# Advanced Strategic Improvement with Decision Interactions

Wenjing Yang[1], Xinpeng Lv[1], Yunxin Mao[1], Liyang Xu[1], Ruochun Jin[1], Huan Chen[1], Jing Ren[1], Jinxuan Yang[3], Yuanlong Chen[2], and Haotian Wang[1] (✉)

[1] College of Computer Science and Technology, National University of Defense Technology, Changsha, China {lvxinpeng, wanghaotian13}@nudt.edu.cn
[2] Faculty of Computing, Harbin Institute of Technology, Harbin, China
[3] Faculty of Engineering, the University of Sydney, Sydney, Australia

**Abstract.** Strategic classification investigates the interaction between a decision-maker (modeled as a jury) and individuals (agents) who may strategically modify their features to obtain favorable outcomes. A key challenge in this setting is *strategic improvement*, which focuses on designing incentive mechanisms that encourage individuals to improve their true qualifications. In real-world scenarios, decision-making often involves multi-dimensional evaluations composed of multiple sub-indicators and a final comprehensive assessment. However, most existing paradigms for strategic classification rely on a single decision model, which is inadequate for capturing the complexity of such settings. To address this gap, we introduce the problem of **S**trategic **I**mprovement with **D**ecision **I**nteractions (*SIDI*), a novel setting that incorporates multiple interacting decision models and an overarching evaluation mechanism. We analyze the influence of decision interactions and reveal how correlations among classifiers can exacerbate manipulative behaviors. Building on these insights, we propose a decorrelation-based strategic improvement framework that leverages decision interactions to promote authentic qualification enhancements. Extensive experiments on both real-world and synthetic datasets demonstrate the effectiveness of our framework in encouraging genuine improvements while maintaining robust accuracy. Our findings highlight the importance of modeling decision interactions and provide new directions for strategic machine learning.

**Keywords:** Strategic Classification · Machine Learning · Decorrelation

## 1 Introduction

As machine learning-based decision making becomes increasingly prevalent, strategic classification [23] has garnered significant attention in recent years. Strategic classification addresses scenarios where individuals intentionally adapt their features to achieve desirable outcomes from intelligent decision systems. This phenomenon is evident in various domains such as credit scoring [1], hiring processes [41], and academic admissions [20]. In these contexts, individuals often

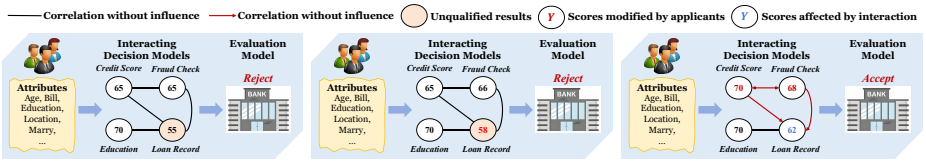



Fig. 1: An illustration of decision interaction. In traditional classification without strategic behavior, individuals fail to pass the loan application (*left*). With strategic behavior only towards a single model, the individual raises their score on the *loan record* metric, but it is still insufficient to pass the application (*center*). In decision interaction, the individual influences the *loan record* outcome by achieving high scores on other metrics, passing the loan approval (*right*).

manipulate their attributes to secure more favorable classifications, thereby challenging the fairness and integrity of these systems. For instance, in credit scoring, applicants might inflate their income or alter other financial metrics to attain a higher credit score.

As these strategic behaviors become more widespread, it is essential to understand and mitigate them in order to maintain the reliability of machine learning systems [37]. Early research primarily focused on developing classifiers that are robust to manipulation [51, 19]. While effective in detecting and resisting adversarial behaviors, such approaches often neglect the incentive structures that motivate individuals to game the system [36].

To overcome this limitation, the notion of *strategic improvement* has been proposed [4]. Instead of merely preventing manipulation, this paradigm aims to design incentive-compatible systems that encourage individuals to genuinely improve their underlying qualifications. For example, in college admissions, educators may wish to promote substantive academic preparation for standardized tests like the SAT, rather than rewarding superficial score inflation.

Existing studies in strategic machine learning have primarily focused on a single decision model. However, real-world strategic classification scenarios are far more diverse and complex. To ensure stability and accuracy, practical decision scenarios typically involve multi-dimensional evaluations of individuals, encompassing multiple sub-indicators and a final comprehensive assessment [2]. For example, in the financial services domain [33], decision makers deploy multiple classifiers (**sub-classifiers**) to evaluate an individual's creditworthiness, detect fraud, and approve loans, among other tasks. The final evaluation score is determined based on the results from these different indicators.

In multi-classifier decision scenarios, classifiers often exhibit significant interdependence [15]. As a result, a favorable outcome in one sub-classifier can influence the predictions of others. For example, in bank loan assessments, a positive evaluation of creditworthiness may affect the outcomes of fraud detection or loan approval components. We refer to this phenomenon as ***decision interaction***, which creates *additional opportunities for strategic manipulation* In particular, decision interaction enables individuals to exploit correlations be-

tween classifiers, creating *shortcut paths* to favorable final decisions that do not require genuine improvement across all relevant dimensions. For instance, individuals may selectively enhance their scores on sub-classifiers that are relatively **easy** to manipulate and less critical to the true outcome, while avoiding improvements on **harder but more important** indicators.

Returning to the bank loan example (see Fig. 1), an applicant who fails to meet the eligibility threshold—even after directly manipulating their score in the "loan repayment record" classifier—may still secure a loan. By recognizing the **interdependencies among decision models**, the applicant can strategically improve their scores on other, more manipulable classifiers. These improvements then propagate through the interaction structure, indirectly boosting their evaluation in the loan repayment metric and ultimately leading to approval. We refer to this form of manipulation, where individuals exploit model interdependencies to achieve favorable outcomes without genuine qualification enhancement, as ***interactive strategic behavior***.

Towards such unique strategic behaviors in interacting classifiers, we explore this new dimension, aiming to uncover the interactions between multiple decision models in this paper. Our work focuses on addressing two key questions in the context of decision interaction and interactive strategic behavior:

1. How to model strategic machine learning problems involving decision interaction in multi-classifier scenarios?
2. How to mitigate the impact of interactive strategic behavior on decision-making models?

To answer the above-mentioned questions, we contribute a new problem called Strategic Improvement with Decision Interaction (*SIDI*). In this context, multiple sub-classifiers first evaluate different aspects of the individuals. Finally, a summary decision model, such as a linear model or a neural network block, integrates the results from multiple sub-classifiers to produce a comprehensive evaluation. **Overall, we summarize our contributions as follows:**

1. **Introduction of the SIDI problem**: We propose the problem of strategic improvement with decision interaction (*SIDI*). This novel problem closely mirrors real-world scenarios and extends the strategic machine learning task to a broader dimension. We conduct a systematic analysis of decision interactions and interactive strategic behaviors, examining the correlation among sub-classifiers and the potential impact on system performance.
2. **Improvement framework with decorrelation**: By investigating the relationship between the correlations among interactive classifiers and strategic behaviors, we demonstrate that appropriate decorrelation facilitates genuine individual improvement. Building on this insight, we propose an effective decorrelation method and design a novel strategic improvement framework for decision interactions.

