Journal of Information Systems Engineering and Management
2025, 10(63s) e-ISSN:
2468-4376
https://www.jisem-journal.com/

Research Article

Toward Trustworthy Churn Prediction: A Comparative Study of Counterfactual Explanation Techniques

Dinesh Kollu[1] Swathi salagalla[2], Vineet Kumar[3]
Email ID: [1]dkollu50@gmail.com, [2]Salagallas@gmail.com

ARTICLE INFO

Received: 04 Oct 2025

Revised: 15 Nov 2025

Accepted: 24 Dec 2025

ABSTRACT

Customer churn prediction has become a mission-critical application of machine learning across telecommunications, banking, subscription services, and software-asa-service (SaaS) platforms. Despite significant advances in predictive modeling, the inherent opacity of high-performing black-box classifiers creates substantial barriers to organizational trust, regulatory compliance, and actionable decision support. Counterfactual explanations have emerged as a particularly compelling class of posthoc interpretability technique, providing stakeholders with algorithmically generated "what-if" scenarios that articulate the minimal feature-level changes necessary to alter a predicted outcome. This paper presents a comprehensive, literature-grounded comparative survey of six prominent counterfactual explanation methodologies applied in the context of customer churn prediction: Wachter et al.'s proximityconstrained optimization, Diverse Counterfactual Explanations (DiCE), the Contrastive Explanations Method (CEM), prototype-based counterfactuals, generative model-based approaches leveraging variational autoencoders, and gradient-based optimization techniques. The evaluation spans seven quality dimensions validity, proximity, sparsity, diversity, plausibility, robustness, and computational efficiency synthesized from the existing literature spanning 2015 to 2025. The paper further situates these methods within the broader discourse on trustworthy AI, fairness, algorithmic recourse, and regulatory compliance, with specific reference to GDPR Article 22 and the EU AI Act. Findings from the surveyed literature consistently indicate that no single counterfactual method achieves universal superiority, and that method selection must be governed by the operational, ethical, and technical priorities of the deployment context. This survey provides researchers and practitioners with a structured analytical foundation for understanding and deploying counterfactual explanations in trustworthy churn prediction systems.

Keywords: Counterfactual Explanations, Explainable AI, Customer Churn Prediction, Trustworthy AI, Algorithmic Recourse, Interpretability, DiCE, Machine Learning Transparency, Fairness-Aware AI

1. INTRODUCTION

Customer churn the voluntary discontinuation of a product or service by a customer represents one of the most financially consequential phenomena in customer relationship management. Industry analyses consistently indicate that retaining an existing customer cost substantially less than acquiring a new one, with acquisition cost ratios reported between five and seven times the retention cost across multiple sectors [1]. In response, organizations deploy predictive churn models to identify at-risk customers in advance, enabling proactive retention interventions such as targeted offers, personalized outreach, and service quality improvements.

Journal of Information Systems Engineering and Management
2025, 10(63s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ **Research Article**

The machine learning community has produced an extensive body of work on churn prediction methodologies. Early contributions employed logistic regression, decision trees, and survival analysis, valued primarily for their interpretability [2], [3]. The subsequent adoption of ensemble methods random forests, gradient boosting machines, and extreme gradient boosting delivered substantial gains in predictive accuracy across telecom, banking, and subscription domains [4], [5]. More recently, deep learning architectures including recurrent neural networks and attention-based transformers have been applied to sequential customer behavior data, achieving state-of-the-art performance on longitudinal churn datasets [6].

However, the transition to high-accuracy black-box models introduced a critical tension between predictive performance and interpretability. In high-stakes customer-facing applications, the inability to explain individual predictions limits organizational trust, constrains human oversight, and creates legal exposure. The European Union's General Data Protection Regulation (GDPR), enforced since 2018, explicitly grants individuals the right to meaningful explanations of automated decisions under Article 22 [7]. The proposed EU AI Act further categorizes automated customer-scoring systems as high-risk applications subject to enhanced transparency requirements [8]. These regulatory pressures have significantly accelerated research into explainable artificial intelligence (XAI) as applied to customer analytics.

Within the XAI landscape, counterfactual explanations occupy a distinctive and practically powerful position. Unlike feature attribution methods such as SHAP [9] or LIME [10], which decompose a model's prediction into constituent feature contributions, counterfactual explanations address a fundamentally different and more actionable question: *what is the minimal change to this customer's profile that would result in a different prediction?* A counterfactual statement such as "if this customer's contract were switched to an annual plan and their average monthly usage increased by 20%, they would no longer be predicted as churning" provides a directly actionable intervention roadmap for retention teams [11] [62].

Despite this practical appeal, the literature on counterfactual explanations in churn prediction remains fragmented. Methodological contributions are frequently evaluated on generic classification benchmarks without systematic attention to the domain-specific characteristics of churn data including class imbalance, feature immutability, temporal structure, and the ethical implications of differential recourse availability [12], [13]. This paper addresses these gaps through a comprehensive comparative survey grounded entirely in published literature from 2015 to 2025, providing a structured multi-dimensional analysis of six counterfactual methods and their suitability for trustworthy churn prediction systems.

2. BACKGROUND

2.1 Customer Churn Prediction

Customer churn prediction is formally framed as a supervised binary classification problem. Given a dataset $\mathcal{D} = \{(\mathbf{x}i, yi)\}^N i=1$, where $\mathbf{x}i \in \mathbb{R}^d$ is a feature vector encoding customer attributes and $yi \in \{0,1\}$ is the churn label, the objective is to learn a mapping f: $\mathbb{R}^d \rightarrow [0,1]$ estimating the probability of churn. The standard training objective minimizes the regularized binary cross-entropy:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[yi\log f\theta(\mathbf{x}i) + (1 - yi)\log(1 - f\theta(\mathbf{x}i))] + \lambda\,\Omega(\theta)$$

where $\Omega(\theta)$ is a regularization term and λ controls regularization strength. A customer is classified as high-churnrisk when $f\theta(\mathbf{x}i) \geq \tau$, with threshold τ calibrated according to the business objective frequently the expected maximum profit criterion [14].

