THE ART OF FAIRNESS:

BUILDING AI ACCOUNTABILITY

THROUGH OBJECTIVE BIAS ANALYSES

By

JARREN BRISCOE

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

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
School of Electrical Engineering and Computer Science

DECEMBER 2024

THE ART OF FAIRNESS:

BUILDING AI ACCOUNTABILITY

THROUGH OBJECTIVE BIAS ANALYSES

Abstract

by Jarren Briscoe, M.S.
Washington State University
December 2024

Chair: Assefaw Gebremedhin and Diane J. Cook

Artificial intelligence (AI) has spurred transformative advancements across diverse domains, yet persistent biases in machine learning (ML) models remain a significant barrier to equitable decision-making. This proposal addresses the pressing need for objective, interpretable frameworks to assess and mitigate bias in ML, critically analyzing classification metrics, and exploring neural frameworks focused on bias mitigation. We focus on many domains, including high-stakes applications such as criminal justice, healthcare, and environmental forecasting.

We introduce the **Objective Fairness Index (OFI)**, a novel metric grounded in legal fairness principles, designed to provide consistent and legally compliant evaluations of systemic biases. By empirically validating OFI on datasets such as COMPAS recidivism predictions and employment classifiers, this research establishes a shift away from reliance on disparate impact metrics, offering a comprehensive lens for detecting systemic disparities.

Complementing this, the study investigates inherent biases in classification metrics, particularly those induced by sample-size variability, which can mask subgroup inequities. To address these challenges, we propose the **Cross-Prior Smoothing (CPS)** method and the **Metric Alignment Trial for Checking Homogeneity (MATCH)** test. These tools ensure robust metric comparisons across imbalanced data distributions, equipping practitioners with actionable solutions to enhance fairness across diverse representation groups.

Next, we develop two unique neural frameworks: the **Conditional HydraGAN** and the **Probabilistic Parallel Time Network (PTN)**. Conditional HydraGAN generates synthetic data with user-defined diversity quotas in multi-objective scenarios. The PTN is a meta-learning framework carefully tailored to mitigate particular types of biases (including underrepresented and algorithmic bias) with multi-step, probabilistic forecasting. We choose this probabilistic approach for uncertainty transparency, allowing the user to better assess risks. By incorporating static data to contextualize time-series predictions, PTN adapts to underrepresented data subsets, such as microclimates. Furthermore, the model's meta-learning capabilities, with differentiable architecture search (DARTS) and hyperparameter optimization (HPO), improve performance across varied domains, including chaotic systems. PTN aims to enhance fairness and reliability in forecasting applications.

The future work of my dissertation will extend OFI's theoretical underpinnings through formal analyses grounded in social choice theory. With the algorithms and architectures of Conditional HydraGAN and PTN complete within this proposal, we also plan to generalize OFI to further analyze them. This will include applying OFI to multiclassification, regression, and probabilistic forecasting. Furthermore, we aim to improve MATCH's precision of metric cumulative-density function approximations.

By integrating these expansions, this research aims to introduce a legally grounded bias metric, establish a unified framework for critical evaluation of sample-size-induced bias in classification metrics, and promote novel methodologies for fairness enhancement. Ultimately, this work aspires to improve the assessment and catalyze the development of ethical, equitable AI systems across a spectrum of critical applications.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 The Need for Objective Assessments and Accountability

Machine learning (ML) models have become integral to decision-making processes across various high-stakes domains, including healthcare, criminal justice, finance, and more. As these systems increasingly influence critical aspects of human life, concerns about inherent biases within ML algorithms have gained significant traction. Left unchecked, biases in algorithmic predictions can perpetuate or even exacerbate existing societal inequities, posing ethical and legal risks that demand rigorous assessments before use (Obermeyer et al., 2019; Angwin et al., 2022; Das et al., 2021a; Amarasinghe et al., 2023).

Despite advancements in fairness-aware ML techniques, a fundamental gap persists in the form of an objective, legally consistent framework that can comprehensively address fairness across applications involving binary classifications and probabilistic forecasts. Without such a framework, fairness assessments remain subjective and are often inconsistently applied, limiting the reliability of bias mitigation efforts. This proposal addresses this gap, aiming to introduce a framework for detecting and evaluating fairness with objectivity, thereby fostering accountability in AI-driven decision-making.

