical recourse than requiring multiple low-impact modifications. On the Student Performance dataset, our framework incurs substantially higher proximity and sparsity costs ($\theta_p = 2.34$, $\theta_s = 10.0$) than UP-AR ($\theta_p = 0.06$, $\theta_s = 2.0$), yet achieves markedly higher preference satisfaction (PSR = 73% vs. 40%). This illustrates that satisfying user intent can be mathematically inefficient, requiring exploration deeper into the counterfactual space rather than following the nearest decision-boundary shortcut exploited by UP-AR. By intentionally accepting higher optimization cost on features users explicitly care about and avoiding those they reject, our framework pays a deliberate "user tax" to deliver substantively more meaningful and actionable recourse.

PEAR consistently incurs high plausibility costs ($\theta_r \geq 3$) across all datasets, reflecting overfitting to individualized subjective cost priors inferred through the Bayesian preference elicitation model. By aggressively minimizing personalized user costs (e.g., categorically avoiding job changes), PEAR is driven toward sparse, outlier regions of the data manifold

that satisfy user constraints but are statistically implausible. In contrast to HARE's manifold anchoring and our framework's causal feasibility filtering, PEAR allows user preferences to override statistical realism. Moreover, PEAR's Bayesian elicitation can be slow to converge on complex manifolds, resulting in efficiency degradation, as observed on German Credit ($\theta_s = 4$, $\theta_p = 1.79$). Our single-step user elicitation instead produces deterministic, structured preference weights upfront, providing a sharper optimization signal and notably improved efficiency. On the German Credit dataset, our framework achieves the lowest proximity (0.004), lowest sparsity (2.0), and highest PSR (0.70), delivering highly targeted suggestions that require minimal effort while satisfying user priorities. When user-preferred features align with the model's decision boundary, user-derived α weights act as a guiding heuristic, steering optimization efficiently toward the global optimum and avoiding irrelevant feature changes. Overall, our framework achieves the highest PSR in two datasets and the most efficient PSR in the third, demonstrating consistent user-aligned performance. Our proposed framework, utilizing min-max normalization for objectives and PSR scores, boosted user satisfaction

CHAPTER SEVEN

A UNIFIED TRUST-CENTRIC FRAMEWORK IN AI-DRIVEN SYSTEMS

The preceding chapters presented distinct but interrelated research objectives addressing multiple dimensions of trust in AI-driven service systems, each evaluated independently with methodological rigor and empirical validation. Beyond these contributions, the broader intellectual value lies in articulating an integrated framework that unifies infrastructure integrity, actionable reasoning, and user-aligned personalization within a cohesive system architecture. This chapter synthesizes the research objectives into a coherent conceptual model of integrated trust, consolidating the theoretical and architectural relationships across the proposed layers. It clarifies how the contributions form a vertically structured trust stack in which each layer enables and constrains the next.

The unified framework conceptualizes trust as a multi-dimensional construct emerging from the interaction of system-level guarantees, causal reasoning, and individualized user alignment (see Fig. 7.1). The framework is formalized as an explicit architecture that makes its interdependencies and systemic implications transparent. This synthesis serves the following purposes. First, it establishes conceptual coherence across the research objectives. Second, it clarifies the dependency structure linking infrastructure security, causal feasibility, and preference-aware optimization within a unified operational pipeline. Finally, it articulates the central thesis of the proposal of *an integrated, trust-centered architecture for AI-driven service systems*.

7.1 From Isolated Trust to Integrated Trust

Trust in AI-driven systems has often been addressed through fragmented interventions. Infrastructure security ensures robustness against adversarial manipulation; explainability provides interpretability; multi-objective counterfactual frameworks enable actionable

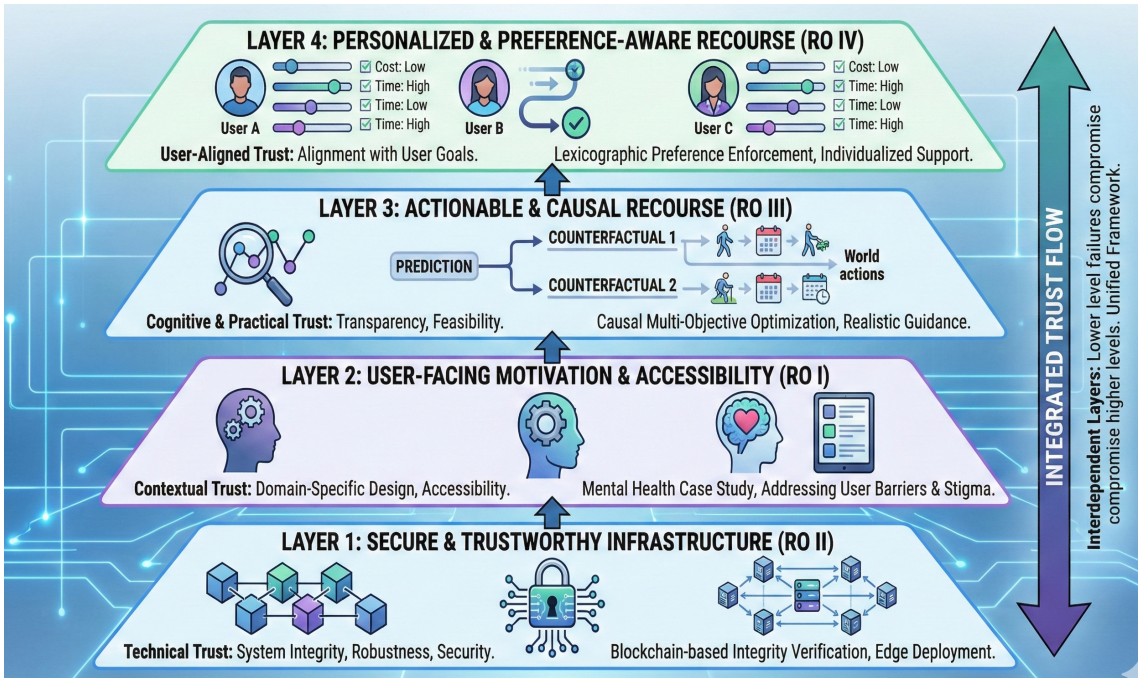


Figure 7.1: A Layered Trust-Centric Framework for Secure, User-Centric AI-driven systems (Image generated using Google Gemini)

recourse under competing objectives; and personalization incorporates user's preferences. Yet these dimensions are commonly treated as independent design goals rather than interdependent system properties. Fragmented approaches demonstrate that each trust dimension is necessary but insufficient: security without recourse is opaque, recourse without personalization is impractical, and personalization without integrity is vulnerable. These interdependencies reveal that trustworthy AI emerges from coordinated layering rather than isolated interventions.