Churn datasets exhibit persistent structural characteristics that complicate modeling and explanation. Class imbalance is endemic: reported churn rates range from 2–5% in mature subscription services to 26–30% in competitive telecom markets [5], [15]. This imbalance necessitates resampling strategies such as SMOTE [16] or

Journal of Information Systems Engineering and Management
2025, 10(63s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ Research Article

costsensitive learning. Feature heterogeneity continuous billing variables, binary service flags, ordinal contract types, and temporal usage sequences requires careful distance metric design [17]. Furthermore, many features relevant to churn prediction are immutable within the intervention horizon: a customer's age, geography, or account opening date cannot be altered by a retention intervention, rendering counterfactuals that modify such features operationally meaningless [18].

The literature documents churn prediction across four primary domains. In telecommunications, Huang et al. [4] and Verbeke et al. [14] established benchmarks using gradient boosting and profit-driven evaluation, respectively. In banking, Zhu et al. [19] applied interpretable models to credit-linked churn, while Lalwani et al. [20] examined ensemble approaches for retail banking attrition. In subscription and SaaS contexts, Backman and Bhatt [21] analyzed behavioral feature engineering for subscription media, while Droftina et al. [22] addressed B2B SaaS churn with survival analysis. These domain studies consistently demonstrate that explanation mechanisms remain an open practical challenge even where predictive accuracy is high.

2.2 Explainable Artificial Intelligence

The field of XAI encompasses techniques that render machine learning model behavior interpretable to human stakeholders [23], [24]. The taxonomy of XAI methods distinguishes global explanations (characterizing model behavior across the input space) from local explanations (explaining individual predictions), and ante-hoc interpretable models from post-hoc explanation methods applied after training [25]. In churn prediction, post-hoc local explanations are most operationally relevant, enabling per-customer decision-making without constraining model selection.

SHAP [9] and LIME [10] are the most widely deployed post-hoc local methods in churn analytics, with SHAP grounded in Shapley values from cooperative game theory and LIME approximating local model behavior with interpretable surrogates. While both provide useful feature importance rankings, they do not inherently provide actionable recourse — information about what could be changed to obtain a different outcome [11]. This limitation motivates counterfactual explanations as a complementary and often superior mechanism for actionable decision support [62].

2.3 Counterfactual Explanations: Foundational Concepts

The concept of counterfactual explanation as applied to machine learning was formalized by Wachter, Mittelstadt, and Russell [11], drawing on the philosophical tradition of counterfactual reasoning in causation theory. A counterfactual explanation for a prediction $f(\mathbf{x}) = y$ is an alternative instance $\mathbf{x}'$ such that $f(\mathbf{x}') = y' \neq y$, where $\mathbf{x}'$ is as similar to $\mathbf{x}$ as possible. Wachter et al. operationalized this as the constrained optimization: $\mathbf{x}^* = \underset{\mathbf{x}'}{\operatorname{argmin}}\ \ell(f(\mathbf{x}'), y') + \mu \cdot d(\mathbf{x}, \mathbf{x}')$

where $\ell(\cdot,\cdot)$ is a loss measuring prediction distance from the target class y', $d(\mathbf{x}, \mathbf{x}')$ is a distance metric, and μ governs the tradeoff between prediction validity and feature proximity. This work was seminal in establishing the legal and technical case for counterfactual explanations under GDPR, arguing that such explanations constitute the "minimum information necessary" for meaningful automated decision explanations [11]. Subsequent literature has substantially extended the foundational formulation across multiple dimensions: diversity [27], plausibility [28], causal validity [29], sparsity [30], and fairness [31], and churn-specific actionable recourse generation [62].

3. COUNTERFACTUAL EXPLANATION METHODOLOGIES

3.1 Wachter et al. Proximity-Constrained Optimization

The method proposed by Wachter et al. [11] remains the canonical baseline in the counterfactual explanation literature. It minimizes the objective in Equation (2) using gradient-based optimization (L-BFGS) with a differentiable proxy for the prediction validity constraint. The distance metric employs feature-wise median absolute deviation (MAD) normalization:

Journal of Information Systems Engineering and Management
2025, 10(63s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ **Research Article**

$$d(\mathbf{x}, \mathbf{x}') = \sum_{j \in \mathcal{C}} \frac{(x_j - x'_j)^2}{\text{MAD}_j^2} + \sum_{j \in \mathcal{B}} \mathbb{1}_{[xj \neq x'j]}$$

where $\mathcal{C}$ and $\mathcal{B}$ denote continuous and binary feature subsets, and $\text{MAD}j = \text{median}i(|xij - \text{median}k(xkj)|)$ provides scale normalization robust to outliers. The primary strengths of Wachter counterfactuals are model-agnosticism and computational efficiency, as the method requires only differentiable access to the model's output probability. Its principal limitations are the generation of only a single counterfactual per instance (no diversity), absence of distributional plausibility constraints, and lack of native support for feature immutability [27], [28], [32].

3.2 Diverse Counterfactual Explanations (DiCE)

Mothilal, Sharma, and Tan [27] introduced DiCE to address the diversity limitation of Wachter's formulation. DiCE generates a set of K counterfactuals by augmenting the optimization objective with a diversity term based on the determinantal point process (DPP):

$$\min_{\{\mathbf{x}'k\}k=1}{}^{K} \sum_{k=1}^{K} \ell\,(f(\mathbf{x}'k), y') + \mu 1 \sum_{k=1}^{K} d\,(\mathbf{x}, \mathbf{x}'k) - \mu 2 \text{logdet}\mathbf{K}$$

where $\mathbf{K}$ is the kernel matrix with entries $Kij = \kappa(\mathbf{x}'i, \mathbf{x}'j)$ encoding pairwise similarity among generated counterfactuals, and $\mu 1$, $\mu 2$ control the validity-proximity-diversity tradeoff. The log-determinant term penalizes near-duplicate counterfactuals, encouraging exploration of diverse regions of the desired-class space. DiCE additionally supports feature immutability constraints and feature range bounds, both critical for churn prediction deployments where demographic and contractual features may be non-actionable. Empirical evaluations in [27] and replications in [32] confirm that DiCE produces substantially more diverse explanation sets than singlecounterfactual methods, at the cost of moderate computation increases and some plausibility reduction relative to prototype-guided methods.