1.2 Significance of Objective Fairness and Bias Detection

The notion of fairness in ML lacks a universally accepted definition, with interpretations and fairness metrics often varying across contexts (Verma and Rubin, 2018; Amazon, 2021). Popular fairness definitions, such as disparate impact and equalized odds, provide essential tools for assessing algorithmic bias but exhibit limitations in their applicability for objective fairness assessments. For instance, these metrics often fail to consider objective-testing laws and statistical implications, leading to bifurcation from a broad international agreement in legal philosophy and a problematic amount of variability.

To address this challenge of objective testing, this proposal introduces the *Objective Fairness Index (OFI)* as a novel metric for assessing fairness that aligns with legal standards on non-discrimination and integrates principles of objective testing. In law, objective testing is the philosophy that correct assessments found via rational, related tests are not discriminatory. OFI aims to simulate a counterfactual test of "what happened" versus "what should have happened", providing an objective basis for assessing bias in ML systems.

There is a broad international agreement to incorporate the principle of objective testing in non-discrimination laws. According to Chopin and Germaine (2017), 34 out of 35 surveyed EU countries (including Spain, Germany, Switzerland, Turkey, and Romania) have such laws in their national legislation. Similarly, beyond the precedents established within the EU, other countries like the United States (with Supreme Court precedents like Griggs v. Duke Power Co. (1971) and Ricci v. DeStefano (2009) et seq.), South Africa (under the Employment Equity Act), Canada (notably through the Public Service Employee Relations Commission v. British Columbia Government Service Employees' Union), and Australia (via the Fair Work Act) uphold similar standards for objective testing and non-discrimination. Our work with the Objective Fairness Index (OFI) aims to build on these principles by establishing a more universally agreeable and interpretable definition of bias, grounded within these comprehensive legal frameworks and precedents.

The necessity for legally compliant fairness metrics is underscored by foundational rulings and laws, such as the U.S. Civil Rights Act and related non-discrimination statutes, which mandate that any assessment impacting protected groups must be both neutral in intent and effect. Traditional bias metrics, such as disparate impact, are often inadequate to fully capture the context of legal standards in fair testing. To this end, OFI is designed to uphold the principles of fairness consistent with legal mandates by measuring bias in a manner that is interpretable within these contexts.

By grounding bias detection in the objective testing principles of legal frameworks, OFI not only aims to improve fairness assessments, but also ensures that these evaluations are defensible under scrutiny, particularly in high-stakes applications involving sensitive attributes. This alignment with legal standards represents a foundational step toward achieving accountability in AI systems, addressing both societal and regulatory expectations for ethical ML.

1.3 Research Objectives

The objectives of this proposal are centered on establishing an objective framework for bias detection and fairness assessment in AI, specifically within binary classification and time-series forecasting domains. Through legally grounded methodologies and novel assessment methods, this research aims to address gaps in fairness evaluation, improve metric reliability among diverse populations, and provide practical solutions for bias mitigation. The specific objectives are as follows:

- **Develop the Objective Fairness Index (OFI):** In Chapter 3, we introduce the Objective Fairness Index (OFI), a novel metric designed to assess fairness in alignment with legal standards, such as those related to disparate impact and objective testing principles in discrimination law. We seek to bridge AI fairness with legally recognized non-discrimination standards, thus enhancing the interpretability and legal defensibility of fairness metrics in sensitive applications, including criminal justice and healthcare.
- **Address Sample-Size-Induced Bias in Classification Metrics:** In Chapter 4, we investigate how sample size affects traditional classification metrics, leading to potential biases, especially in imbalanced datasets. The research introduces Cross-Prior Smoothing (CPS) and the Metric Alignment Trial for Checking Homogeneity (MATCH) test to stabilize classification metrics across varying sample sizes. These tools are designed to ensure that fairness assessments remain reliable, even when applied to subpopulations with limited data, thus supporting equitable model evaluations across diverse groups.
- **Introduce Probabilistic Parallel Time Networks (PTN) for Bias-Adaptive Forecasting:** In Chapter 5, a new forecasting model called Probabilistic Parallel Time Networks (PTN) is presented. PTN leverages probabilistic forecasting and adaptive modeling to address algorithmic uncertainty and increase user understandability. By incorporating contextual static data, PTN offers dynamic bias corrections in underrepresented data subsets, such as microclimates. Furthermore, PTN's meta-learning and neural architecture search capabilities are tailored to optimize its performance across diverse domains, enhancing its potential for equitable predictions in time-sensitive applications.
- **Transparently Convey Next Steps:** In Chapter 6, we include all plans to complete my doctoral work. First, we propose to build upon the legal principles of OFI into multiclassification, regression, and multi-step probabilistic forecasting domains. Furthermore, we