3. **Experimental verification**: Our extensive experiments on real-world and synthetic datasets validate the efficacy and robustness of our method in decision interaction environments.

## 2 Related Work

### 2.1 Strategic Classification

Foundational work in strategic classification [23] commenced by examining how individuals might manipulate their features to obtain favorable outcomes from classifiers. Focusing on the adverse consequences of strategic behavior, some studies aim to develop algorithms that are resilient to such manipulation [9, 44]. Given the complexity of strategic classification, some research attempts to address unknown manipulations or limited information [12, 19]. Recent approaches have proposed stochastic classifiers [43], differentiable optimization-based defenses [34], and graph-based models [14] to improve robustness and handle inter-agent dependencies. Multi-agent extensions further explore strategic externalities and interaction effects [27]. Beyond robustness, to avoid disproportionate disadvantages for certain demographic groups, ongoing studies have also investigated fairness in strategic machine learning [50, 17, 29].

### 2.2 Strategic Improvement

While early research framed strategic behavior as adversarial, an emerging direction views it as an opportunity to promote genuine self-improvement. Strategic improvement [36] introduces a causal framework that aligns agents' incentives with authentic qualification gains. Building on this idea, several studies design mechanisms that encourage changes in causal features or improvable features, rather than superficial manipulation [25, 8, 26, 45, 13, 7]. To address feedback long-term benefits in dynamic systems, performative prediction [39, 22, 38] examines how model deployment can influence the distribution of agents over time. Others explore strategies such as maximizing long-term social welfare [21, 16, 46], aiming to regulate strategic behavior for broader societal benefit [4].

## 3 Preliminaries

We present the essential background on strategic machine learning and causal decorrelation methods. Throughout this paper, we denote random variables by uppercase letters (e.g., $X$ and $Y$) and their realizations by lowercase letters (e.g., $x$ and $y$). Bold symbols (e.g., $\mathbf{x}$ and $\mathbf{X}$) are used for vectors or matrices.

[4] More related work is included in Appendix G.

### 3.1 Strategic Classification

The strategic classification (SC) problem is commonly formulated as a Stackelberg game [5] involving two players: a **decision maker** (modeled as a jury) and **the classified individuals** (modeled as agents) [23].

The decision maker defines a classification function $f : \mathbb{R}^d \to \{0, 1\}$, mapping a feature vector $\mathbf{x}$ to a binary outcome.

**Definition 1 (Strategic Manipulation).** *Agents may strategically change their features to $\mathbf{x}'$ at a cost $c(\mathbf{x}, \mathbf{x}')$. Strategic manipulation is given by*

$$\mathbf{x}' = b(\mathbf{x}) = \arg\max_{\mathbf{x}' \in \mathcal{D}} \left[ f(\mathbf{x}') - \lambda c(\mathbf{x}, \mathbf{x}') \right], \tag{1}$$

*where $f(\mathbf{x}') \in \{0, 1\}$ is the classification result after modification, $c(\mathbf{x}, \mathbf{x}')$ is the manipulation cost, $\lambda > 0$ is a trade-off parameter, and $\mathcal{D}$ is the feature space. Typically, the cost is modeled as the Mahalanobis distance [18].*

**Definition 2 (Decision Optimization).** *To mitigate manipulation, the decision maker optimizes $f$ to maximize expected accuracy against strategic manipulation:*

$$f^* \in \arg\max_{f \in \mathcal{F}} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}} \big[ \mathbb{1}\big( f(b(\mathbf{x})) = y \big) \big], \tag{2}$$

*where $\mathcal{F}$ is the set of all feasible classification rules, $\mathbb{1}$ denotes the indicator function $\mathbb{1}(\cdot) \to \{-1, 1\}$, and $y$ is the observed label.*

*Improvement Against Gaming.* Traditional strategic classifiers often suffer from suboptimal accuracy because they cannot effectively distinguish between genuine improvement and gaming behavior [36, 50]. For example, consider a loan approval scenario: a model predicts whether a customer will repay a loan. In this case, a model designer benefits when $y = +1$, indicating that a borrower will repay the loan.

Within the strategic improvement framework [8, 26], the strategic manipulation of agents can modify their true qualifications. Therefore, if an agent adapts the feature from $x$ to $x'$, with qualification becoming $y(x')$, it may differ from $y(x)$. As a result, strategic improvement corresponds to training a classifier $f^*$ that maximizes the following ideal objective:

$$f^* \in \arg\max_{f \in \mathcal{F}} \mathbb{E}_{x \sim \mathcal{D}} \left[ \mathbb{1}(f(x') = y(x')) \right], \tag{3}$$

where $x'$ is the agent's adapted feature, and $y(x')$ is the true qualification after strategic behavior.

[5] In Stackelberg framework, the interaction unfolds in two stages: (i) the decision maker publishes its policy (classification rule $f$); (ii) decision subjects (agents), after observing the policy, decide whether to modify their features.

### 3.2 Decorrelation Method

Decorrelation is a fundamental concept in machine learning, aimed at reducing or eliminating the correlation between variables to maximize their independence. This technique plays a crucial role in addressing multicollinearity issues among variables and minimizing dependencies between features in specific tasks.

Our work aims to propose a decorrelation method applicable to decision interactions based on the following lemmas [6].

**Lemma 1 (Variable Independence [6]).** *Variables $X$ and $Y$ are independent if $\mathbb{E}(X^k \cdot Y^l) = \mathbb{E}(X^k) \cdot \mathbb{E}(Y^l)$ holds for all $k$ and $l$ in $\mathbf{N}$ with discretization condition.*

Inspired by the sample weighting methods in the causal literature [3, 31], we investigate the following methods for linear correlation.

**Lemma 2 (Independence via Reweighting [32]).** *Let $X_a$ and $X_b$ be random variables with all existing moments. There exists a reweighting scheme with weights $W$ such that $X_a$ and $X_b$ become independent under the weighted distribution. The optimal weights $W^*$ can be obtained by solving the following moment-matching objective:*

$$W^* = \arg\min_{W} \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} \left\| \mathbb{E}\left[X_a^k \cdot \Sigma_W \cdot X_a^k\right] - \mathbb{E}\left[X_a^k \cdot W\right] \mathbb{E}\left[X_b^k \cdot W\right] \right\|_2^2, \quad (4)$$

*where $\Sigma_W$ denotes the covariance structure under the weighted distribution.*

**Lemma 3 (Decorrelation via Fourier Transform [49]).** *By applying the Fourier transform to map the variables with linear and nonlinear correlations in the original domain into the frequency domain, the resulting features can be decorrelated using standard linear decorrelation methods.*

## 4 Problem Statement

### 4.1 Decision Interaction

In many real-world applications, decision makers rely on multiple sub-classifiers to evaluate agents based on various criteria. These sub-classifiers are often not independent: **a favorable outcome from one classifier can influence the results of others, ultimately affecting the final decision.** We refer to this phenomenon as **decision interaction**.