3.3 Contrastive Explanations Method (CEM)

Dhurandhar et al. [30] proposed CEM to address the sparsity and interpretability dimensions of counterfactual explanation. CEM frames the explanation problem in terms of pertinent negatives the minimal set of features that, when modified, changes the classification and pertinent positives. The pertinent negative optimization solves:

$$\min_{\boldsymbol{\delta}} c \cdot \ell(f(\mathbf{x} + \boldsymbol{\delta}), y') + \lambda 1 \,\|\, \boldsymbol{\delta} \,\|1 + \lambda 2 \,\|\, \boldsymbol{\delta} \,\|^2 2 + \lambda 3 \cdot \text{AE}(\mathbf{x} + \boldsymbol{\delta})$$

where $\boldsymbol{\delta}$ is the perturbation vector, $\|\, \boldsymbol{\delta} \,\|1$ promotes sparsity by encouraging perturbation components to be exactly zero, $\|\, \boldsymbol{\delta} \,\|^2 2$ bounds the total perturbation magnitude, and $\text{AE}(\cdot)$ is an autoencoder reconstruction loss penalizing deviations from the training data manifold. CEM's $\ell 1$ regularization is the key mechanism enabling sparse explanations. In churn prediction applications, CEM's pertinent negatives frequently identify one to three features as decisive, enabling concise communications of the form: "switching to an annual contract would change your risk classification" [30], [34]. Benchmarks in [32] and [35] confirm that CEM achieves higher sparsity scores than DiCE or Wachter counterfactuals.

3.4 Prototype-Based Counterfactuals

Van Looveren and Klaise [28] introduced prototype-based counterfactuals to directly address the out-of-distribution plausibility problem in Wachter-style optimization. Their method incorporates class prototype information learned from the training distribution:

$$\min_{\mathbf{x}'} c1 \cdot \ell(f(\mathbf{x}'), y') + c2 \cdot \|\, \mathbf{x}' - \mathbf{x} \,\|^2 2 + c3 \cdot \|\, \text{Enc}(\mathbf{x}') - \text{proto}y' \,\|^2 2 + c4 \cdot \text{AE}(\mathbf{x}')$$

Journal of Information Systems Engineering and Management
2025, 10(63s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ **Research Article**

where Enc(·) is an encoder mapping to a latent space, $\text{proto}_{y'} = \frac{1}{}\sum_{x_i \in |\mathcal{D}y'|} \mathcal{D}y'$ Enc $(\mathbf{x}i)$ is the prototype of the desired class in the latent space, and $\text{AE}(\mathbf{x}')$ is an autoencoder reconstruction loss. The prototype term pulls counterfactuals toward the centroid of desired-class instances in the latent space, substantially reducing the probability of generating implausible feature combinations. Van Looveren and Klaise [28] report that prototype counterfactuals achieve reconstruction losses 40–60% lower than Wachter counterfactuals on the same datasets, indicating substantially greater distributional realism. The tradeoff is that manifold constraints may exclude valid counterfactuals lying off the training distribution, reducing validity relative to unconstrained methods [35].

3.5 Generative Counterfactual Methods (REVISE)

Joshi et al. [36] proposed REVISE, a generative counterfactual method that uses a variational autoencoder (VAE) to project instances into a continuous latent space where optimization is performed. Rather than searching the original feature space, REVISE solves:

$$\min_{\mathbf{z}} \ell(f(\text{Dec}(\mathbf{z})), y') + \mu \cdot \| \mathbf{z} - \text{Enc}(\mathbf{x}) \|^2 2$$

where Enc(·) and Dec(·) are the VAE encoder and decoder, $\mathbf{z}$ is the latent vector, and the counterfactual in original feature space is recovered as $\mathbf{x}' = \text{Dec}(\mathbf{z}^*)$. The VAE's training objective enforces that decoded points are approximately drawn from the training distribution, providing a distributional realism guarantee by construction [36]. REVISE has been evaluated on tabular classification tasks consistently demonstrating higher plausibility scores than gradient-based methods [36], [37]. The method's primary limitation for churn prediction deployment is computational: VAE training and latent-space optimization require GPU infrastructure for practical generation speeds [37].

3.6 Gradient-Based Optimization (GradCF)

Gradient-based counterfactual generation applies direct gradient ascent on differentiable classifiers by iteratively modifying the input toward the target class:

$$\mathbf{x}(t+1) = \Pi\mathcal{F}(\mathbf{x}(t) - \alpha \cdot \nabla\mathbf{x}\ell(f(\mathbf{x}(t)), y'))$$

where α is the step size and $\Pi\mathcal{F}$ denotes projection onto the feasible feature set $\mathcal{F}$ defined by bounds and immutability constraints. This approach, studied in the context of adversarial examples [38] and adapted for counterfactual generation in [39], achieves high validity and minimal computation time for differentiable models. The critical weakness of unconstrained GradCF is its susceptibility to producing adversarially-style counterfactuals: technically valid instances that exploit model artifacts, lie off the data manifold, and are realistically implausible as actual customer profiles [39]. Studies comparing GradCF to prototype-based and generative methods consistently find substantially lower plausibility and robustness [35], [40]. Presenting customers with implausible counterfactuals destroys trust and reduces adoption of retention recommendations [33].

4. EVALUATION DIMENSIONS FOR COUNTERFACTUAL EXPLANATIONS

The literature has progressively converged on a set of quality dimensions for evaluating counterfactual explanations, with contributions from Verma et al. [41], Guidotti [42], Carvalho et al. [43], and Albini et al. [44].