In light of these limitations, this dissertation consolidates its contributions by establishing that assurance in AI-driven service ecosystems is architecturally stratified rather than monolithic. It rests upon three interdependent domains:

- Foundational system guarantees: secured through verifiable integrity, robustness, and resilient deployment mechanisms
- Operational soundness: achieved through causally coherent and practically feasible recourse generation
- Individual congruence: realized through the incorporation of user-specific priorities, constraints, and decision preferences

Each domain mitigates a distinct vulnerability in AI service delivery, yet none is sufficient in isolation. These reciprocal dependencies define the conditions under which AI-driven services are dependable, actionable, and aligned with user intent.

7.1.1 Limitations of Isolated Trust Mechanisms

The necessity of integration becomes evident when considering the limitations of isolated mechanisms.

- Process-Centric Limitations: Trust evolves over the lifecycle of an AI system [142]. A system that is initially secure may fail to maintain user confidence if causal reasoning becomes inconsistent or personalization needs shift [143, 144]. Fragmented mechanisms cannot adapt coherently to dynamic environments, leaving gaps in reliability.
- Interaction Perspective: Trust emerges from the interplay between system components and users. Even when individual mechanisms perform well in isolation, uncoordinated interactions can produce unpredictable or contradictory outcomes [145, 146].
- Risk Management: Isolated mechanisms create compounded vulnerabilities [147]. Security alone cannot prevent misleading recommendations; causal reasoning alone

cannot prevent exploitation; personalization alone cannot safeguard integrity. Without integrated layering, these limitations propagate systemically, producing risks that no single component can mitigate.

- Cognitive Load and User Trust: Human trust depends on perceiving the system as dependable, actionable, and aligned with personal needs [148, 149]. Misaligned robustness, reasoning, or personalization leads to confusion, frustration, and erosion of trust.

These limitations demonstrate that trustworthy AI requires architecturally integrated layering, where infrastructure integrity anchors causal reasoning, causal feasibility enables effective personalization, and the interaction of all layers ensures reliable, usable, and user-aligned outcomes.

7.2 Layered Architecture of the Integrated Trust-Centric Framework

The integrated trust-centric framework consists of four interdependent layers, each corresponding to a research objective (RO I – RO IV) and addressing a distinct dimension of system trustworthiness (see Fig. 7.1). The layers are ordered according to functional dependency, reflecting how higher-level capabilities presuppose guarantees established at lower levels.

7.2.1 Layer 1: Secure and Trustworthy Infrastructure

Establishes system-level integrity through robustness, verifiability, and resistance to adversarial manipulation. It ensures that model parameters, predictions, and system interactions remain reliable under adversarial or distributed conditions, serving as the foundational anchor for all higher-level recourse mechanisms.

7.2.2 Layer 2: User-Facing Motivation and Accessibility

Mediates between system capabilities and user context, ensuring that outputs from robust infrastructure translate into accessible and interpretable guidance aligned with domain constraints.

7.2.3 Layer 3: Actionable and Causal Recourse

Introduces causal reasoning to counterfactual explanations to ensure feasible intervention pathways. Using structural modeling and multi-objective optimization, it ensures that recommendations respect causal dependencies and real-world constraints.

7.2.4 Layer 4: Personalized and Preference-Aware Recourse

Incorporates individual user preferences into recourse generation. Through preference modeling and lexicographic enforcement, it aligns recommendations with user-specific priorities such as cost, effort, time, or risk tolerance.

Each layer presupposes the guarantees established at lower levels, forming a sequentially dependent trust stack in which higher-level personalization and adoption rely on foundational integrity and causal coherence. Together, the four layers form the proposed architecture in which security, causal reasoning, and user alignment are interdependent, enabling AI recourse that is robust, actionable, and personalized.

7.3 Cross-Layer Dependency Analysis and Systemic Trade-Offs

Building on this layered structure, the framework enables analysis of how guarantees propagate across levels and how failures cascade through the stack.

Layer 1 establishes the foundation; without technical integrity, higher-level reasoning and personalization lack credibility, and failures propagate upward. Layer 2 translates

technically valid outputs into usable guidance, bridging system reliability and cognitive adoption; without accessibility, even valid recommendations remain unused. Layer 3 enforces structural causal constraints, ensuring that proposed interventions are feasible and coherent, thereby enabling meaningful personalization. Layer 4 aligns interventions with individual preferences, reconciling trade-offs among cost, time, and risk. Because each layer presupposes the guarantees of those beneath it, degradation at any foundational level produces cascading effects across the stack. Trustworthiness, therefore, emerges not from isolated assurances but from cross-layer coherence.

These interdependencies also expose inherent tensions in AI system design. Optimizing one layer often influences constraints in another, such as, stronger security mechanisms may introduce computational overhead; increased personalization may interact with fairness constraints; enhanced transparency may conflict with privacy requirements; and complex multi-objective optimization may reduce usability. By structuring trust across technical, practical, and user-aligned dimensions, the framework makes such trade-offs explicit and provides a principled basis for balancing system-level guarantees with actionable, user-centered outcomes.

7.4 Contributions of the Trust-Centric Framework

The trust-centric framework bridges theory and practice, providing both a structured foundation for research and actionable guidance for designing robust, user-aligned, and socially responsible AI systems. The contributions from this framework are described in the following subsections.

7.4.1 Theoretical Contributions

The following are the theoretical contributions of this research to advancing trust in AI-driven service systems.

1. Multi-dimensional Trust Formalization: By an explicit distinction of technical, practical, and user-aligned trust, the framework moves beyond monolithic conceptions of trust and offers a structured taxonomy for future research.

2. Layered Dependency Modeling: Trust is formalized as an emergent system property arising from hierarchical dependencies across infrastructure, personalization, and reasoning layers, enabling analytical examination of cross-layer trade-offs.