**Definition 3 (Decision Interaction).** *Let $\mathcal{X}$ denote the population with distribution $\mathcal{D}$. Suppose there are $n$ interacting classifiers $\{h_1, h_2, \ldots, h_n\}$, i.e., sub-classifiers, where each classifier produces an outcome according to*

$$y_i = h_i(\mathbf{x}, y_{-i}), \quad (5)$$

[6] The proofs of these lemmas are included in Appendix A.

*where $y_{-i}$ denotes the outputs of all classifiers except $h_i$, for $i = 1, \ldots, n$. These outputs may either be treated as fixed from a previous evaluation or dynamically updated, depending on the system.*

Let $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ be the vector of outputs. The final comprehensive decision is computed as $s = g(\mathbf{Y})$, where $g : \mathbb{R}^n \to \{-1, +1\}$ is an aggregation function that maps the sub-classifier outputs to a final decision. The function $g$ can be implemented as a linear model, a neural network, or a rule-based evaluator.

*Remark 1.* We say that *decision interaction* exists if a change in $y_i$ (for some $i$) can directly or indirectly influence the output of another classifier $y_j$ (for $j \neq i$) through their functional dependencies, thereby affecting the final decision $s$.

Under decision interaction, the decision boundaries of classifiers are no longer isolated. Instead, they form an interdependent structure where a modification in one feature can propagate through multiple classifiers, amplifying its effect on the final decision outcome.

### 4.2 Interactive Strategic Behavior

In a decision interaction environment, agents can exploit dependencies among multiple classifiers to achieve favorable outcomes at a reduced cost. This phenomenon, which we term *interactive strategic behavior*, arises when agents coordinate their feature modifications to influence one classifier's output in a way that also affects others, thereby amplifying the overall effect of their strategic actions.

Let $x \in \mathcal{X}$ be the original feature vector of a agent, and let $c(x, x')$ denote the cost of modifying $x$ to $x'$. Suppose there are $n$ interacting classifiers $\{h_1, h_2, \ldots, h_n\}$, where each classifier's output may depend not only on $\mathbf{x}$ but also on the outputs of the others.

To isolate the effect of strategic behavior on a particular classifier, we adopt a *freeze-and-optimize* framework. First, we compute the outputs $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)$ on the original input $x$ using a fixed-point evaluation procedure. Then, for a given index $i$, we freeze all outputs $y_j^*$ for $j \neq i$, and allow only the $i$-th classifier to respond to modifications in the feature vector.

**Definition 4 (Interactive Strategic Behavior).** *Let $\mathbf{y}_{-i}^* = (y_1^*, \ldots, y_{i-1}^*, y_{i+1}^*, \ldots, y_n^*)$ denote the frozen outputs of all classifiers except the $i$-th. The agent exhibits interactive strategic behavior with respect to the $i$-th classifier by selecting a modified feature vector $x'$ according to:*

$$x' = \arg\max_{x' \in \mathcal{D}} \left\{ g\big(h_1(x', \mathbf{y}_{-1}^*), \ldots, h_n(x', \mathbf{y}_{-n}^*)\big) - \lambda\, c(x, x') \right\}, \tag{6}$$

*where $\lambda > 0$ is a trade-off parameter. Although the overall decision function $g$ depends on all classifiers, only the $i$-th classifier's output changes in response to $x'$; all others remain fixed by construction.*

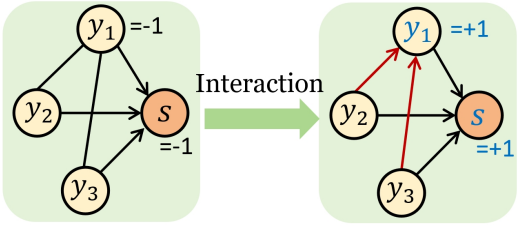



Fig. 2: A loan approval example with interacting decision models. The red arrows represent decision interactions.

*Remark 2.* This definition formalizes a targeted form of strategic behavior in interactive settings. The agents optimize their features to influence a particular sub-classifier, while accounting for the fixed influence of others. This setup reveals how local manipulations can leverage inter-classifier dependencies to affect the final decision, even without directly altering all sub-models.

*Example 1 (Loan Approval with Interacting Classifiers).* As illustrated in Fig. 2, consider a loan application evaluated by three interacting classifiers: *creditworthiness* ($y_1$), *fraud detection* ($y_2$), and *indebtedness* ($y_3$). Initially, the applicant is rejected ($s = -1$), primarily due to a low creditworthiness score ($y_1 = -1$). Under traditional strategic manipulation, the applicant may attempt to improve $y_1$ directly (e.g., by repaying part of a loan), but the effect is limited and insufficient to change the final decision. By recognizing the dependencies among classifiers, the applicant instead improves $y_2$ and $y_3$, which indirectly influence $y_1$ through decision interactions. As a result, $y_1$ increases beyond the approval threshold, ultimately yielding a favorable final decision ($s = +1$). This illustrates how *interactive strategic behavior* can leverage inter-model dependencies to achieve positive outcomes at a lower manipulation cost.

In the context of decision interaction, as illustrated in Example 1, agents consider the outcomes of multiple decision models when deciding how to manipulate their features. By leveraging these interdependent relationships, they can achieve desired results at a reduced cost compared to manipulating each classifier in isolation.

We define our overarching problem in this multi-classifier environment.

*Problem 1 (Strategic Improvement with Decision Interaction).* Consider a decision interaction environment $e$ with a population $\mathcal{X}^e$ drawn from a distribution $\mathcal{D}^e$, and $n$ interacting classifiers $h_1^e, \ldots, h_n^e$ producing outcomes $\mathbf{Y}^e$. The comprehensive evaluation is given by $s^e = g(\mathbf{Y}^e)$. Our task is to learn an algorithm $\mathcal{A}$ that minimizes interactive strategic behavior and encourages authentic improvements.

To solve this novel problem, we provide an incentive improvement framework for decision interactions and design a novel decorrelation method for interactive strategic behavior.

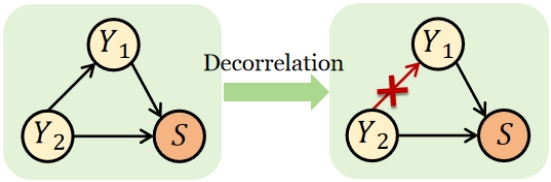



Fig. 3: Schematic of the decision interaction. $Y_1$ and $Y_2$ denote the results of the sub-classifiers, while $S$ represents the aggregate score. The red arrows signify interactions between the classifiers.