4.1 Validity

Validity measures the proportion of generated counterfactuals that achieve the desired class prediction. For a set of K counterfactuals generated for instance $\mathbf{x}$:

$$V = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}[f(\mathbf{x}'k) = y']$$

Journal of Information Systems Engineering and Management
2025, 10(63s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ **Research Article**

Validity is the most fundamental requirement: a counterfactual that does not flip the prediction is technically invalid [41]. However, high validity alone is insufficient — an adversarial example achieves perfect validity while failing all other quality criteria [38].

4.2 Proximity

Proximity quantifies feature-level closeness between an original instance and its counterfactual using a mixed-type distance metric [27]:

$$\text{Prox}(\mathbf{x}, \mathbf{x}') = \underset{\text{MAD}j}{\mathbb{1}[xj \neq x'j]} \quad \sum_{j \in \mathcal{C}} \frac{|x_j - x'_j|}{} + \sum_{j \in \mathcal{B}}$$

Lower proximity values indicate counterfactuals closer to the original instance. Proximity and validity are in tension: proximity constraints restrict the search space, potentially excluding valid regions [41].

4.3 Sparsity

Sparsity, also termed compactness [42], measures the number of features changed in the counterfactual:

$$\mathcal{S}(\mathbf{x}, \mathbf{x}') = 1 - \frac{\| \mathbf{x} - \mathbf{x}' \|_0}{d}$$

where $\|\cdot\|_0$ is the $\ell 0$ pseudo-norm (count of non-zero differences) and d is total feature dimensionality. Human factors research reviewed by Miller [45] confirms that explanations involving fewer changed variables are substantially easier to understand and act upon, motivating sparsity as a first-class quality criterion.

4.4 Diversity

Diversity, introduced by Mothilal et al. [27], quantifies dissimilarity among the K counterfactuals generated for a single instance:

$$D = \frac{2}{K(K-1)} \sum_{i=1}^{K} \sum_{j>i}^{K} d(\mathbf{x}'i, \mathbf{x}'j)$$

Higher diversity provides decision-makers with a richer menu of actionable options, accommodating heterogeneous customer constraints and preferences [27], [33].

4.5 Plausibility

Plausibility measures the extent to which generated counterfactuals resemble actual instances from the training distribution. Density-based measures using kernel density estimation are used in evaluation [43]:

$$\text{Pl}(\mathbf{x}') = p\text{KDE}(\mathbf{x}')$$

where $p\text{KDE}$ is a kernel density estimator fitted on training data. Implausible counterfactuals that suggest impossible or internally inconsistent feature combinations reduce user trust and are legally problematic in customer-facing systems [11], [46].

4.6 Robustness

Robustness, formalized by Alvarez-Melis and Jaakkola [47] and extended to counterfactuals by Slack et al. [48], measures explanation stability under small input perturbations:

$$\text{Rob}(\mathbf{x}) = 1 - \mathbb{E}_{\tilde{\mathbf{x}} \in \mathcal{B}(\mathbf{x}, \epsilon)} \left[\frac{d(\mathbf{x}^*(\mathbf{x}), \mathbf{x}^*(\tilde{\mathbf{x}}))}{} \right] \epsilon$$

Journal of Information Systems Engineering and Management
2025, 10(63s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ Research Article

where $\mathcal{B}(\mathbf{x}, \epsilon) = \{\mathbf{x} + \boldsymbol{\eta}: \| \boldsymbol{\eta} \|2 \leq \epsilon\}$ is the perturbation ball. Slack et al. [48] demonstrated that low-robustness explanations can be exploited to generate systematically misleading explanations for auditors a critical fairness and governance concern.

4.7 Actionability and Feasibility

Actionability requires that counterfactual feature changes be realistically achievable within the intervention context [18], [29]. In churn prediction, recent work has further explored actionability through t-way sampling with IPOG to generate feasible combinations of feature changes for customer recourse [62]. Ustun et al. [18] introduced this as a formal constraint in the algorithmic recourse literature, distinguishing mutable features (payment method, contract tier) from immutable features (age, account open date, geographic region). Karimi et al. [29] further decompose actionability into causal feasibility whether the suggested change sequence respects causal relationships among features distinguishing direct actions from confounded observational correlations.

5. RESEARCH GAP

Analysis of the surveyed literature reveals three critical gaps that motivate this comparative study.

Gap 1: Absence of Domain-Specific Comparative Evaluation. The most comprehensive comparative studies of counterfactual methods [41], [42], [32] evaluate methods on generic classification benchmarks (Adult Income, German Credit, COMPAS) without systematic attention to churn-specific data characteristics. Customer churn datasets exhibit structural properties — high class imbalance, heterogeneous mixed-type features, temporal behavioral sequences, and legally significant feature immutability constraints — that materially affect the relative performance of counterfactual methods. The absence of churn-domain evaluation in foundational comparative studies limits the direct applicability of their findings to practitioners.

Gap 2: Incomplete Multi-Criteria Assessment. Existing comparative studies typically evaluate two to four quality dimensions. Wachter et al. [11] focused on validity and proximity; Mothilal et al. [27] added diversity; Van Looveren and Klaise [28] added plausibility. No single study synthesizes all seven dimensions validity, proximity, sparsity, diversity, plausibility, robustness, and actionability within a unified comparative framework grounded in the churn prediction literature.

Gap 3: Under integration of Trustworthiness and Fairness. The intersection of counterfactual explanations with broader trustworthy AI principles particularly fairness in recourse availability across demographic groups has been studied separately in the algorithmic fairness literature [31], [49] but has not been systematically integrated into comparative surveys of churn prediction explainability. Given that churn prediction models interact with protected attributes in complex ways, this integration is ethically and legally necessary.

6. Comparative Analysis of Counterfactual Methods

6.1 Summary of Methods and Key Properties

Table 1 presents a structured comparison of the six surveyed counterfactual methods across key methodological properties, synthesized from the original papers and replications in the literature.