3. Integration of Causal Reasoning and Personalization: The framework demonstrates how structural causal models can be reconciled with individualized user preferences, providing a pathway to reconcile system-level guarantees with heterogeneous human priorities.

4. Bridging Theory and Practice: By connecting theoretical constructs (e.g., causal validity, user alignment) with operational layers (e.g., accessibility, infrastructure security), the framework establishes a foundation for both analytical evaluation and applied system design.

These insights extend prior research by reframing trust as a systemic property rather than a set of independent interventions, with implications for AI ethics and actionable recourse design.

7.4.2 Practical Contributions to AI System Design

The proposed framework provides actionable guidance for designing and deploying trustworthy AI systems, categorized into five operational principles:

1. Integrity Before Intelligence: System verification and robust underlying infrastructure are foundational. Decision-making relies on integrity guarantees such as distributed ledger consistency, blockchain-based validation, and secure deployment, making infrastructure verification a prerequisite for actionable AI.

2. Explanation Must Enable Action: Transparency alone is insufficient; explanations must translate into feasible, actionable guidance that users can follow, accounting for cognitive and contextual barriers. Achieving this requires user-centric mediation that incorporates contextual, accessibility, and cognitive constraints into interface design.

3. Causality Constrains Recourse: Personalized suggestions must respect underlying causal dependencies, ensuring that modifications are structurally coherent and feasible.

4. Preferences Are Hierarchical: User priorities, constraints, and trade-offs should be enforced lexicographically, balancing multiple objectives such as cost, time, and risk. Hierarchical enforcement ensures key user priorities are met before addressing secondary objectives, producing feasible and goal-aligned interventions.

5. Trust Is Layer-Dependent: Higher-level guarantees, such as personalization and adoption, depend on the integrity, feasibility, and accessibility ensured by lower layers. Because performance and reliability propagate across layers, evaluation must extend beyond model accuracy to include infrastructure robustness, recourse feasibility, and user-aligned usability metrics. This provides a multi-dimensional evaluation of system performance.

By following these principles, system architects and developers can operationalize layered trust, producing AI interventions that are robust, actionable, and user-aligned, while managing trade-offs between security, usability, transparency, and fairness.

7.4.3 Societal Contribution

Trustworthy AI-driven systems operate at the intersection of technical performance, user empowerment, and social responsibility. The layered framework highlights several societal and ethical implications:

1. Equity and Accessibility: By incorporating contextual mediation and accessibility into its design, the framework promotes inclusive AI adoption, enabling diverse populations to benefit equitably from system recommendations.

2. Transparency and Accountability: Layered causal reasoning provides traceability from predictions to actionable recourse, supporting interpretability, stakeholder oversight, and societal accountability.

3. Personal Autonomy: Preference-aware personalization safeguards individual agency, ensuring interventions reflect user-defined priorities rather than externally imposed norms.

4. Systemic Risk Mitigation: By embedding security and integrity at the foundation, the framework reduces exposure to adversarial manipulation and erroneous outputs, contributing to safer AI deployment with minimized societal harm.

Collectively, these contributions show that trust is a socio-technical property, not merely a technical one. The framework offers a blueprint for AI-driven service systems that are operationally robust and socially responsible, aligning innovation with ethical and societal imperatives.

7.5 Limitations of the Trust-Centric Framework

While the proposed trust-centric framework advances both the conceptual and operational understanding of AI-driven systems, several limitations warrant explicit consideration:

1. Abstraction vs. Implementation Complexity: The layered architecture is intentionally conceptual, providing a high-level blueprint rather than a fully instantiated system. Translating the framework into production-grade AI pipelines requires domain-specific adaptation, detailed engineering, and iterative testing, which may introduce unforeseen trade-offs between layers.

2. Causal Model Assumptions: The framework relies on causal inference for actionable recourse. This depends on the correctness of causal models and the availability of reliable observational or experimental data. Structural mis-specifications or latent confounders could reduce the validity of recommended interventions, which undermines the quality of the explanations given to users.

3. Dynamic User Preferences: Preference-aware personalization assumes relatively stable user priorities. In reality, user goals and constraints can evolve rapidly, potentially limiting the effectiveness of static or predefined preference models. This necessitates exploring adaptive mechanisms for real-time preference updates.

4. Computational Scalability: Multi-objective optimization across causal, technical, and user-alignment constraints can be computationally intensive, particularly for large-scale or high-dimensional datasets. While approximate methods can alleviate some burdens, computational efficiency remains a practical consideration.

Acknowledging these limitations highlights opportunities for future research while situating the current framework as a theoretically grounded and methodologically rigorous starting point for layered trust evaluation in AI-driven systems.

CHAPTER EIGHT

CONCLUSION

This research advanced a trust-centric perspective on AI-driven systems, demonstrating that reliable, actionable, and user-aligned AI emerges through the integration of interrelated capabilities rather than isolated interventions. Progressing from user-facing trust requirement to verifiable infrastructure guarantees, causally coherent interventions, and personalized, preference-aware recourse, the research establishes a layered framework in which trust arises from cross-layer dependencies, systemic coherence, and principled trade-offs between technical, practical, and user-aligned objectives. This chapter synthesizes the key findings, draws conclusions, and outlines directions for advancing adaptive, scalable, and ethically grounded AI systems.

8.1 Key Findings

The research yielded significant and quantifiable improvements across all stages of investigation, advancing the overarching trust-centric architecture. Within the trust problem space of AI-driven systems, using mental healthcare as a motivating scenario, existing diagnostic approaches are methodologically diverse but architecturally fragmented. Even technically robust models struggle to achieve real-world reliability, user confidence, and sustained adoption in the absence of an integrated, scalable, privacy-preserving, accessible, and quality-aware system design.

Grounded in this user-centric trust foundation, the work in User-Facing Motivation and Accessibility (RO I) establishes a foundational understanding of how ambiguity, limited observability, and barriers to access shape user trust in sensitive, high-impact domains. Drawing from these insights, Secure and Trustworthy Infrastructure (RO II) addresses system-level trust by embedding verifiability, tamper-resistance, and resilience against

adversarial manipulation through a blockchain-based framework that validates AI model parameters in edge environments. This layer strengthens the credibility and reliability of AI-driven services, forming the technical anchor upon which operational and user-aligned trust mechanisms are constructed.