## 5 Methods for Strategic Improvement with Decision Interaction

We argue that interactive strategic behavior arises due to the correlation among interacting classifiers. In this section, we analyze this correlation and subsequently establish a novel framework for strategic improvement under decision interaction.

### 5.1 Decorrelation Promoting Improvement

We begin by noting that in SIDI, interacting classifiers exhibit a strong correlation. As depicted in Fig. 3, a change in one classifier (e.g., $Y_2$) can influence the output of another classifier (e.g., $Y_1$), ultimately affecting the final evaluation score $S$. To analyze the relationship between correlation and strategic behavior, we propose the following hypothesis:

**Hypothesis 1 (Correlation Exacerbates Strategic Behavior)** *In SIDI, correlation among interacting classifiers exacerbates strategic (manipulative) behavior of agents, enabling them to achieve favorable outcomes by modifying their features without true improvement.*

*Remark 3.* This hypothesis is motivated by the observation that when interactions exist among sub-classifiers, even small modifications to individual features are amplified through decision interactions, leading to more pronounced changes in the final decision scores.

We quantify this phenomenon by introducing the **local sensitivity**.

**Definition 5 (Local Sensitivity in Decision Interaction).** *Let* $F(\mathbf{x}) = g\big(h_1(\mathbf{x}, y_{-1}), \ldots, h_n(\mathbf{x}, y_{-n})\big)$ *be the comprehensive decision function with* $y_i = h_i(\mathbf{x}, y_{-i})$. *The* local sensitivity *at point* $\mathbf{x}$ *is defined as*

$$\gamma(\mathbf{x}) = \left\| Dg\big(\mathbf{h}(\mathbf{x})\big) \cdot \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}) \right\|, \tag{7}$$

*where* $Dg\big(\mathbf{h}(\mathbf{x})\big) \in \mathbb{R}^{1\times n}$ *is the gradient of* $g$ *at* $\mathbf{h}(\mathbf{x})$, *and* $\nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x}) \in \mathbb{R}^{n\times d}$ *is the Jacobian matrix of* $\mathbf{h}(\mathbf{x})$ *with respect to* $\mathbf{x}$ [7].

[7] The specific derivation is in Appendix D.1.

*Remark 4.* Local sensitivity $\gamma(\mathbf{x})$ characterizes the impact on the comprehensive evaluation $F(\mathbf{x})$ following a change in feature $\mathbf{x}$. A larger value of $\gamma(\mathbf{x})$ indicates a greater influence.

Consequently, consider a decorrelation transformation $Q \in \mathbb{R}^{n\times n}$ used to de-correlate the interacting classifier outputs, i.e.,$\tilde{\mathbf{h}}(\mathbf{x}) = Q\,\mathbf{h}(\mathbf{x})$ and $\tilde{g}\big(\tilde{\mathbf{h}}(\mathbf{x})\big) = g\big(Q^{-1}\tilde{\mathbf{h}}(\mathbf{x})\big)$. The decorrelation local sensitivity is denoted as:

$$\tilde{\gamma}(\mathbf{x}) = \left\| D\tilde{g}\big(Q^{-1}\tilde{\mathbf{h}}(\mathbf{x})\big) \cdot Q^{-1} \cdot \nabla_{\mathbf{x}}\tilde{\mathbf{h}}(\mathbf{x}) \right\|. \tag{8}$$

We argue that $\tilde{\gamma}(\mathbf{x}) < \gamma(\mathbf{x})$ (see Appendix D.2 for details), which indicates that decorrelation reduces the utility of strategic behavior.

**Proposition 1 (Decorrelation Promotes Improvement [8]).** *In the SIDI problem, an appropriate decorrelation method that attenuates only the spurious correlations will diminish the gains from manipulative strategic behavior, thereby promoting substantive improvement.*

### 5.2 Decorrelation between Interacting Classifiers

Note that in the context of the comprehensive evaluation process, the outcomes of interacting classifiers are often discrete. Accordingly, we present a decorrelation theorem for SIDI based on Lemma 1.

**Theorem 1 (Independence in SIDI).** *Let $\mathbf{Y}_i$ and $\mathbf{Y}_j$ be the outcomes of two interacting classifiers, each taking a finite number of discrete values. For each positive integer $k$, denote by $\mathbf{Y}_i^{(k)}$ the $k$-th order moment of $\mathbf{Y}_i$, and similarly for $\mathbf{Y}_j^{(k)}$. Suppose there exists a sequence $\{\epsilon(k)\}_{k\in\mathbb{N}}$ with $\epsilon(k) > 0$ for all $k \in \mathbb{N}$ and $\lim \epsilon(k) = 0$, such that for every $k \in \mathbb{N}$ the following inequality holds:*

$$\left\| \mathbb{E}\Big[\big(\mathbf{Y}_i^{(k)}\big)^T \mathbf{Y}_j^{(k)}\Big] - \mathbb{E}\Big[\mathbf{Y}_i^{(k)}\Big] \cdot \mathbb{E}\Big[\mathbf{Y}_j^{(k)}\Big] \right\|_2^2 < \epsilon(k). \tag{9}$$

*Then, the outcomes $\mathbf{Y}_i$ and $\mathbf{Y}_j$ can be regarded as approximately independent [9].*

*Remark 5.* In SIDI, both linear and nonlinear correlations coexist. To address linear and nonlinear correlations more effectively, we transform the decorrelation problem into a linear one in a high-dimensional space, leveraging the principles outlined in Lemma 3.

**Definition 6 (Fourier Transform in Interacting Classifiers).** *Let $u_k(Y_i)$ denote the Fourier features obtained from a function $u_k \in \mathcal{H}_{RFF}$. Then, the Fourier features of $\mathbf{Y}_i$ are given by:*

$$\mathbf{u}(\mathbf{Y}_i) = \big(u_1(\mathbf{Y}_i), u_2(\mathbf{Y}_i), \cdots, u_m(\mathbf{Y}_i)\big) \in \mathcal{H}_{RFF}. \tag{10}$$

[8] The proof is included in Appendix E.
[9] The proof can be found in Appendix B.

*The functions* $\{u_1, u_2, \cdots, u_m\}$ *are sampled from the space of Random Fourier Features* $\mathcal{H}_{RFF}$*:*

$$\mathcal{H}_{RFF} = \{\, f : x \to \sqrt{2}\cos(\omega x + \phi) \mid \ \omega \sim N(0,1),\ \phi \sim \mathit{Uniform}(0, 2\pi)\}. \quad (11)$$

*where* $\omega$ *is sampled from the standard Normal distribution and* $\phi$ *is sampled from the Uniform distribution.*

According to Lemma 2, we introduce adaptive weights $\mathbf{W}$ to regulate the correlations between interacting classifiers. Given the Fourier features $\mathbf{u}(\mathbf{Y}_i)$ and $\mathbf{u}(\mathbf{Y}_j)$ for interacting classifiers, we propose the following measure:

$$\begin{aligned}\Delta_{\mathbf{W}}^{\mathcal{H}_{\mathrm{RFF}}}(k) = \sum_{i=1}^{n}\sum_{j\neq i}\Big\|\mathbb{E}\Big[(\mathbf{u}(\mathbf{Y}_i^{(k)}))^T \cdot \Sigma_{\mathbf{W}} \cdot \mathbf{u}(\mathbf{Y}_j^{(k)})\Big] \\ -\mathbb{E}\Big[(\mathbf{u}(\mathbf{Y}_i^{(k)}))^T \cdot \mathbf{W}\Big]\,\mathbb{E}[(\mathbf{u}(\mathbf{Y}_j^{(k)}))^T \cdot \mathbf{W}\Big]\Big\|_2^2,\end{aligned} \quad (12)$$

where $\Sigma_{\mathbf{W}}$ denotes the covariance structure under the weighted distribution.