Table 1: Comparative Analysis of Counterfactual Explanation Methods — Methodological Properties

Method	Optimization	Diversity Support	Plausibility Mechanism	Immutability Support	Model Requirement
Wachter et al. [11]	L-BFGS gradient descent	None (single CF)	None	Manual post-hoc	Differentiable or black-box

Journal of Information Systems Engineering and Management
2025, 10(63s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ Research Article

DiCE [27]	DPP-regularized gradient	DPP diversity term	None native	Built-in constraint	Differentiable or black-box
CEM [30]	Elasticnet perturbation	None (single CF)	Autoencoder loss	Manual post-hoc	Differentiable
Prototype CF [28]	Latent prototype guidance	None native	Prototype + AE loss	Manual post-hoc	Differentiable
REVISE [36]	VAE latent space search	Moderate (stochastic)	VAE generative prior	Built-in constraint	Differentiable
GradCF [39]	Direct gradient ascent	None	None	Projection constraint	Differentiable

6.2 Evaluation Across Quality Dimensions

Table 2 synthesizes quality dimension evaluations from across the literature. Ratings (High/Moderate/Low) are derived from reported results in the cited primary and comparative studies [27], [28], [30], [32], [35], [36], [40], [41], [62].

Table 2: Comparative Evaluation of Counterfactual Methods Across Quality Dimensions (Synthesized from Literature [27], [28], [30], [32], [35], [36], [40], [41, [62]])

Method	Validity	Proximity	Sparsity	Diversity	Plausibility	Robustness	Computational Cost
Wachter et al. [11]	High	Moderate	Moderate	Low	Low	Moderate	Low
DiCE [27]	High	Moderate	Moderate	High	Moderate	Moderate	Moderate
CEM [30]	Moderate–High	High	**High**	Low	Moderate–High	Moderate–High	Moderate
Prototype CF [28]	Moderate	**High**	Moderate–High	Low	**High**	**High**	Moderate–High
REVISE [36]	Moderate	Moderate	Moderate	Moderate	**High**	High	High
GradCF [39]	High	Low–Moderate	Low	Low	Low	Low	Low

Journal of Information Systems Engineering and Management
2025, 10(63s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ **Research Article**

Note: "Proximity: High" indicates low distance from original (closer). Bold denotes best performance per column. Ratings synthesized from cited comparative studies.

Validity. Gradient-based methods (Wachter, DiCE, GradCF) consistently achieve the highest validity rates in the literature, exceeding 90% in tabular classification benchmarks [32], [41]. Methods incorporating strong manifold constraints (Prototype CF, REVISE) exhibit lower validity, with Pawelczyk et al. [32] reporting reductions of 8–15% relative to unconstrained methods. CEM occupies an intermediate position dependent on the autoencoder constraint strength [30].

Proximity. CEM and Prototype CF achieve the best proximity scores, as their respective $\ell 1$ and prototype-distance regularization terms directly penalize feature-space distance [28], [30]. GradCF, lacking a proximity regularizer, frequently produces counterfactuals at larger distances from the original [39].

Sparsity. CEM's $\ell 1$ penalty is the most effective sparsity mechanism documented in the literature, producing a median of one to two feature changes in Dhurandhar et al.'s original evaluation [30] and confirmed by Carvalho et al. [43]. DiCE and GradCF, which lack explicit sparsity constraints, change a median of 4–7 features in tabular benchmarks [32], substantially reducing interpretability.

Diversity. DiCE's DPP-based diversity term produces demonstrably more varied counterfactual sets than all other methods, with Mothilal et al. [27] reporting pairwise counterfactual distances 3–5× higher than single-counterfactual methods. REVISE achieves moderate diversity through stochastic VAE decoding [36].

Plausibility. Prototype CF and REVISE consistently achieve the highest plausibility scores in evaluations using density estimation metrics [28], [36], [40]. Van Looveren and Klaise [28] report reconstruction losses 40–60% lower than Wachter counterfactuals on the same datasets, indicating substantially greater distributional realism.

Robustness. Prototype CF and REVISE exhibit the highest explanation robustness, as their manifold constraints stabilize the optimization landscape [48], [50]. GradCF's adversarial-style search produces explanations highly sensitive to small input perturbations [48].

Computational Cost. Wachter and GradCF are the fastest methods due to their simple optimization landscapes and absence of auxiliary model training [32]. REVISE incurs the highest cost due to VAE training requirements [36].

7. DOMAIN-SPECIFIC CHALLENGES IN CHURN PREDICTION

7.1 Class Imbalance and Counterfactual Quality

Class imbalance characteristic of virtually all churn datasets interacts with counterfactual explanation quality in ways insufficiently studied in the existing literature. When the desired target class (retained customers) constitutes 70–98% of the training distribution, prototype-based methods may compute biased prototypes dominated by the majority-class geometry [15], [28]. Lundberg et al. [51] note that SHAP values are systematically affected by class imbalance; analogous analysis for counterfactual proximity and plausibility metrics has been called for by Rawal and Ghassemi [52] but remains an open research problem. Practical deployments typically apply class weights or SMOTE oversampling before training both the churn classifier and any auxiliary explanation models [16].

7.2 Feature Immutability and Actionable Recourse

Feature immutability the distinction between features a retention intervention can change and those that are fixed is a critical practical constraint that several counterfactual methods address inadequately. Wachter et al. [11] acknowledged immutability as a limitation of their original formulation. Ustun et al. [18] formalized this as the "actionable recourse" problem, requiring that all suggested feature changes be actionable and feasible. In the telecom churn context, Keane and Smyth [53] identified that naively generated counterfactuals frequently suggest modifications to contract start dates or geographic location features that cannot be changed within the intervention horizon. DiCE natively supports immutability constraints as a hard constraint in its optimization [27], while other methods require post-hoc filtering that can substantially reduce counterfactual count and degrade quality dimensions [32].