Extending the user-centric perspective of RO I and the infrastructure guarantees of RO II, Actionable and Causal Recourse (RO III) advances operational trust by ensuring that AI-driven interventions are both feasible and meaningful. By integrating structural causal modeling with multi-objective optimization, this contribution generates realistic, diverse, and cost-efficient counterfactual explanations that respect causal dependencies and practical constraints, showing an average improvement of approximately 38.8% compared to existing frameworks. In doing so, RO III bridges the secure infrastructure layer with user-aligned personalization, enhancing the reliability, interpretability, and actionable value of AI recommendations.

Finally, Personalized and Preference-Aware Recourse (RO IV) leverages the causal and operational guarantees established in RO III to ensure user-aligned reliability. By embedding elicited user priorities within a causally grounded, lexicographic multi-objective optimization framework, this layer produces counterfactual explanations that are feasible, personalized, and aligned with user intent, achieving a 35% improvement relative to prior literature. RO IV completes the layered framework, uniting secure infrastructure, actionable recourse, and individualized alignment with user priorities into a cohesive, end-to-end system that operationalizes the central thesis of a fully integrated, dependable, and user-centric AI service architecture

Together, these objectives demonstrate how layered, interdependent mechanisms produce reliable, actionable, and user-aligned trust, forming the foundation for the conclusions drawn in the following section.

8.2 Conclusion

This work reframes trust in AI-driven systems as a systemic property emerging from the coordinated integration of secure infrastructure, actionable reasoning, and user-aligned personalization. Beyond individual methodological advances, this research establishes a layered, trust-centric architecture formalizing dependencies among technical integrity, operational feasibility, and user alignment as a principled blueprint for AI system design. The conceptual and operational insights from this architecture illuminate how trust can be engineered across AI services, providing principles for designing systems that are not only technically robust but also interpretable, actionable, and socially responsible.

By synthesizing theoretical, practical, and societal perspectives, this work underscores that trustworthy AI requires holistic system design rather than isolated interventions. It provides a blueprint for future research on adaptive, scalable, and ethically grounded AI ecosystems, including the incorporation of dynamic user preferences, generalized causal modeling, and cross-domain deployment. Ultimately, this research contributes both a unifying conceptual framework and a set of operational principles that can guide the development of next-generation AI systems where trust is embedded by design.

8.3 Future Work

The findings from this research open several promising avenues for future research, aiming to enhance the robustness, intelligence, user alignment, and real-world applicability of the proposed counterfactual and trust-centric frameworks.

A key direction will focus on extending the counterfactual generation framework to handle incomplete or imperfect causal models, incorporate heterogeneous and unstructured data beyond tabular formats, and refine recourse efficiency metrics to account for plausibility, proximity, and weighted effort. These extensions are essential to ensure that counterfactual

explanations remain feasible, personalized, and actionable across diverse and evolving user contexts.

Another important direction focuses on enhancing intelligence and adaptivity within the layered trust-centric architecture. This includes implementing mechanisms to continuously update user preferences in real time and dynamically validate causal relationships, closing the loop between secure infrastructure, operational feasibility, and user-aligned personalization. This involves designing pipelines that mediate trade-offs among privacy, fairness, and transparency, enabling context-aware deployment across domains beyond mental healthcare. By operationalizing these feedback-informed mechanisms, future research aims to create AI services that are operationally reliable, socially responsible, and contextually adaptive, moving closer to fully integrated, trust-by-design ecosystems.

REFERENCES

[1] L. M. Candanedo and V. Feldheim, "Accurate occupancy detection of an office room from light, temperature, humidity and co2 measurements using statistical learning models," *Energy and Buildings*, vol. 112, pp. 28–39, 2016.

[2] L. Cao, "Ai in finance: challenges, techniques, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–38, 2022.

[3] M. Y. Shaheen, "Applications of artificial intelligence (ai) in healthcare: A review," *ScienceOpen Preprints*, 2021.

[4] E. K. Kelan, "Algorithmic inclusion: Shaping the predictive algorithms of artificial intelligence in hiring," *Human Resource Management Journal*, vol. 34, no. 3, pp. 694–707, 2024.

[5] N. Montealegre-López, "Exploring the role of trust in ai-driven decision-making: a systematic literature review," *Management Review Quarterly*, pp. 1–51, 2025.

[6] O. Vereschak, F. Alizadeh, G. Bailly, and B. Caramiaux, "Trust in ai-assisted decision making: Perspectives from those behind the system and those for whom the decision is made," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2024.

[7] A. Kovari, "Ai for decision support: Balancing accuracy, transparency, and trust across sectors," *Information*, vol. 15, no. 11, p. 725, 2024.

[8] P. R. Brandao, "The impact of artificial intelligence on modern society," *AI*, vol. 6, no. 8, p. 190, 2025.

[9] B. C. Cheong, "Transparency and accountability in ai systems: safeguarding wellbeing in the age of algorithmic decision-making," *Frontiers in Human Dynamics*, vol. 6, p. 1421273, 2024.

[10] J. Law, "The ethical imperative of algorithmic fairness in ai-enabled hiring: a critical analysis of bias, accountability, and justice," *AI and Ethics*, vol. 6, no. 1, p. 65, 2026.

[11] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, and P. S. Yu, "Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity," *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–39, 2022.

[12] H. Lyons, E. Velloso, and T. Miller, "Conceptualising contestability: Perspectives on contesting algorithmic decisions," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–25, 2021.

[13] H. Bleher and M. Braun, "Diffused responsibility: attributions of responsibility in the use of ai-driven clinical decision support systems," *AI and Ethics*, vol. 2, no. 4, pp. 747–761, 2022.

[14] A. Boch, "The engineered illusion: ethical risks and negotiable benefits of ai deception," in *Expert Essentials*, Oxford University Press.

[15] O. Vereschak, F. Alizadeh, G. Bailly, and B. Caramiaux, "Trust in ai-assisted decision making: Perspectives from those behind the system and those for whom the decision is made," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2024.

[16] S. Afroogh, A. Akbari, E. Malone, M. Kargar, and H. Alambeigi, "Trust in ai: progress, challenges, and future directions," *Humanities and Social Sciences Communications*, vol. 11, no. 1, pp. 1–30, 2024.