plan to assess OFI within social choice theory, providing novel insights in the discussion on classification bias. Additionally, we aim to extend MATCH to improve the beta-distribution approximation of Joint Ratio Metrics, enhancing the reliability of fairness assessments across diverse datasets. Finally, we plan to conduct ablation studies on PTN and use OFI to assess its performance across domains such as microclimate forecasting and chaotic systems.

In summary, this proposal seeks to advance the state of AI fairness evaluation by developing innovative methodologies grounded in legal principles and formal theory, improving metric reliability across diverse contexts, and addressing biases in probabilistic forecasting. By integrating these objectives, the research aspires to contribute a robust, legally-informed framework for equitable AI system development, while ensuring practical applicability in real-world scenarios. The results aim to enhance the transparency, reliability, and fairness of AI models, ultimately supporting their adoption in high-stakes domains where bias mitigation is critical.

Chapter 2

BACKGROUND

2.1 Confusion Matrices and Classification

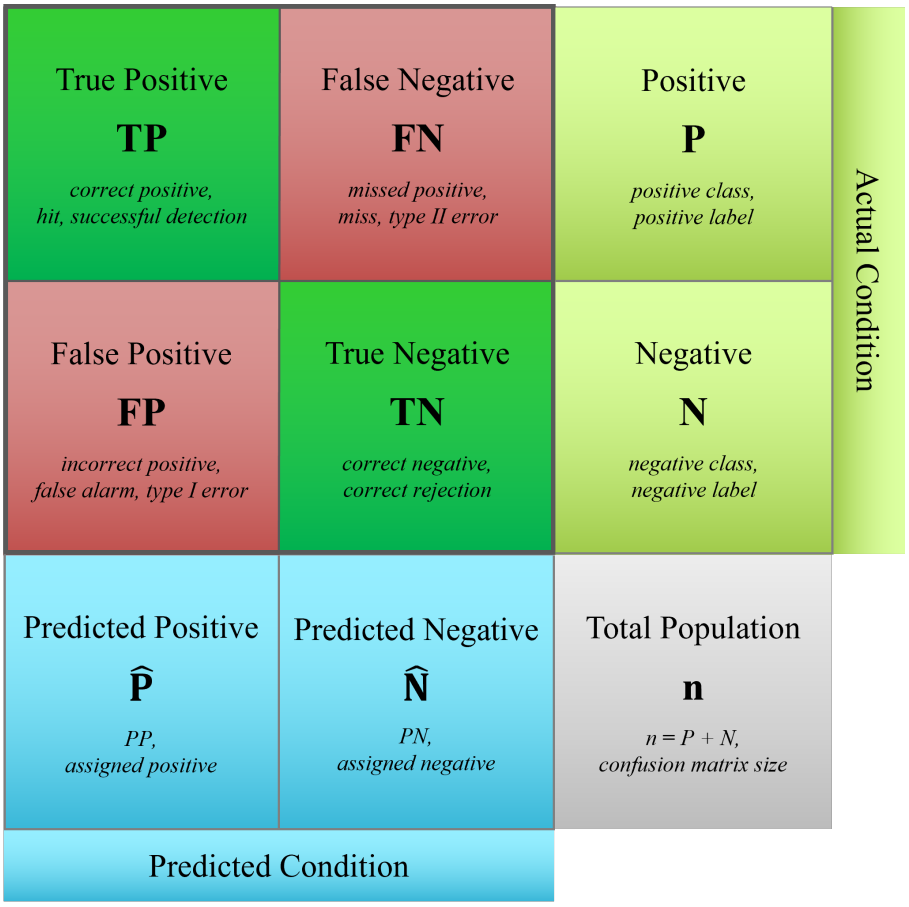


Figure 2.1: The binary confusion matrix illustrates the four primary outcomes of a binary classifier: True Positives, False Negatives, False Positives, and True Negatives. The additional cells summarize overall populations, providing context at a glance.