*Remark 6.* According to the basic condition for linear independence [24, 48], it is sufficient to consider only the cases $k = 1$ and $k = 2$ in Theorem 1.

**Proposition 2 (Adaptive Weight in Decision Interaction [10]).** *In the SIDI context, given the outcomes* $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ *of* $n$ *interacting classifiers, the correlations among them can be reduced by designing adaptive weights* $\mathbf{W}$*, which can be learned via the following objective:*

$$\mathbf{W}^* = \arg\max_{\mathbf{W}\in\Delta_n}\sum_{k=1}^{2}\Delta_{\mathbf{W}}^{\mathcal{H}_{RFF}}(k), \quad \textit{where} \quad \Delta_n = \left\{\mathbf{w}\in\mathbb{R}_+^n \mid \sum_{i=1}^{n} w_i = n\right\}. \quad (13)$$

### 5.3 Decision Interaction Improvement Mechanism

Traditional strategic classification methods and improvement mechanisms have not yet accounted for decision interactions. Based on the outline of the strategic classification in Subsection 3.1, we propose a new strategic improvement mechanism that explicitly addresses the SIDI problem.

**Definition 7 (Utility Function in SIDI).** *Given* $n$ *interacting classifiers* $h_1, h_2, \ldots, h_n$*, and a comprehensive evaluator* $g$*, the utility function is defined as:*

$$U(\mathbf{x}, \mathbf{x}') = g(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n) - \lambda c(\mathbf{x}, \mathbf{x}'), \quad (14)$$

*where* $\mathbf{x}'$ *is the modified feature vector from strategic behavior,* $\mathbf{y}_i$ *is the outcome of the* $i$*-th interacting classifier* $h_i$*, and* $c$ *is the cost function.*

[10] The proof can be found in Appendix F.

**Definition 8 (Strategic Behavior in SIDI).** *In strategic improvement, a agent's strategic behavior is categorized into two types $b_I$ and $b_M$:*

$$\mathbf{x}'_I = b_I(\mathbf{x}) = \arg\max_{x,x' \in \mathcal{D}} U(\mathbf{x}, \mathbf{x}'), \qquad \mathbf{x}'_M = b_M(\mathbf{x}) = \arg\max_{x,x' \in \mathcal{D}} U(\mathbf{x}, \mathbf{x}'), \tag{15}$$

*where $b_I$ represents strategic behaviors that genuinely enhance the agent's true qualification, while $b_M$ represents strategic behaviors aimed solely at deceiving the decision model.*

*Remark 7.* We write $\mathbf{x} = (\mathbf{x}_I, \mathbf{x}_M, \mathbf{x}_U)$, which denotes the categories of features:*Improvable features ($\mathbf{x}_I$)*, *Manipulable features ($\mathbf{x}_M$)*, and *Immutable features ($\mathbf{x}_U$)* [11].

**Definition 9 (Improvement for Interacting Classifier).** *A new objective function can be introduced for each interacting classifier in SIDI.*

$$h_i^* \in \arg\max_{h_i \in \mathcal{H}} R_{DI}(h_i) := \kappa R_I(h_i) + R_M(h_i), \tag{16}$$

*where $R_I(h_i) = \mathbb{E}_{x \sim \mathcal{D}}[\mathbb{1}(h_i(\mathbf{x}'_I, \mathbf{x}_U, \mathbf{y}_{-i}) = +1)]$ is the improvement objective, rewarding agents for achieving genuine improvement, while $R_M(h_i) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\mathbb{1}(h_i(\mathbf{x}'_I, \mathbf{x}'_M, \mathbf{x}_U, \mathbf{y}_{-i}) = y_i)]$ is the manipulation objective. The parameter $\kappa > 0$ balances the two competing objectives.*

**Definition 10 (Decision Optimization in SIDI).** *Let $\mathbf{y}_i^*$ denote the outcome of the optimized interacting classifier $h_i^*$. We incorporate adaptive weights into $g$ to obtain a weighted evaluator $g_W$. We optimize $g_W$ via the following objective:*

$$g_W^* \in \arg\max_{g_W \in \mathcal{G}} \mathbb{E}_{(\mathbf{x},\mathbf{y}) \sim \mathcal{D}} \left[ \mathbb{1} \left( g_W(\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*) = \mathbf{s}_{true} \right) \right]. \tag{17}$$

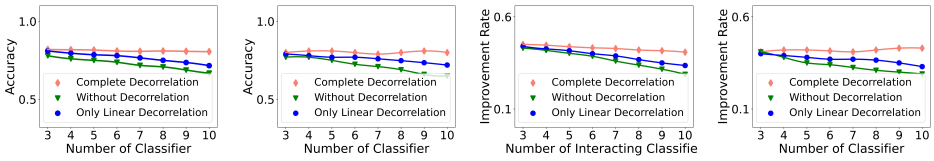
*Remark 8.* By employing a comprehensive evaluator with adaptive weights and integrating an improvement objective through interacting classifiers, we establish a novel strategic improvement framework in SIDI.

## 6 Experiment

In this section, we evaluate the efficacy of our method for the new problem, i.e., strategic improvement with decision interaction. In this new environment, our main experiments are divided into four parts:

1. Compare the performance of the single model and the interactive models across different classification scenarios.
2. Verify the relationship between decorrelation and strategic improvement
3. Evaluate the performance with three different methods. ***TC***, traditional strategic classification ignore the improvement in agents. ***TI***, traditional strategic improvement, does not consider decision interaction. ***SIDI***, our method, focuses on strategic improvement with decision interaction,
4. Ablation studies on the number of interacting classifiers.

[11] The method for feature classification is included in Appendix C.




(a) Small-scale (b) Large-scale (c) Small-scale (d) Large-scale

Fig. 4: Performance comparison in the SIDI problem under different degrees of decorrelation across small-scale and large-scale datasets.