Journal of Information Systems Engineering and Management
2025, 10(63s) e-ISSN:
2468-4376
https://www.jisem-journal.com/ Research Article

7.3 Temporal and Sequential Feature Structure

A dimension underrepresented in the counterfactual explanation literature is the temporal structure of customer behavior data. Churn prediction models that incorporate recurrent neural networks or temporal attention mechanisms [6] introduce complications for counterfactual optimization: the feature space is no longer a static vector but a sequence, and "changing" a feature involves reasoning about time-indexed interventions. Delaney et al. [54] proposed counterfactual explanations for time-series classifiers using a native guide approach, but the specific application to behavioral churn sequences remains largely unaddressed a meaningful methodological gap between the academic counterfactual literature and production churn models.

8. TRUSTWORTHINESS, FAIRNESS, AND REGULATORY ALIGNMENT

8.1 Trustworthy AI Principles

The European Commission's High-Level Expert Group on Artificial Intelligence identified seven trustworthy AI requirements [55]: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and non-discrimination, societal and environmental well-being, and accountability. Counterfactual explanations directly address transparency and human agency. Technical robustness operationalized by the robustness score in Equation (9) shows significant variation across methods [47], [48]. Transparency requires not only that explanations be generated but that they be comprehensible to stakeholders of varying technical literacy, motivating sparsity as a proxy for human understandability [45], [53].

8.2 Fairness in Algorithmic Recourse

A critical and rapidly developing area of the trustworthy AI literature concerns fairness in the distribution of counterfactual recourse across demographic groups. Ustun et al. [18] demonstrated that algorithmic recourse can be systematically less available to members of protected groups even when the underlying classifier satisfies standard fairness criteria. If members of a protected group require larger feature changes to achieve a favorable classification, the proximity cost of recourse is disparately distributed a form of fairness violation not captured by prediction-level fairness metrics.

Karimi et al. [29] extended this analysis within a structural causal model (SCM) framework, showing that standard counterfactual methods can suggest recourse paths that are causally invalid. In banking churn prediction, this has direct legal significance as credit-related recommendations interact with fair lending regulations [49]. Gupta et al.
[49] demonstrated that jointly enforcing causal validity and demographic parity of recourse cost substantially narrows but does not eliminate the recourse disparity gap, underscoring that fairness-aware counterfactual generation requires both appropriate mathematical formulation and domain-specific causal knowledge.

8.3 Regulatory Alignment

The GDPR Article 22 right to explanation has been interpreted by Wachter et al. [11] and Edwards and Veale [46] as consistent with, and potentially requiring, counterfactual explanations constituting the minimum viable explanation without requiring disclosure of the model's internal architecture. Subsequent legal analysis has refined this position, noting that GDPR's "meaningful information" standard requires comprehensibility to ordinary individuals a requirement mapping most naturally to CEM's sparse single-feature explanations [56]. The proposed EU AI Act's Article 13 transparency requirements for high-risk systems go further, requiring explanations that enable human oversight of individual decisions, with counterfactual explanations specifically cited in interpretive documents [57]. The EU AI Act's robustness requirement additionally motivates the robustness evaluation dimension: explanations must remain consistent under minor distributional shifts [8, 62].

9. PRACTICAL IMPLICATIONS FOR CHURN ANALYTICS

9.1 Method Selection by Deployment Context

The comparative analysis in Section 6 reveals that no single counterfactual method is universally optimal, and that method selection must be governed by the specific operational requirements of the deployment context. Three primary deployment archetypes can be distinguished from the literature.

Real-Time Agent-Assisted Retention. Customer service agents interacting with customers in real time require

explanations generated within seconds in concise, immediately actionable formats. For this context, Wachter counterfactuals or CEM are most suitable — the former for computational speed and the latter for sparsity [11], [30], [53]. DiCE's moderate computational cost may be acceptable for near-real-time batch pre-computation.

Batch Overnight Retention Scoring. Nightly batch pipelines processing thousands of at-risk customers can accommodate higher computational costs in exchange for richer explanation quality. DiCE is the most appropriate method for this context, providing diverse explanation sets enabling retention managers to select from multiple intervention options adapted to each customer's situation [27], [33].

Regulatory Compliance and Audit. In regulated sectors where explanations may be reviewed by regulators or presented in legal proceedings, plausibility and robustness are paramount. Prototype CF or REVISE provide the most legally defensible explanations as their manifold-constrained counterfactuals represent demonstrably realistic customer profiles rather than adversarially constructed artifacts [28], [36], [46].

9.2 Integration with Business Processes

The practical integration of counterfactual explanations into churn management involves challenges beyond technical quality. Keane and Smyth [53] conducted a user study with domain experts evaluating counterfactual explanations for churn, finding that sparse explanations (1–2 feature changes) were significantly preferred over comprehensive explanations (5+ feature changes), even when the latter more precisely characterized the decision boundary. Tsirtsis and Rodriguez [58] similarly found that user-perceived actionability was more strongly predicted by sparsity than by proximity or plausibility. Arya et al. [59] documented organizational challenges in deploying explainable AI in banking, noting that explanation systems must be integrated with CRM platforms, calibrated to business intervention logic, and maintained as the underlying model is retrained a temporal stability dimension largely unaddressed in the methodology literature.

10. LIMITATIONS OF THE SURVEYED LITERATURE

Several important limitations in the existing body of counterfactual explanation literature merit acknowledgment.

Benchmark Generalizability. The majority of methodological papers evaluate counterfactual methods on the UCI Adult Income, German Credit, or COMPAS datasets [11], [27], [28], [30]. These datasets differ from production churn datasets in dimensionality, feature types, class balance, and the availability of domain-specific constraint knowledge. The extent to which benchmark results generalize to telecom, banking, or SaaS churn scenarios remains inadequately validated [32].

Absence of Longitudinal Evaluation. All surveyed methods generate counterfactuals from static cross-sectional customer snapshots. Real churn prediction systems operate on streaming data where customer behavior evolves continuously. The validity of generated counterfactuals over time as the model is retrained and customer features evolve has not been systematically studied [54].

User Study Coverage. Despite methodological papers citing interpretability and comprehensibility as motivating goals, only a small subset of the surveyed literature includes user studies validating that target stakeholder's retention agents, customers, or compliance officers find generated explanations useful and understandable [45], [53], [58]. Objective quality metrics do not reliably proxy for human judgment of explanation quality [45].