2.1.1 The Binary Confusion Matrix

The binary confusion matrix, CM, is defined in Definition 2.1. A detailed visualization is provided in Figure 2.1, which further illustrates the components and interactions within CM. This figure and accompanying discussion offer a comprehensive understanding of the matrix, highlighting the importance of each cell in evaluating classifier performance.

Definition 2.1 (Confusion Matrix)**.** *The binary confusion matrix,* CM*, summarizes the performance of a binary classifier and is defined as:*

Table 2.1: Confusion Matrix

	Predicted Positive	**Predicted Negative**
Actual Positive	TP	FN
Actual Negative	FP	TN

Overview: The confusion matrix above categorizes the outcomes of a binary classifier into four primary groups: **True Positives (TP)**, **False Positives (FP)**, **False Negatives (FN)**, and **True Negatives (TN)**. Each category helps evaluate the classifier's performance by describing its predictions compared to the actual labels.

Components of the Confusion Matrix:

- **True Positive (TP)**: Instances correctly predicted as positive.
  - *Alternative Names*: "Correct Positive", "Hit", "Successful Detection".
- **False Negative (FN)**: Instances that are actually positive but predicted as negative.
  - *Alternative Names*: "Missed Positive", "Miss", "Type II Error".
- **False Positive (FP)**: Instances that are actually negative but predicted as positive.
  - *Alternative Names*: "Incorrect Positive", "False Alarm", "Type I Error".
- **True Negative (TN)**: Instances correctly predicted as negative.
  - *Alternative Names*: "Correct Negative", "Correct Rejection".

Additional Notations:

- $\hat{\mathbf{P}}$ **(Predicted Positive Total)**: The total number of instances that the model predicts as positive, regardless of their actual label, also referred to as **PP**.
- $\hat{\mathbf{N}}$ **(Predicted Negative Total)**: The total number of instances that the model predicts as negative, regardless of their actual label, also referred to as **PN**.

- **P (Actual Positive Total)**: The count of all actual positive instances in the dataset.
- **N (Actual Negative Total)**: The count of all actual negative instances in the dataset.
- **n (Total Population)**: The overall size of the dataset, equivalent to $\mathbf{n} = \mathbf{P} + \mathbf{N}$ and $\mathbf{n} = \hat{\mathbf{P}} + \hat{\mathbf{N}}$, representing the count of all instances.

Derived Metrics: Derived metrics, including accuracy, precision, recall, and specificity, serve as scalar summaries of model performance derived from the counts in the confusion matrix. These metrics reduce the complexity of a classifier's predictions into single-valued representations, making them more interpretable and facilitating comparisons across different models and datasets. However, their applicability is context-sensitive; the relevance and informativeness of each metric depend on the specific problem domain and the objectives of the analysis. Each metric emphasizes a distinct aspect of model performance, enabling a more nuanced evaluation of classifier behavior. For example:

- **Accuracy** measures overall correctness.
- **Precision** assesses the relevancy of positive predictions.
- **Recall** evaluates sensitivity to actual positives.
- **Specificity** captures the ability to identify negatives accurately.

These derived metrics offer an accessible and scalar representation of a model's effectiveness, which is particularly useful for comparative analysis and for tracking performance improvements over time. As discussed in greater detail in Chapter 4, we classify these performance metrics into three primary categories: Binomial Metrics, Joint Ratio Metrics (JRM), and specialized metrics, such as fairness measures. Each category provides unique insights that contribute to a comprehensive evaluation of the classifier, thereby refining our understanding of model performance across various dimensions.