### 6.1 Setup

**Dataset.** We utilized six real datasets and one synthetic dataset to validate the efficacy of our method. ***Credit***, a dataset contains individual credit information, including credit history, loan purpose, loan amount, employment duration, and personal information [47]. ***Student Performance***, a dataset includes student performance data in mathematics and Portuguese language courses [11]. ***Adult***, a census-based dataset for predicting adult annual incomes [5]. ***PhiUSIIL***, a substantial dataset comprising 134,850 legitimate and 100,945 phishing URLs [40]. ***German***, a dataset to assess credit risk in loans from the UCI ML Repository [30]. ***Synthetic*** [35], a synthetic dataset generated using the PaySim simulator, which mimics mobile financial transactions and fraud patterns based on real-world data.

**Metric.** In addition to ***accuracy***, we introduce the improvement and cheatment rates to assess the effects of strategic behavior. ***Improvement rate*** is defined as the proportion of the population initially facing rejection but subsequently being accepted following incentive adaptation. ***Cheatment rate*** is defined as the proportion of the population initially facing rejection, but subsequently being accepted by exploiting strategic behavior without genuine improvement.

To verify the effect of correlation, when linear and nonlinear decorrelation together, the adaptive weight $W$ is trained with Eq (13). If only linear correlation, the adaptive weight $W$ is trained with the following objective:

$$\mathbf{W}' = \arg\min_{\mathbf{W}} \big(\Delta_{\mathbf{W}}(1) + \Delta_{\mathbf{W}}(2)\big), \tag{18}$$

where we set:

$$\Delta_{\mathbf{W}}(k) = \sum_{i=1}^{n} \sum_{j \neq i} \Big\| \mathbb{E}[(\mathbf{Y}_i^{(k)})^T \cdot \Sigma_{\mathbf{W}} \cdot \mathbf{Y}_j^{(k)}] - \mathbb{E}[(\mathbf{Y}_i^{(k)})^T \cdot \mathbf{W}] \cdot \mathbb{E}[(\mathbf{Y}_j^{(k)})^T \cdot \mathbf{W}] \Big\|_2^2. \tag{19}$$

### 6.2 Implementation Details

All experiments are conducted on a single NVIDIA TITAN V 12GB GPU. The adaptive weights used in our decorrelation method are optimized via the Adam

Table 1: Performance comparison between single model and interacting models

| Metrics | Methods | Datasets | | | | | |
|---|---|---|---|---|---|---|---|
| | | *Credit* | *Student* | *Adult* | *German* | *PhiUSIIL* | *Synthetic* |
| **Non-strategic** | | | | | | | |
| *Accuracy* | *Single Model* | $71.83_{\pm 1.65}$ | $77.42_{\pm 1.30}$ | $82.12_{\pm 1.05}$ | $91.28_{\pm 1.35}$ | $64.82_{\pm 1.10}$ | $81.35_{\pm 1.20}$ |
| | *Interacting models* | $74.23_{\pm 1.81}$ | $79.85_{\pm 1.50}$ | $83.45_{\pm 1.20}$ | $91.17_{\pm 1.60}$ | $65.15_{\pm 1.25}$ | $85.73_{\pm 1.35}$ |
| **Strategic** | | | | | | | |
| *Accuracy* | *Single model* | $69.52_{\pm 2.70}$ | $73.15_{\pm 2.13}$ | $79.23_{\pm 1.30}$ | $88.13_{\pm 1.80}$ | $63.53_{\pm 1.60}$ | $73.75_{\pm 1.90}$ |
| | *Interacting models* | $71.22_{\pm 1.97}$ | $76.78_{\pm 1.85}$ | $81.56_{\pm 1.13}$ | $90.61_{\pm 1.54}$ | $64.29_{\pm 1.02}$ | $80.66_{\pm 1.58}$ |
| *Imp. rate* | *Single model* | $38.53_{\pm 5.30}$ | $34.93_{\pm 3.80}$ | $35.52_{\pm 3.30}$ | $34.85_{\pm 4.50}$ | $30.83_{\pm 4.10}$ | $27.93_{\pm 4.20}$ |
| | *Interacting models* | $47.13_{\pm 4.78}$ | $39.74_{\pm 3.60}$ | $42.29_{\pm 4.12}$ | $43.05_{\pm 3.67}$ | $36.41_{\pm 2.91}$ | $38.54_{\pm 3.30}$ |

*Note:* "Non-strategic" indicates that the models do not take into account agents' strategic behaviors during both training and inference, whereas "Strategic" means that the model integrates these strategic behaviors. "Single Model" comes from the collection of existing methods [19, 10, 8, 28]. "Interacting models" are designed with decision interaction. "Imp. rate" is short for the improvement rate.

Table 2: Performance comparison of different methods in SIDI.

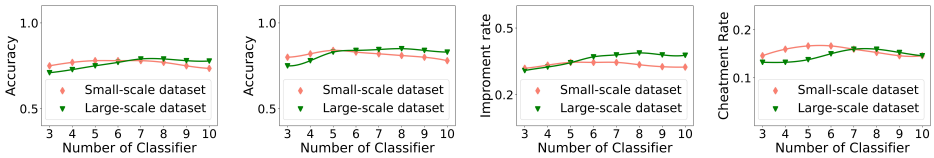
| Methods | Metrics | Datasets | | | | | |
|---|---|---|---|---|---|---|---|
| | | *Credit* | *Student* | *Adult* | *German* | *PhiUSIIL* | *Synthetic* |
| *TC [19, 37, 42, 10]* | Accuracy | $70.89_{\pm 2.77}$ | $76.50_{\pm 2.06}$ | $81.61_{\pm 1.22}$ | $89.36_{\pm 1.87}$ | $63.81_{\pm 1.71}$ | $74.08_{\pm 1.82}$ |
| | Imp. rate | $32.14_{\pm 4.54}$ | $28.67_{\pm 3.48}$ | $30.72_{\pm 3.29}$ | $29.35_{\pm 3.91}$ | $24.56_{\pm 2.56}$ | $25.66_{\pm 3.68}$ |
| | Cheatment | $30.88_{\pm 2.76}$ | $23.71_{\pm 1.89}$ | $25.30_{\pm 2.12}$ | $24.35_{\pm 1.83}$ | $22.72_{\pm 1.54}$ | $28.53_{\pm 2.68}$ |
| *TI [21, 36, 8, 28]* | Accuracy | $68.43_{\pm 1.89}$ | $74.50_{\pm 2.27}$ | $79.10_{\pm 1.63}$ | $87.54_{\pm 2.14}$ | $61.50_{\pm 1.42}$ | $72.62_{\pm 1.93}$ |
| | Imp. rate | $40.71_{\pm 5.54}$ | $37.65_{\pm 3.00}$ | $39.36_{\pm 3.36}$ | $38.35_{\pm 4.75}$ | $32.53_{\pm 4.03}$ | $33.27_{\pm 4.09}$ |
| | Cheatment | $23.68_{\pm 2.57}$ | $20.24_{\pm 1.96}$ | $22.10_{\pm 2.35}$ | $20.53_{\pm 2.19}$ | $18.33_{\pm 1.62}$ | $20.33_{\pm 2.52}$ |
| ***SIDI(ours)*** | Accuracy | $\mathbf{71.66}_{\pm 2.27}$ | $\mathbf{77.21}_{\pm 2.15}$ | $\mathbf{82.00}_{\pm 1.43}$ | $\mathbf{90.57}_{\pm 2.04}$ | $\mathbf{64.73}_{\pm 1.32}$ | $\mathbf{81.10}_{\pm 1.88}$ |
| | Imp. rate | $\mathbf{48.05}_{\pm 5.93}$ | $\mathbf{41.59}_{\pm 4.75}$ | $\mathbf{43.21}_{\pm 5.27}$ | $\mathbf{42.97}_{\pm 3.82}$ | $\mathbf{36.33}_{\pm 4.06}$ | $\mathbf{38.59}_{\pm 3.45}$ |
| | Cheatment | $\mathbf{19.68}_{\pm 1.81}$ | $\mathbf{16.24}_{\pm 1.67}$ | $\mathbf{18.10}_{\pm 1.51}$ | $\mathbf{16.50}_{\pm 1.72}$ | $\mathbf{15.24}_{\pm 1.51}$ | $\mathbf{16.33}_{\pm 2.05}$ |