Causal Completeness. Standard counterfactual methods, including DiCE and CEM, operate on observational features without modeling causal relationships. The generated counterfactuals may recommend changes that are observationally associated with retention but causally inert or counterproductive under intervention [29]. Causal counterfactual methods [29], [60] address this gap but require causal graph knowledge not routinely available in commercial deployments.

11. FUTURE RESEARCH DIRECTIONS

Causal Counterfactuals for Churn. The integration of structural causal models into counterfactual optimization for churn prediction following the framework of Karimi et al. [29] is a pressing research need. Domain-specific causal graphs for telecom, banking, and SaaS churn can be partially specified from business domain knowledge and validated through expert elicitation. Hybrid approaches combining data-driven causal discovery with expertspecified constraints represent a tractable path to causal counterfactual churn explanations.

Fairness-Aware Counterfactual Generation. The recourse fairness literature [18], [31], [49] has established that standard counterfactual methods produce disparate recourse costs across demographic groups. Developing fairness-

constrained counterfactual optimization calibrated to the feature structures and protected attribute interactions characteristic of churn datasets is a critical research need with direct regulatory relevance.

Temporal Counterfactuals for Sequential Churn Models. As churn prediction increasingly incorporates temporal behavioral features and sequential models [6], methodology for generating counterfactuals that suggest realistic time-indexed intervention sequences must be developed. Delaney et al.'s [54] work on time-series counterfactuals provides a foundation, but extension to the mixed tabular-sequential feature spaces common in churn datasets requires further investigation.

Federated Counterfactual Learning. Privacy regulations increasingly restrict centralization of customer data, motivating federated approaches to both churn model training and explanation generation. Federated learning frameworks [61, 62] for training counterfactual explanation models without sharing individual customer records represent an underexplored direction with significant practical relevance for multi-entity organizations.

Standardized Churn-Domain Evaluation Benchmarks. The field would benefit substantially from standardized churn-domain evaluation benchmarks that encode realistic immutability constraints, provide groundtruth causal structures where possible, and include human preference annotations from domain expert user studies analogous to the COMPAS and Adult Income datasets used in general counterfactual evaluation.

12. CONCLUSION

This paper has presented a comprehensive comparative survey of six counterfactual explanation techniques Wachter proximity-constrained optimization, DiCE, CEM, prototype-based counterfactuals, REVISE, and gradient-based optimization as applied to customer churn prediction, grounded entirely in published literature from 2015 to 2025. The analysis spans telecommunications, banking, subscription services, and SaaS contexts, synthesizing methodological contributions, comparative evaluations, and domain-specific applications from across the literature.

The comparative analysis across seven quality dimensions reveals that no single method achieves universal superiority. Wachter counterfactuals and GradCF lead on validity and computational efficiency; CEM excels in sparsity, making it most suitable for customer-facing communications; DiCE offers the strongest diversity, enabling multi-option retention strategy generation; and prototype-based methods and REVISE achieve the highest plausibility and robustness, making them most appropriate for regulated environments where explanation realism carries legal significance. These tradeoffs are fundamental properties of the underlying mathematical objectives, not implementation artifacts.
The survey further documents significant domain-specific challenges unique to churn prediction: class imbalance effects on explanation quality, feature immutability constraints that render many generated counterfactuals operationally void, temporal structure in behavioral data that current counterfactual methods do not adequately address, and fairness implications of differential recourse availability across customer demographic groups. These challenges define a rich agenda for future research at the intersection of explainability, causal inference, and ethical AI.

As regulatory frameworks including GDPR Article 22 and the EU AI Act increasingly mandate explainability for automated customer decisions, the selection and deployment of appropriate counterfactual explanation methods will transition from an academic question to an organizational governance imperative. This survey provides the analytical foundation for informed, context-sensitive method selection by researchers advancing the state of the art and practitioners deploying trustworthy churn prediction systems.

REFERENCES

[1] V. Kumar and W. J. Reinartz, *Customer Relationship Management: Concept, Strategy, and Tools*, 3rd ed. Berlin, Germany: Springer, 2018.

[2] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and solutions," in *Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD)*, New York, NY, USA, 1998, pp. 73–79.

[3] G. Mozer, R. Wolniewicz, D. B. Grimes, E. Johnson, and H. Kaushansky, "Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry," *IEEE Trans. Neural Networks*, vol. 11, no. 3, pp. 690–696, May 2000.

[4] B. Huang, M. T. Kechadi, and B. Buckley, "Customer churn prediction in telecommunications," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1414–1425, 2012.

[5] A. Amin, S. Anwar, A. Adnan, M. Nawaz, N. Howard, J. Qadir, A. Hawalah, and A. Hussain, "Comparing oversampling techniques to handle the class imbalance problem for healthcare decision making," *IEEE Access*, vol. 4, pp. 7940–7957, 2016.

[6] Q. Liang, Z. Rong, J. Zhang, J. Liu, and Z. Xiong, "Customer churn prediction using a hierarchical ensemble of tree models," *IEEE Access*, vol. 9, pp. 65175–65185, 2021.

[7] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 — General Data Protection Regulation," *Official Journal of the European Union*, Apr. 2016.

[8] European Commission, "Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)," COM/2021/206 final, Apr. 2021.

[9] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774.

[10] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[11] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.

[12] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[13] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *Data Mining and Knowledge Discovery*, vol. 37, pp. 1719–1778, 2023.

[14] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211–229, 2012.

[15] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *European Journal of Operational Research*, vol. 269, no. 2, pp. 760–772, 2018.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[17] A. De Caigny, K. Coussement, K. W. De Bock, and S. Lessmann, "Incorporating textual information in customer churn prediction models based on a convolutional neural network," *Int. Journal of Forecasting*, vol. 36, no. 4, pp. 1563–1578, 2020.

[18] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, 2019, pp. 10–19.