2.1.2 Metric Definitions

In this section, we present a detailed overview of the classification metrics used throughout the paper. These metrics are categorized into three groups: Binomial Metrics, Joint Ratio Metrics (JRM), and Other Metrics, including fairness and specialized measures. Each table provides a formula, abbreviations, and brief description of the metric to clarify their definitions and usage. We use the binary confusion matrix from Definition 2.1.

Binomial Metrics

Binomial metrics, such as Accuracy and Prevalence, are fundamental to assessing the overall performance of a classification model. These metrics are based on binomial distributions and provide general insights into model behavior across various datasets. Refer to Table 2.2 for a complete breakdown of these metrics.

Joint Ratio Metrics (JRM)

Joint ratio metrics (JRM) are metrics derived from the confusion matrix, providing insight into the relationships between different types of classification error. These metrics include widely used measures such as True Positive Rate (TPR), False Positive Rate (FPR), and Positive Predictive Value (PPV). A comprehensive list of JRMs, their formulas, and descriptions can be found in Table 2.3.

Other Metrics

In addition to the JRMs and Binomial Metrics, we also include specialized metrics such as our Objective Fairness Index (discussed in Chapter 3), Matthews Correlation Coefficient (Matthews, 1975), and Prevalence Threshold. These metrics extend beyond basic performance evaluation and focus on more complex aspects, including fairness and correlation in imbalanced datasets. See Table 2.4 for further details.

Table 2.2: Binomial Metrics: General performance metrics derived from the binomial distribution of outcomes, commonly used for basic model evaluation.

Metric	Formula	Description
Binomial Metric	$\frac{c_i + c_j}{n}$	A ratio of cells c_i and c_j to the total count n.
Accuracy **ACC**	$\frac{\text{TP} + \text{TN}}{n}$	Overall proportion of correct predictions (both positive and negative).
Prevalence **PREV**	$\frac{\text{TP} + \text{TP}}{n}$	Proportion of actual positive instances in the population. (Synonym: P)
Predicted Positive Rate **PPR**	$\frac{\text{TP} + \text{FP}}{n}$	Fraction of instances predicted as positive by the classifier. (Synonyms: $\hat{\text{P}}$, PP)
Inaccuracy **INACC**	$\frac{\text{FP} + \text{TP}}{n}$	Fraction of incorrect predictions out of total instances. (Synonyms: Error Rate, Misclassification Rate)
Negative Prevalence **NPREV**	$\frac{\text{TN} + \text{FP}}{n}$	Proportion of actual negative instances in the population. (Synonyms: Complement of Prevalence, N)
Predicted Negative Rate **PNR**	$\frac{\text{TN} + \text{TP}}{n}$	Proportion of predicted negative instances in the population. (Synonyms: $\hat{\text{N}}$, PN)

2.2 Probability Distributions

Here, we give necessary preliminaries on probability distributions.

2.2.1 Common Terms

In probability and statistics, a **distribution** describes how the values of a random variable are distributed. This section defines key terms and concepts related to distributions.

A probability distribution is classified as a **discrete distribution** if a random variable X assumes a countable set of distinct values. Conversely, a probability distribution is referred to as a **continuous distribution** if the random variable X can take on infinitely many values within a continuous range.

Key notations associated with distributions are as follows:

- X represents the random variable.
- $\mathcal{X}$ is the *support* of the distribution, defining the set of all possible values that X can assume.

Table 2.3: Joint Ratio Metrics (JRM): Metrics that describe the relationships between confusion matrix components as ratios, commonly used to assess various types of classification performance errors.