optimizer (learning rate 0.0005), with softmax normalization applied per mini-batch. The weighting scheme converges stably within 128 steps in our experiments. For the decorrelation computation, we use 32-dimensional random Fourier features. Each sub-classifier $h_i$ is instantiated as a logistic regression model. The final aggregator $g$ is implemented as a linear model [12]. The full model is trained using Adam (learning rate 0.0001) with early stopping.

### 6.3 Results and Analysis

Table 1 shows that, across various datasets, models using decision interactions outperform single-model decisions in both *non-strategic* and *strategic* environments. Under strategic conditions, interacting models achieve a significantly

[12] We also consider different interaction types. The details and results are included in Appendix H.




(a) Non-strategic (b) Strategic (c) Improvement (d) Cheatment

Fig. 5: Effect of the number of interactive classifiers on performance in the SIDI problem.

higher improvement rate, suggesting that decision interactions more effectively motivate agents to enhance their qualifications.

Table 2 validates our framework on multiple datasets in a decision interaction context. Our method outperforms existing strategic classification and improvement approaches in accuracy, improvement rate, and cheatment rate, the latter reduction attributed to adaptive weight decorrelation mitigating strategic behavior, while the higher improvement rate indicates more genuine qualification enhancements.

Fig. 4 compares results on datasets of varying scales using adaptive weights in both linear and nonlinear decorrelation settings. Both decorrelation types significantly affect accuracy and improvement rate, reducing the impact of strategic behavior in the SIDI problem and encouraging qualification improvements.

Fig. 5 examines the effect of the number of interactive classifiers in both strategic and non-strategic settings. The accuracy and improvement rate initially rise with more classifiers, though an excessive number causes a slight decline in accuracy, while the cheatment rate remains stable. These findings imply that an optimal number of classifiers enhances overall performance.

## 7 Conclusion

In this work, we introduce a novel problem in strategic machine learning: **S**trategic **I**mprovement with **D**ecision **I**nteraction (*SIDI*). We frame the problem and uncover novel strategic behaviors among agents in decision interaction. By analyzing the correlations among interacting classifiers, we introduce a comprehensive decorrelation method and propose a new strategic improvement framework. In future work, we aim to incorporate fairness considerations into our framework.

## Acknowledgement

This work is supported in part by the National Natural Science Foundation of China under Grants No. 62372459 and 62302503, the Natural Science Foundation of Heilongjiang Province under Grant No. LH2023C069, the NUDT Youth Independent Innovation Science Fund under Grant No. ZK23-15, and the Open Research Fund of the State Key Laboratory of High Performance Computing of China under Grant No. 202401-09.

## References


1. Abdoli, M., Akbari, M., Shahrabi, J.: Bagging supervised autoencoder classifier for credit scoring. Expert Systems with Applications **213**, 118991 (2023)
2. Alagic, A., Zivic, N., Kadusic, E., Hamzic, D., Hadzajlic, N., Dizdarevic, M., Selmanovic, E.: Machine learning for an enhanced credit risk analysis: A comparative study of loan approval prediction models integrating mental health data. Machine Learning and Knowledge Extraction **6**(1), 53–77 (2024). https://doi.org/10.3390/make6010004, https://www.mdpi.com/2504-4990/6/1/4
3. Athey, S., Imbens, G.W., Wager, S.: Approximate residual balancing: debiased inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society Series B: Statistical Methodology **80**(4), 597–623 (2018)
4. Bechavod, Y., Ligett, K., Wu, S., Ziani, J.: Gaming helps! learning from strategic interactions in natural dynamics. In: International Conference on Artificial Intelligence and Statistics. pp. 1234–1242. PMLR (2021)
5. Becker, B., Kohavi, R.: Adult. UCI Machine Learning Repository (1996), DOI: https://doi.org/10.24432/C5XW20
6. Bisgaard, T.M., Sasvári, Z.: When does e (xk· yl)= e (xk)· e (yl) imply independence? Statistics probability letters **76**(11), 1111–1116 (2006)
7. Chang, T., Warrenburg, L., Park, S.H., Parikh, R., Makar, M., Wiens, J.: Who's gaming the system? a causally-motivated approach for detecting strategic adaptation. Advances in Neural Information Processing Systems **37**, 42311–42348 (2024)
8. Chen, Y., Wang, J., Liu, Y.: Learning to incentivize improvements from strategic agents. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=W98AEKQ38Y
9. Chen, Y., Liu, Y., Podimata, C.: Grinding the space: Learning to classify against strategic agents. arXiv preprint arXiv:1911.04004 (2019)
10. Cohen, L., Mansour, Y., Moran, S., Shao, H.: Learnability gaps of strategic classification (2024), https://arxiv.org/abs/2402.19303
11. Cortez, P.: Student Performance. UCI Machine Learning Repository (2014), DOI: https://doi.org/10.24432/C5TG7T
12. Dong, J., Roth, A., Schutzman, Z., Waggoner, B., Wu, Z.S.: Strategic classification from revealed preferences. In: Proceedings of the 2018 ACM Conference on Economics and Computation. pp. 55–70 (2018)
13. Efthymiou, V., Podimata, C., Sen, D., Ziani, J.: Incentivizing desirable effort profiles in strategic classification: The role of causality and uncertainty. arXiv preprint arXiv:2502.06749 (2025)
14. Eilat, I., Finkelshtein, B., Baskin, C., Rosenfeld, N.: Strategic classification with graph neural networks. arXiv preprint arXiv:2205.15765 (2022)
15. Emmanuel, I., Sun, Y., Wang, Z.: A machine learning-based credit risk prediction engine system using a stacked classifier and a filter-based feature selection method. Journal of Big Data **11**(1), 23 (2024)
16. Estornell, A., Chen, Y., Das, S., Liu, Y., Vorobeychik, Y.: Incentivizing recourse through auditing in strategic classification. In: IJCAI (2023)
17. Estornell, A., Das, S., Liu, Y., Vorobeychik, Y.: Group-fair classification with strategic agents. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. p. 389–399. FAccT '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3593013.3594006, https://doi.org/10.1145/3593013.3594006