[17] X. Zhong, Y. Chen, W. Chen, Y. Liu, S. Gui, J. Pu, D. Wang, Y. He, X. Chen, X. Chen, *et al.*, "Identification of potential biomarkers for major depressive disorder: based on integrated bioinformatics and clinical validation," *Molecular Neurobiology*, vol. 61, no. 12, pp. 10355–10364, 2024.

[18] Kaiser Family Foundation, "Mental health care health professional shortage areas (hpsas)." `https://www.kff.org/other/state-indicator/mental-health-care-health-professional-shortage-areas`. [accessed: 09.10.2021].

[19] J. Lally, A. ó Conghaile, S. Quigley, E. Bainbridge, and C. McDonald, "Stigma of mental illness and help-seeking intention in university students," *The Psychiatrist*, vol. 37, no. 8, p. 253–260, 2013.

[20] Y. Hou, S. Garg, L. Hui, D. N. K. Jayakody, R. Jin, and M. S. Hossain, "A data security enhanced access control mechanism in mobile edge computing," *IEEE Access*, vol. 8, pp. 136119–136130, 2020.

[21] J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data security and privacy-preserving in edge computing paradigm: Survey and open issues," *IEEE Access*, vol. 6, pp. 18209–18237, 2018.

[22] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *Information Fusion*, vol. 81, pp. 59–83, 2022.

[23] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[24] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the conference on fairness, accountability, and transparency*, pp. 10–19, 2019.

[25] M. Bhan, J.-N. Vittaut, N. Chesneau, and M.-J. Lesot, "Tigtec: Token importance guided text counterfactuals," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 496–512, Springer, 2023.

[26] G. D. Toni, P. Viappiani, S. Teso, B. Lepri, and A. Passerini, "Personalized algorithmic recourse with preference elicitation," *Transactions on Machine Learning Research*, 2024.

[27] D. Xu, T. Li, Y. Li, X. Su, S. Tarkoma, T. Jiang, J. Crowcroft, and P. Hui., "Edge intelligence: Architectures, challenges, and applications," *Networking and Internet Architecure (cs.NI); Artificial Intelligence (cs.AI)*. arXiv:2003.12172v2 [cs.NI].

[28] M. Mukherjee, R. Matam, C. X. Mavromoustakis, H. Jiang, G. Mastorakis, and M. Guo, "Intelligent edge computing: Security and privacy challenges," *IEEE Communications Magazine*, vol. 58, no. 9, pp. 26–31, 2020.

[29] S. M. H. Bamakan, A. Motavali, and A. B. Bondarti., "A survey of blockchain consensus algorithms performance evaluation criteria," *Expert Systems with Applications.*, vol. 154, pp. 2567 – 2572, September 2020. https://doi.org/10.1016/j.eswa.2020.113385.

[30] P. Rasouli and I. Chieh Yu, "Care: Coherent actionable recourse based on sound counterfactual explanations," *International Journal of Data Science and Analytics*, vol. 17, no. 1, pp. 13–38, 2024.

[31] S. Dasgupta, S. M. Halim, J. Arias, E. Salazar, and G. Gupta, "Mc3g: Model agnostic causally constrained counterfactual generation," in *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, vol. 284, pp. 926–937, 2025.

[32] S. S. Kancheti, R. Vigneswaran, B. Mishra, and V. N. Balasubramanian, "HARE: Human-in-the-loop algorithmic recourse," *Transactions on Machine Learning Research*, 2025.

[33] D. Kim and N. Moniz, "Relevance-aware algorithmic recourse," in *International Symposium on Intelligent Data Analysis*, pp. 444–455, Springer, 2025.

[34] A. Jeyasothy, T. Laugel, M.-J. Lesot, C. Marsala, and M. Detyniecki, "A general framework for personalising post hoc explanations through user knowledge integration," *International Journal of Approximate Reasoning*, vol. 160, p. 108944, 2023.

[35] A. Roy and S. Madria, "Secure and privacy-preserving traffic monitoring in vanets," in *2020 IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pp. 567–575, 2020.

[36] S. Singh, R. Sulthana, T. Shewale, V. Chamola, A. Benslimane, and B. Sikdar, "Machine-learning-assisted security and privacy provisioning for edge computing: A survey," *IEEE Internet of Things Journal*, vol. 9, no. 1, pp. 236–260, 2022.

[37] D. Yue, R. Li, Y. Zhang, W. Tian, and Y. Huang, "Blockchain-based verification framework for data integrity in edge-cloud storage," *Journal of Parallel and Distributed Computing*, vol. 146, pp. 1–14, 2020.

[38] H. Wang, J. Zhang, Y. Lin, and H. Huang, "Zss signature based data integrity verification for mobile edge computing," in *2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, pp. 356–365, 2021.

[39] T. Ma, B. Du, M. Li, J. Li, Y. Shi, and H. Fan, "Toward data authenticity and integrity for blockchain-based mobile edge computing," *IEEE Sensors Journal*, vol. 22, no. 10, pp. 9967–9980, 2022.

[40] X. Xu, H. Wang, Y. Lin, and F. Xiao, "A lightweight data integrity verification with data dynamics for mobile edge computing," *Hindawi, Security and Communication Networks*, vol. 2022, 2022.

[41] D. Mingxiao, M. Xiaofeng, Z. Zhe, W. Xiangwei, and C. Qijun, "A review on consensus algorithm of blockchain," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2567–2572, 2017.

[42] S. M. Alrubei, E. Ball, and J. M. Rigelsford, "A secure blockchain platform for supporting ai-enabled iot applications at the edge layer," *IEEE Access*, vol. 10, pp. 18583–18595, 2022.

[43] L. Yuan, Q. He, S. Tan, B. Li, J. Yu, F. Chen, H. Jin, and Y. Yang, "Coopedge: A decentralized blockchain-based platform for cooperative edge computing," in *Proceedings of the Web Conference 2021*, WWW '21, (New York, NY, USA), p. 2245–2257, Association for Computing Machinery, 2021.

[44] Y. Fan, H. Wu, and H.-Y. Paik, "Dr-bft: A consensus algorithm for blockchain-based multi-layer data integrity framework in dynamic edge computing system," *Future Generation Computer Systems*, vol. 124, pp. 33–48, 2021.

[45] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, vol. 38, no. 5, pp. 2770–2824, 2024.

[46] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *International Conference on Machine Learning*, pp. 2376–2384, PMLR, 2019.