Metric	Formula	Description
Joint Ratio Metric **JRM**	$\frac{c_i}{c_i + c_j}$	A ratio between a cell c_i and the sum of c_i and another cell c_j.
True Positive Rate **TPR**	$\frac{\text{TP}}{\text{TP} + \text{TP}}$	Proportion of actual positives correctly identified. (Synonyms: Sensitivity, Recall, Hit Rate)
False Positive Rate **FPR**	$\frac{\text{FP}}{\text{FP} + \text{TN}}$	Proportion of actual negatives incorrectly classified as positive. (Synonym: Fall-Out)
True Negative Rate **TNR**	$\frac{\text{TN}}{\text{TN} + \text{FP}}$	Proportion of actual negatives correctly identified. (Synonym: Specificity)
False Negative Rate **FNR**	$\frac{\text{TP}}{\text{FN} + \text{TP}}$	Proportion of actual positives incorrectly classified as negative. (Synonym: Miss Rate)
Positive Predictive Value **PPV**	$\frac{\text{TP}}{\text{TP} + \text{FP}}$	Proportion of predicted positives that are actual positives. (Synonym: Precision)
Negative Predictive Value **NPV**	$\frac{\text{TN}}{\text{TN} + \text{TP}}$	Proportion of predicted negatives that are actual negatives.
False Discovery Rate **FDR**	$\frac{\text{FP}}{\text{FP} + \text{TP}}$	Proportion of predicted positives that are false positives.
False Omission Rate **FOR**	$\frac{\text{TP}}{\text{FN} + \text{TN}}$	Proportion of predicted negatives that are false negatives. (Synonym: False Reassurance Rate)

Table 2.4: Other Metrics: Specialized metrics focused on fairness, correlation, and balancing different forms of classification errors.

Metric	Formula	Description
F_1 Score $\mathbf{F_1}$	$\frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{TP}}$	Harmonic mean of precision and recall (simplified).
Matthews Correlation Coefficient **MCC**	$\frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{TP}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{TP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{TP})}}$	Balanced metric accounting for all confusion matrix values, robust for imbalanced data.
Prevalence Threshold **PT**	$\frac{\sqrt{\mathrm{TPR} \cdot \mathrm{FPR}} - \mathrm{FPR}}{\mathrm{TPR} - \mathrm{FPR}}$	Threshold at which the positive prediction rate balances misclassification rates.
Treatment Equality **TE**	$\frac{\mathrm{TP}_1}{\mathrm{FP}_1} - \frac{\mathrm{TP}_2}{\mathrm{FP}_2}$	Compares false negatives and false positives between two subgroups.
Marginal Benefit $\mathcal{B}$	$\frac{\mathrm{FP} - \mathrm{TP}}{n}$	Considers objective testing principles in law; represents the benefit gained or lost (cost) for a group.
Objective Fairness Index **OFI**	$\mathcal{B}_1 - \mathcal{B}_2$	The disparity between two subgroups' marginal benefits.

- x denotes a specific realization or observed value of X.
- $X \sim D$ means that X is sampled from the distribution D.

The **probability mass function (PMF)** is used for discrete random variables. It gives the probability that X takes on a specific value x:

$$P(X = x) = f_X(x), \quad \text{where } \sum_x f_X(x) = 1. \tag{2.1}$$

For example, the PMF of a fair six-sided die is:

$$f_X(x) = \begin{cases} \frac{1}{6}, & x \in \{1, 2, 3, 4, 5, 6\}, \\ 0, & \text{otherwise.} \end{cases}$$

The **probability density function (PDF)** is used for continuous random variables. It describes the relative likelihood of a random variable X taking on a value near x. For example, assume $f_X(a) = 0.1$ and $f_X(b) = 0.2$ for $(a, b) \in \mathbb{R}^2$. Then, the following calculates how likely a is compared to b.

$$\text{Relative likelihood of a to b:} \quad \frac{f_X(a)}{f_X(b)} = \frac{1}{2}. \tag{2.2}$$

So, a value X being near a is half as likely to occur as X being near b.

To find the probability that X lies within a range $[a, b]$, we use:

$$P(a \leq X \leq b) = \int_a^b f_X(x)\, dx.$$

The **cumulative distribution function (CDF)** applies to both discrete and continuous random variables. It gives the probability that a random variable X is less than or equal to a value x:

$$F_X(x) = P(X \leq x) = \begin{cases} \sum_{k \leq x} f_X(k), & \text{if } \mathcal{X} \text{ is discrete,} \\ \int_{-\infty}^{x} f_X(t)\, dt, & \text{if } \mathcal{X} \text{ is continuous.} \end{cases} \tag{2.3}$$