[19] X. Zhu, H. Li, and S. Wang, "Interpretable customer churn prediction using SHAP values in banking," in *Proc. IEEE Int. Conf. Big Data*, 2021, pp. 4033–4038.

[20] P. Lalwani, M. K. Mishra, J. S. Chadha, and P. Sethi, "Customer churn prediction system: A machine learning approach," *Computing*, vol. 104, no. 2, pp. 271–294, 2022.

[21] J. Backman and N. Bhatt, "Predicting subscription churn for a freemium streaming service using behavioral features," in *Proc. IEEE Int. Conf. Data Science and Advanced Analytics (DSAA)*, 2020, pp. 598–607.

[22] U. Droftina, M. Stukelj, and K. Koncnik, "Churn prediction in subscription-based B2B SaaS with survival analysis," in *Proc. Int. Conf. Information Society (i-Society)*, 2015, pp. 24–29.

[23] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "XAI — Explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, 2019.

[24] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[25] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd ed. Munich, Germany: Leanpub, 2022.

[26] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[27] R. K. Mothilal, A. Sharma, and D. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, 2020, pp. 607–617.

[28] A. Van Looveren and J. Klaise, "Interpretable counterfactual explanations guided by prototypes," in *Proc. ECML-PKDD*, 2021, pp. 650–665.

[29] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse: From counterfactual explanations to interventions," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, 2021, pp. 353–362.

[30] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, A. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018, pp. 592–603.

[31] S. Venkatasubramanian and M. Alfano, "The philosophical basis of algorithmic recourse," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, 2020, pp. 284–293.

[32] M. Pawelczyk, K. Broelemann, and G. Kasneci, "Learning model-agnostic counterfactual explanations for tabular data," in *Proc. The Web Conf. (WWW)*, 2020, pp. 3126–3132.

[33] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[34] P. Hall, N. Gill, and B. Cox, "Responsible machine learning: Actionable strategies for mitigating risks & driving adoption," O'Reilly Media, 2020.

[35] R. Guidotti, "Counterfactual explanations and how to find them: Literature review and benchmarking," *Data Mining and Knowledge Discovery*, vol. 38, pp. 2770–2824, 2024.

[36] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, "Towards realistic individual recourse and actionable explanations in black-box decision making systems," *arXiv preprint arXiv:1907.09615*, 2019.

[37] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2014.

[38] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2014.

[39] A. Fischetti and J. Jo, "Deep neural networks and tabular data: A survey," *IEEE Trans. Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 6499–6519, 2022.

[40] M. Pawelczyk, S. Bielawski, J. van den Heuvel, T. Richter, and G. Kasneci, "CARLA: A Python library to benchmark algorithmic recourse and counterfactual explanation algorithms," in *Proc. NeurIPS Workshop on Algorithmic Fairness through the Lens of Causality and Robustness*, 2021.

[41] S. Verma, J. Dickerson, and K. Hines, "Counterfactual explanations for machine learning: A review," *arXiv preprint arXiv:2010.10596*, 2020.

[42] R. Guidotti, "Counterfactual explanations and how to find them: Literature review and benchmarking," *Data Mining and Knowledge Discovery*, vol. 38, pp. 2770–2824, 2024.

[43] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[44] E. Albini, J. Long, D. Dervovic, and D. Magazzeni, "Counterfactual Shapley additive explanations," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, 2022, pp. 1054–1070.

[45] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[46] L. Edwards and M. Veale, "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for," *Duke Law & Technology Review*, vol. 16, no. 1, pp. 18–84, 2017.

[47] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," in *Proc. ICML Workshop on*

Human Interpretability in Machine Learning, 2018.

[48] D. Slack, A. Hilgard, S. Singh, H. Lakkaraju, and E. Strubell, "Reliable post hoc explanations: Modeling uncertainty in explainability," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 9391–9404.

[49] U. Gupta, A. T. Liu, and B. Ustun, "Algorithmic recourse under imperfect causal knowledge: A probabilistic approach," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 265–277.

[50] F. Hamman, E. Noorani, S. Mishra, D. Magazzeni, and S. Dutta, "Robust counterfactual explanations for neural networks with probabilistic guarantees," in *Proc. Int. Conf. Machine Learning (ICML)*, 2023.

[51] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv preprint arXiv:1802.03888*, 2018.

[52] K. Rawal and M. Ghassemi, "Beyond individualized recourse: Interpretable and interactive summaries of actionable recourse," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 12187–12198.

[53] M. T. Keane and B. Smyth, "Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI," in *Proc. Int. Conf. Case-Based Reasoning (ICCBR)*, 2020, pp. 163–178.

[54] E. Delaney, D. Greene, and M. T. Keane, "Instance-based counterfactual explanations for time series classification," in *Proc. Int. Conf. Case-Based Reasoning (ICCBR)*, 2021, pp. 32–47.

[55] High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI," European Commission, Brussels, Belgium, Apr. 2019.

[56] G. Malgieri and G. Comandé, "Why a right to legibility of automated decision-making exists in the general data protection regulation," *Int. Data Privacy Law*, vol. 7, no. 4, pp. 243–265, 2017.

[57] M. Brundage et al., "Toward trustworthy AI development: Mechanisms for supporting verifiable claims," *arXiv preprint arXiv:2004.07213*, 2020.

[58] S. Tsirtsis and M. Rodriguez, "Decisions, counterfactual explanations and strategic behavior," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 16749–16760.

[59] V. Arya et al., "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques," *arXiv preprint arXiv:1909.03012*, 2019.

[60] A.-H. Karimi, J. von Kügelgen, B. Schölkopf, and I. Valera, "Probabilistic and causal adaptations of algorithmic recourse," *arXiv preprint arXiv:2008.06677*, 2020.

[61] P. Kairouz et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[62] Khadka, S., Thapa, P., Sharma, P., Bhattarai, U., & Silwal, S. (2025). t-Recourse: Actionable counterfactual explanations for customer churn via t-way sampling with IPOG. *Available at SSRN 5441694*.