18. Gavish, M., Talmon, R., Su, P.C., Wu, H.T.: Optimal recovery of precision matrix for mahalanobis distance from high dimensional noisy observations in manifold learning (2021), https://arxiv.org/abs/1904.09204
19. Ghalme, G., Nair, V., Eilat, I., Talgam-Cohen, I., Rosenfeld, N.: Strategic classification in the dark. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 3672–3681. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/ghalme21a.html
20. Grosz, M.: Admissions policies, cohort composition, and academic success: Evidence from california. Journal of Human Resources **58**(4), 1242–1272 (2023)
21. Haghtalab, N., Immorlica, N., Lucier, B., Wang, J.Z.: Maximizing welfare with incentive-aware evaluation mechanisms. arXiv preprint arXiv:2011.01956 (2020)
22. Hardt, M., Jagadeesan, M., Mendler-Dünner, C.: Performative power. Advances in Neural Information Processing Systems **35**, 22969–22981 (2022)
23. Hardt, M., Megiddo, N., Papadimitriou, C., Wootters, M.: Strategic classification. In: Proceedings of the 2016 ACM conference on innovations in theoretical computer science. pp. 111–122 (2016)
24. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classification. In: European Conference on Computer Vision. pp. 459–472. Springer (2012)
25. Harris, K., Ngo, D.D.T., Stapleton, L., Heidari, H., Wu, S.: Strategic instrumental variable regression: Recovering causal relationships from strategic responses. In: International Conference on Machine Learning. pp. 8502–8522. PMLR (2022)
26. Horowitz, G., Rosenfeld, N.: Causal strategic classification: A tale of two shifts. In: International Conference on Machine Learning. pp. 13233–13253. PMLR (2023)
27. Hossain, S., Micha, E., Chen, Y., Procaccia, A.: Strategic classification with externalities. arXiv preprint arXiv:2410.08032 (2024)
28. Jin, K., Zhang, X., Khalili, M.M., Naghizadeh, P., Liu, M.: Incentive mechanisms for strategic classification and regression problems. In: Proceedings of the 23rd ACM Conference on Economics and Computation. pp. 760–790 (2022)
29. Keswani, V., Celis, L.E.: Addressing strategic manipulation disparities in fair classification. In: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. EAAMO '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3617694.3623252, https://doi.org/10.1145/3617694.3623252
30. Khan, M.M.R., Arif, R.B., Siddique, M.A.B., Oishe, M.R.: Study and observation of the variation of accuracies of knn, svm, lmnn, enn algorithms on eleven different datasets from uci machine learning repository. In: 2018 4th International Conference on Electrical Engineering and Information Communication Technology (iCEEiCT). pp. 124–129 (2018). https://doi.org/10.1109/CEEICT.2018.8628041
31. Kuang, K., Cui, P., Li, B., Jiang, M., Yang, S., Wang, F.: Treatment effect estimation with data-driven variable decomposition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
32. Kuang, K., Xiong, R., Cui, P., Athey, S., Li, B.: Stable prediction with model misspecification and agnostic distribution shift. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 4485–4492 (2020)
33. Kurysheva, A., van Rijen, H.V., Dilaver, G.: How do admission committees select? do applicants know how they select? selection criteria and transparency at a dutch university. Tertiary education and management **25**(4), 367–388 (2019)
34. Levanon, S., Rosenfeld, N.: Strategic classification made practical. In: International Conference on Machine Learning. pp. 6243–6253. PMLR (2021)

35. Lopez-Rojas, E., Elmir, A., Axelsson, S.: Paysim: A financial mobile money simulator for fraud detection. In: 28th European modeling and simulation symposium, EMSS, Larnaca. pp. 249–255. Dime University of Genoa (2016)
36. Miller, J., Milli, S., Hardt, M.: Strategic classification is causal modeling in disguise. In: International Conference on Machine Learning. pp. 6917–6926. PMLR (2020)
37. Milli, S., Miller, J., Dragan, A.D., Hardt, M.: The social cost of strategic classification. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 230–239 (2019)
38. Mofakhami, M., Mitliagkas, I., Gidel, G.: Performative prediction with neural networks. In: International Conference on Artificial Intelligence and Statistics. pp. 11079–11093. PMLR (2023)
39. Perdomo, J., Zrnic, T., Mendler-Dünner, C., Hardt, M.: Performative prediction. In: International Conference on Machine Learning. pp. 7599–7609. PMLR (2020)
40. Prasad, A., Chandra, S.: PhiUSIIL Phishing URL (Website). UCI Machine Learning Repository (2024), DOI: https://doi.org/10.1016/j.cose.2023.103545
41. Raghavan, M., Barocas, S., Kleinberg, J., Levy, K.: Mitigating bias in algorithmic hiring: Evaluating claims and practices. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 469–481 (2020)
42. Shavit, Y., Edelman, B., Axelrod, B.: Causal strategic linear regression. In: International Conference on Machine Learning. pp. 8676–8686. PMLR (2020)
43. Singh, M.K., Kulkarni, A.A.: Optimal stochastic decision rule for strategic classification. In: 2024 National Conference on Communications (NCC). pp. 1–6. IEEE (2024)
44. Sundaram, R., Vullikanti, A., Xu, H., Yao, F.: PAC-Learning for Strategic Classification. arXiv e-prints arXiv:2012.03310 (Dec 2020). https://doi.org/10.48550/arXiv.2012.03310
45. Vo, K.Q., Aadil, M., Chau, S.L., Muandet, K.: Causal strategic learning with competitive selection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 15411–15419 (2024)
46. Xie, T., Zhang, X.: Non-linear welfare-aware strategic learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. vol. 7, pp. 1660–1671 (2024)
47. Yeh, I.C.: Default of Credit Card Clients. UCI Machine Learning Repository (2016), DOI: https://doi.org/10.24432/C55S3H
48. Zhang, H., Zhang, K., Zhou, S., Guan, J., Zhang, J.: Testing independence between linear combinations for causal discovery. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 6538–6546 (2021)
49. Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., Shen, Z.: Deep stable learning for out-of-distribution generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5372–5382 (2021)
50. Zhang, X., Khalili, M.M., Jin, K., Naghizadeh, P., Liu, M.: Fairness interventions as (Dis)Incentives for strategic manipulation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 26239–26264. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/zhang22l.html
51. Zrnic, T., Mazumdar, E., Sastry, S., Jordan, M.: Who leads and who follows in strategic classification? Advances in Neural Information Processing Systems **34**, 15257–15269 (2021)