The CDF satisfies:

$$F_X(-\infty) = 0, \quad F_X(\infty) = 1.$$

Key Properties

- **Mean (μ)**: The mean represents the expected value of the random variable. It provides a measure of central tendency, indicating where the distribution is centered. For a random variable X, the mean is defined as:

$$\mu = \mathbb{E}[X] = \begin{cases} \sum_{x \in \mathcal{X}} xP(X = x) & \text{for discrete distributions,} \\ \int_{-\infty}^{\infty} x f_X(x)\, dx & \text{for continuous distributions.} \end{cases} \tag{2.4}$$

- **Variance (σ^2)**: The variance quantifies the spread of the distribution around the mean. It is useful for understanding the variability of outcomes, which is critical for risk assessment and uncertainty quantification. Its defined as:

$$\sigma^2 = \mathbb{E}[(X - \mu)^2]. \tag{2.5}$$

- **Skewness (γ_1)**: Skewness measures the asymmetry of the probability distribution. Positive skewness indicates a longer tail on the right, while negative skewness reflects a longer tail on the left. This helps in identifying directional biases in the distribution. It is calculated as:

$$\gamma_1 = \frac{\mathbb{E}[(X - \mu)^3]}{\sigma^3}. \tag{2.6}$$

- **Kurtosis (β_2)**: Kurtosis measures the "tailedness" of the distribution, capturing the sharpness of its peak and the heaviness of its tails. This is to identify bias toward the center or extremes. It is given by:

$$\beta_2 = \frac{\mathbb{E}[(X - \mu)^4]}{\sigma^4}. \tag{2.7}$$

- **Excess Kurtosis (γ_2)**: Excess kurtosis refines the concept of kurtosis by quantifying the deviation of the distribution's tail heaviness and peak sharpness relative to the normal distribution, which has an excess kurtosis of 0. It helps identify whether the data exhibits fatter tails (positive excess kurtosis) or thinner tails (negative excess kurtosis) compared to the normal distribution. We subtract three since the normal distribution's kurtosis is 3.

$$\gamma_2 = \beta_2 - 3 \tag{2.8}$$

These properties are crucial in distribution analyses. They provide insights into the distribution's shape, central tendency, and variability, enabling a comprehensive understanding of the underlying data.

2.2.2 The Gaussian Distribution

The Gaussian distribution, also known as the **normal distribution**, is one of the most fundamental probability distributions in statistics and machine learning. It is characterized by its bell-shaped curve and is mathematically defined in Definition 2.2.

Definition 2.2 (Gaussian Distribution)**.** *The probability density function (PDF) of the Gaussian distribution is given by:*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}, \tag{2.9}$$

where:

- $\mu \in \mathbb{R}$ *is the mean of the distribution,*
- $\sigma^2 > 0$ *is the variance of the distribution.*

Using the PDF, the CDF of a Gaussian random variable X is:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt. \tag{2.10}$$

The Gaussian distribution possesses several key properties that make it widely applicable. Firstly, the distribution is symmetric around the mean, with the mean, median, and mode all coinciding at the center of the distribution. Additionally, the Gaussian distribution is solely characterized by two parameters: the mean μ (location) and variance σ^2 (spread). The standard normal distribution is a common, special case of the Gaussian distribution with $\mu = 0$ and $\sigma^2 = 1$, formally $\mathcal{N}(0, 1)$. The PDF and CDF for $\mathcal{N}(0, 1)$ are publicly available in precomputed standard normal tables.

Additionally, the normal approximation is widely used in practice due to the Central Limit Theorem, which states that the sum of a large number of independent and identically distributed random variables tends to a normal distribution. Furthermore, the Gaussian distribution is linear for independent variables, offering great conveniences. For example, if $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$

and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then:

$$aX_1 + bX_2 \sim \mathcal{N}(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2), \tag{2.11}$$

where $(a, b) \in \mathbb{R}^2$.

We list the first four moments and excess kurtosis of the Gaussian distribution in Table 2.5.