[47] A. Lucic, H. Haned, and M. de Rijke, "Why does my model fail?: contrastive local explanations for retail forecasting," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, p. 90–98, ACM, Jan. 2020.

[48] C. Russell, "Efficient search for diverse coherent explanations," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, (New York, NY, USA), p. 20–28, Association for Computing Machinery, 2019.

[49] E. M. Kenny and M. T. Keane, "On generating plausible counterfactual and semi-factual explanations for deep learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 11575–11585, May 2021.

[50] J. Zhu, H. Xie, J. Li, and W. Abd-Almageed, "Diffusioncounterfactuals: Inferring high-dimensional counterfactuals with guidance of causal representations," 2024.

[51] E. Panagiotou, M. Heurich, T. Landgraf, and E. Ntoutsi, "Tabcf: Counterfactual explanations for tabular data using a transformer-based vae," in *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 274–282, 2024.

[52] K. Kanamori, T. Takagi, K. Kobayashi, and H. Arimura, "Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization.," in *IJCAI*, pp. 2855–2862, 2020.

[53] S. Sharma, J. Henderson, and J. Ghosh, "Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 166–172, 2020.

[54] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the conference on fairness, accountability, and transparency*, pp. 10–19, 2019.

[55] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," in *International conference on parallel problem solving from nature*, pp. 448–469, Springer, 2020.

[56] S. Dash, V. N. Balasubramanian, and A. Sharma, "Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 915–924, 2022.

[57] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "Face: feasible and actionable counterfactual explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, 2020.

[58] A.-H. Karimi, J. Von Kügelgen, B. Schölkopf, and I. Valera, "Algorithmic recourse under imperfect causal knowledge: a probabilistic approach," *Advances in neural information processing systems*, vol. 33, pp. 265–277, 2020.

[59] M. Schleich, Z. Geng, Y. Zhang, and D. Suciu, "Geco: quality counterfactual explanations in real time," *Proc. VLDB Endow.*, vol. 14, p. 1681–1693, May 2021.

[60] C. Russell, "Efficient search for diverse coherent explanations," in *Proceedings of the conference on fairness, accountability, and transparency*, pp. 20–28, 2019.

[61] J. Yetukuri, I. Hardy, and Y. Liu, "Towards user guided actionable recourse," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 742–751, 2023.

[62] Z. J. Wang, J. Wortman Vaughan, R. Caruana, and D. H. Chau, "Gam coach: Towards interactive and user-centered algorithmic recourse," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2023.

[63] S. Esfahani, G. De Toni, B. Lepri, A. Passerini, K. Tentori, and M. Zancanaro, "Preference elicitation in interactive and user-centered algorithmic recourse: an initial exploration," (New York, NY, USA), Association for Computing Machinery, 2024.

[64] Anxiety and Depression Association of America, "Understand anxiety and depression, facts and statistics." `https://adaa.org/understanding-anxiety/facts-statistics`. [accessed: 09.10.2021].

[65] Mental Health America, "2021: Covid-19 and mental health: A growing crisis." `https://mhanational.org/sites/default/files/Spotlight2021-COVID-19andMentalHealth.pdf`. [accessed: 02.16.2021].

[66] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent(wysa) for digital mental well-being: Real-world data evaluation mixed-methods study," *JMIR Mhealth Uhealth*, Nov 2018.

[67] D. Cupkova, E. Kajati, J. Mocnej, P. Papcun, J. Koziorek, and I. Zolotova, "Intelligent human-centric lighting for mental wellbeing improvement," *International Journal of Distributed Sensor Networks*, vol. 15, August 2019.

[68] R. Johansson and G. Andersson, "Internet-based psychological treatments for depression," *Expert Review of Neurotherapeutics*, 2012. [12:7, 861-870, DOI: 10.1586/ern.12.63].

[69] H. A. M. Motlagh, B. M. Bidgoli, and A. A. P. Fard, "Design and implementation of a web-based fuzzy expert system for diagnosing depressive disorder," *Appl Intell (2018)*, 2017.

[70] G. Dosovitsky, B. S. Pineda, N. C. Jacobson, C. Chang, M. Escoredo, and E. L. Bunge, "Artificial intelligence chatbot for depression: Descriptive study of usage," *JMIR Form Res 2020;4(11):e17065*, 2020. [doi: 10.2196/17065].

[71] T. Alhanai, M. Ghassemi, and J. Glass, "Detecting depression with audio/text sequence modeling of interviews," *Interspeech*, September 2018.

[72] I. A. Alshawwa, M. Elkahlout, H. Q. El-Mashharawi, and S. S. Abu-Naser, "An expert system for depression diagnosis," *International Journal of Academic Health and Medical Research (IJAHMR)*, vol. 3, pp. 20–27, April 2019.

[73] M. D. Nemesure, M. V. Heinz, R. Huang, and N. C. Jacobson, "Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence," *Scientific Reports (2021)11:1980 https://doi.org/10.1038/s41598-021-81368-4*, Jan 2021.

[74] N. C. Jacobson and M. D. Nemesure, "Using artificial intelligence to predict change in depression and anxiety symptoms in a digital intervention: Evidence from a transdiagnostic randomized controlled trial," *Psychiatry Research https://doi.org/10.1016/j.psychres.2020.113618*, vol. 295, 2021.

[75] M. D. Nemesure, M. V. Heinz, R. Huang, and N. C. Jacobson, "Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence," *Scientific reports*, vol. 11, no. 1, p. 1980, 2021.

[76] E. Victor, Z. M. Aghajan, A. R. Sewart, and R. Christian, "Detecting depression using a framework combining deep multimodal neural networks with a purpose-built automated evaluation.," *Psychological assessment*, vol. 31, no. 8, p. 1019, 2019.

[77] henkai Yang, C. Chen, H. Li, L. Yao, and X. Zhao, "Unsupervised classifications of depression levels based on machine learning algorithms perform well as compared to traditional normbased classifications," *Frontiers in Psychiatry* |, 2020. doi: 10.3389/fpsyt.2020.00045.

[78] A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth, "Semi-supervised approach to monitoring clinical depressive symptoms in social media," *IEEE/ACM Advances in Social Networks Analysis and Mining*, july 2017.

[79] V. M. Brown, L. Zhu, A. Solway, J. M. Wang, K. L. McCurry, B. King-Casas, and P. H. Chiu, "Reinforcement learning disruptions in individuals with depression and sensitivity to symptom change following cognitive behavioral therapy," *JAMA Psychiatry*, july 2021.

[80] S. Khedkar, V. Subramanian, G. Shinde, and P. Gandhi, "Explainable ai in healthcare," *ICAST*, 2019.

[81] R. Zainab and R. Chandramouli, "Detecting and explaining depression in social media text with machine learning," *KDD 2020*, August 2020.

[82] C. Guestrin, S. Singh, and M. T. Ribeiro, "Why should i trust you? explaining the predictions of any classifier," *ACM*, August 2016.

[83] J. Yu, C. Chiu, Y. Wang, E. Dzubur, W. Lu, and J. Hoffman, "A machine learning approach to passively informed prediction of mental health risk in people with diabetes: Retrospective case-control analysis," *JIMIR*, August 2021.

[84] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[85] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE Transactions on Visualization and Computer Graphics*, October 2019.

[86] IBM Watson Labs. `https://cloud.ibm.com/docs/personality-insights?topic=personality-insights-about`. Last Updated: 08.26.2021.

[87] I. E. C. (IEC), "Edge intelligence," October 2017. last accessed: 03/13/2023.

[88] Y. Hao, Y. Li, X. Dong, L. Fang, and P. Chen, "Performance analysis of consensus algorithm in private blockchain," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 280–285, 2018.

[89] N. Hudson, M. J. Hossain, M. Hosseinzadeh, H. Khamfroush, M. Rahnamay-Naeini, and N. Ghani, "A framework for edge intelligent smart distribution grids via federated learning," in *2021 International Conference on Computer Communications and Networks (ICCCN)*, pp. 1–9, 2021.

[90] J. Peña Queralta, T. Nguyen gia, H. Tenhunen, and T. Westerlund, "Edge-ai in lora-based health monitoring: Fall detection system with fog computing and lstm recurrent neural networks," 07 2019.

[91] O. Oyeniyi, S. Dhandhukia, A. Sen, and K. K. Fletcher, "A study of ai frameworks for major depressive disorder diagnosis," in *International Workshop on AI-enabled Process Automation*, 2021.

[92] J. Zhang and K. B. Letaief, "Mobile edge intelligence and computing for the internet of vehicles," *Proceedings of the IEEE*, vol. 108, no. 2, pp. 246–261, 2020.

[93] J. Liu, J. Xiang, Y. Jin, R. Liu, J. Yan, and L. Wang, "Boost precision agriculture with unmanned aerial vehicle remote sensing and edge intelligence: A survey," *Remote Sensing*, vol. 13, no. 21, 2021.

[94] D. Van Huynh, S. R. Khosravirad, A. Masaracchia, O. A. Dobre, and T. Q. Duong, "Edge intelligence-based ultra-reliable and low-latency communications for digital twin-enabled metaverse," *IEEE Wireless Communications Letters*, vol. 11, no. 8, pp. 1733–1737, 2022.

[95] S. L. D. W. Sandro), "3 ai marketplaces everyone has to know [one will define the century]," nov 2020. last accessed: 03.25.2023.

[96] W. Tong, W. Chen, B. Jiang, F. Xu, Q. Li, and S. Zhong, "Privacy-preserving data integrity verification for secure mobile edge storage," *IEEE Transactions on Mobile Computing*, vol. 22, no. 9, pp. 5463–5478, 2023.

[97] P. Garcia Lopez, A. Montresor, D. Epema, A. Datta, T. Higashino, A. Iamnitchi, M. Barcellos, P. Felber, and E. Riviere, "Edge-centric computing: Vision and challenges," *SIGCOMM Comput. Commun. Rev.*, vol. 45, p. 37–42, Sept. 2015.

[98] M. A. Rahman, M. S. Hossain, G. Loukas, E. Hassanain, S. S. Rahman, M. F. Alhamid, and M. Guizani, "Blockchain-based mobile edge computing framework for secure therapy applications," *IEEE Access*, vol. 6, pp. 72469–72478, 2018.

[99] A. Stanciu, "Blockchain based distributed control system for edge computing," in *2017 21st International Conference on Control Systems and Computer Science (CSCS)*, pp. 667–671, 2017.

[100] M. Zhaofeng, M. Jialin, W. Jihui, and S. Zhiguang, "Blockchain-based decentralized authentication modeling scheme in edge and iot environment," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2116–2123, 2021.

[101] B. Li, Q. He, F. Chen, H. Jin, Y. Xiang, and Y. Yang, "Auditing cache data integrity in the edge computing environment," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 5, pp. 1210–1223, 2021.

[102] H. HaddadPajouh, R. Khayami, A. Dehghantanha, C. K.-K. Raymond, and R. M. Parizi, "Ai4safe-iot: an ai-powered secure architecture for edge layer of internet of things," *Neural Computing and Applications*, vol. 32, no. 20, p. 16119–16133, 2019.

[103] "Vosk." `https://alphacephei.com/vosk/models`, 2025.

[104] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[105] Mozilla, "Deepspeech," 2019.

[106] A. A. Menon, T. Saranya, S. Sureshbabu, and A. S. Mahesh, "A comparatıve analysis on three consensus algorithms," in *Computer Networks and Inventive Communication Technologies* (S. Smys, R. Bestak, R. Palanisamy, and I. Kotuliak, eds.), (Singapore), pp. 369–383, Springer Nature Singapore, 2022.

[107] Anonymous, "User centric counterfactual generation framework repository." `https://anonymous.4open.science/r/User_Centric_CFE_Generation-E338/`, 2026. GitHub repository.

[108] R. L. Shinmoto Torres, D. C. Ranasinghe, Q. Shi, and A. P. Sample, "Sensor enabled wearable rfid technology for mitigating the risk of falls near beds," in *2013 IEEE International Conference on RFID (RFID)*, pp. 191–198, 2013.

[109] V. Arzamasov, K. Böhm, and P. Jochem, "Towards concise models of grid stability," in *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pp. 1–6, 2018.

[110] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[111] A. Salam and A. E. Hibaoui, "Comparison of machine learning algorithms for the power consumption prediction : - case study of tetouan city –," in *2018 6th International Renewable and Sustainable Energy Conference (IRSEC)*, pp. 1–5, 2018.

[112] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, *et al.*, "Explainable ai (xai): Core ideas, techniques, and solutions," *ACM computing surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[113] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[114] 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*, (New York, NY, USA), p. 1135–1144, Association for Computing Machinery, 2016.

[115] M. T. Keane and B. Smyth, "Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai)," in *Case-Based Reasoning Research and Development: 28th International Conference,*

ICCBR 2020, Salamanca, Spain, June 8–12, 2020, Proceedings, (Berlin, Heidelberg), p. 163–178, 2020.

[116] B. Smyth and M. T. Keane, "A few good counterfactuals: Generating interpretable, plausible and diverse counterfactual explanations," in *Case-Based Reasoning Research and Development: 30th International Conference, ICCBR 2022, Nancy, France, September 12–15, 2022, Proceedings*, (Berlin, Heidelberg), p. 18–32, 2022.

[117] S. Dasgupta, F. Shakerin, J. Arias, E. Salazar, and G. Gupta, "C3g: Causally constrained counterfactual generation," in *Practical Aspects of Declarative Languages* (E. Erdem and G. Vidal, eds.), (Cham), pp. 215–232, Springer Nature Switzerland, 2025.

[118] D. Dua and C. Graff, "UCI machine learning repository," 2019.

[119] I.-C. Yeh, "Default of Credit Card Clients." UCI Machine Learning Repository, 2009.

[120] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," in *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, (New York, NY, USA), pp. 19–24, ACM, 2015.

[121] H. Hofmann, "Statlog (German Credit Data)." UCI Machine Learning Repository, 1994.

[122] "Home equity line of credit(heloc)." https://tinyurl.com/2mhujrmn, 2018.

[123] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 2020.

[124] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, Aug. 2018.

[125] L. Wang, S. Huang, S. Wang, J. Liao, T. Li, and L. Liu, "A survey of causal discovery based on functional causal model," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108258, 2024.

[126] S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvärinen, Y. Kawahara, T. Washio, P. O. Hoyer, and K. Bollen, "Directlingam: A direct method for learning a linear non-gaussian structural equation model," *J. Mach. Learn. Res.*, vol. 12, p. 1225–1248, July 2011.

[127] P.-E. Danielsson, "Euclidean distance mapping," *Computer Graphics and image processing*, vol. 14, no. 3, pp. 227–248, 1980.

[128] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part i: Solving problems with box constraints," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 4, pp. 577–601, 2014.

[129] A. Tversky, "Elimination by aspects: A theory of choice.," *Psychological review*, vol. 79, no. 4, p. 281, 1972.

[130] O. Oyeniyi, "Actionable counterfactuals explanations." `github.com/oluwafeyisayo/Actionable-CFEs.git`, November 2025.

[131] D. Ley, U. Bhatt, and A. Weller, "Diverse, global and amortised counterfactual explanations for uncertainty estimates," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 7390–7398, 2022.

[132] M. Mersha, K. Lam, J. Wood, A. K. AlShami, and J. Kalita, "Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction," *Neurocomputing*, vol. 599, p. 128111, 2024.

[133] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, vol. 38, no. 5, pp. 2770–2824, 2024.

[134] M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian, "Causal inference and counterfactual prediction in machine learning for actionable healthcare," *Nature Machine Intelligence*, vol. 2, no. 7, pp. 369–375, 2020.

[135] J. Gan, S. Zhang, C. Zhang, and A. Li, "Automated counterfactual generation in financial model risk management," in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 4064–4068, 2021.

[136] Anonymous, "Improving the practical utility of counterfactual explanations for machine learning models." Manuscript Under Review. IEEE Transactions on Artificial Intelligence, 2025.

[137] E. AlJalaud and M. Hosny, "Enhancing explainable artificial intelligence: Using adaptive feature weight genetic explanation (afwge) with pearson correlation to identify crucial feature groups," *Mathematics*, vol. 12, no. 23, 2024.

[138] M. Suffian, J. M. Alonso-Moral, and A. Bogliolo, "Introducing user feedback-based counterfactual explanations (ufce)," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 123, 2024.

[139] C. Abrate, F. Siciliano, F. Bonchi, and F. Silvestri, "Human-in-the-loop personalized counterfactual recourse," in *World Conference on Explainable Artificial Intelligence*, pp. 18–38, Springer, 2024.

[140] C. Paulo and A. M. G. da Silva, "Using data mining to predict secondary school student performance." Proceedings of the 5th Annual Future Business Technology Conference, 2008.

[141] A. Chernev, U. Böckenholt, and J. Goodman, "Choice overload: A conceptual review and meta-analysis," *Journal of Consumer Psychology*, vol. 25, no. 2, pp. 333–358, 2015.

[142] Q. Dang and G. Li, "Unveiling trust in ai: The interplay of antecedents, consequences, and cultural dynamics," *AI & SOCIETY*, pp. 1–24, 2025.

[143] J. Han and D. Ko, "Trust formation, error impact, and repair in human–ai financial advisory: A dynamic behavioral analysis," *Behavioral Sciences*, vol. 15, no. 10, p. 1370, 2025.

[144] S. Dhuliawala, V. Zouhar, M. El-Assady, and M. Sachan, "A diachronic perspective on user trust in ai under uncertainty," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5567–5580, 2023.

[145] AiThority Guest Author, "Why the next era of AI demands explainability: Building trust to avoid a costly rebuild," 2025. Accessed: 2026-02-13.

[146] Y. Li, B. Wu, Y. Huang, and S. Luan, "Developing trustworthy artificial intelligence: insights from research on interpersonal, human-automation, and human-ai trust," *Frontiers in psychology*, vol. 15, p. 1382693, 2024.

[147] D. Schwarz, "Unvalidated trust: Cross-stage vulnerabilities in large language model architectures," *arXiv preprint arXiv:2510.27190*, 2025.

[148] I. G. Mahlangu, A. Roy, and K. Hosseini, "Cognitive load and trust in ai-augmented workflows: Evaluating human factors in copilot design," 2025.

[149] W. Jiang, D. Li, and C. Liu, "Understanding dimensions of trust in ai through quantitative cognition: Implications for human-ai collaboration," *PloS one*, vol. 20, no. 7, p. e0326558, 2025.