Table 2.5: Moments and Excess Kurtosis of the Normal Distribution

Property	Symbol	Alternative Notation	Mathematical Expression
Mean	μ	$\mathbb{E}[X]$	μ
Variance	σ^2	$\mathrm{Var}(X)$	σ^2
Skewness	γ_1	$\mathrm{Skew}(X)$	0
Kurtosis	β_2	$\mathrm{Kurt}(X)$	3
Excess Kurtosis	γ_2	—	0

2.2.3 The Binomial Distribution

The binomial distribution is a discrete probability distribution that describes the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success. A Bernoulli trial is a random experiment with exactly two possible outcomes: "success" (with probability p) and "failure" (with probability $q = 1 - p$).

Definition 2.3 (The Binomial Distribution)**.** *Mathematically, the binomial distribution can be expressed with its PMF:*

$$P(X = k) = \binom{n}{k} p^k q^{n-k}, \quad \textit{for } k = 0, 1, 2, \ldots, n,$$

where:

- n *is the total number of trials,*
- k *is the number of successes,*
- p *is the probability of success on a single trial,*
- $q = 1 - p$ *is the probability of failure on a single trial,*
- $\binom{n}{k}$ *is the binomial coefficient, given by:*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \tag{2.12}$$

which represents the number of ways to choose k successes from n trials. The binomial coefficient is said as "n choose k", and is also expressed as: $C(n, k)$ or nCk.

Example Suppose a coin is flipped 10 times, and the probability of getting heads (success) in each flip is 0.5. Let X denote the number of heads obtained. Then $X \sim \text{Binomial}(10, 0.5)$. The probability of getting exactly 4 heads is:

$$P(X = 4) = \binom{10}{4}(0.5)^4(1 - 0.5)^6 = \frac{10!}{4! \cdot 6!} \cdot (0.5)^4 \cdot (0.5)^6 \approx 0.205. \tag{2.13}$$

2.2.4 Key Properties

The binomial distribution is characterized by several fundamental properties and assumptions, which form the basis for its application in probabilistic modeling. These include:

- **Support:** The random variable X takes discrete values in the set $\{0, 1, \ldots, n\}$, representing the possible number of successes in n trials.
- **Independence:** Each trial is independent of all others, ensuring that the outcome of one trial does not influence the outcomes of others.
- **Summation Property:** The probabilities of all possible outcomes sum to 1:

$$\sum_{k=0}^{n} P(X = k) = 1, \tag{2.14}$$

  satisfying the requirements of a probability distribution.

The moments and excess kurtosis of the binomial distribution, as outlined in Table 2.6, provide critical insights into the behavior of binomial random variables. These properties serve as essential tools for analyzing and interpreting outcomes in scenarios involving binary experiments.

In Chapter 4, we use these properties to analyze binomial metrics (Table 2.2) and their biases.

2.2.5 The Beta Distribution

The Beta distribution is a continuous probability distribution defined over the interval $[0, 1]$, frequently employed to model probabilities and proportions. It is parameterized by two positive

Table 2.6: Moments and Excess Kurtosis of the Binomial Distribution

Property	Symbol	Alternative Notation	Mathematical Expression
Mean	μ	$\mathbb{E}[X]$	np
Variance	σ^2	$\mathrm{Var}(X)$	npq
Skewness	γ_1	$\mathrm{Skew}(X)$	$\frac{q-p}{\sqrt{npq}}$
Kurtosis	β_2	$\mathrm{Kurt}(X)$	$\frac{1-6pq}{npq} + 3$
Excess Kurtosis	γ_2	—	$\frac{1-6pq}{npq}$

shape parameters, α and β, commonly referred to as "concentrations". These parameters represent pseudo-counts: α corresponds to the number of successes, and β to the number of failures. As unnormalized quantities, they convey the strength of prior evidence; larger values imply stronger beliefs. When $\alpha = \beta$, the distribution is symmetric around 0.5.

In Bayesian statistics, the relationship $\alpha + \beta = n + 1$ holds, where n denotes the number of trials, and the constant 1 reflects a non-informative prior belief derived from the assumption that all outcomes are equally likely.

Practical Example Suppose a (possibly biased) coin is flipped $n = 10$ times, with $k = 7$ heads observed. Using a uniform prior $\mathrm{Beta}(1, 1)$, which assumes no prior knowledge about the coin's bias, the posterior becomes:

$$p \mid k, n \sim \mathrm{Beta}(1 + 7, 1 + 3) = \mathrm{Beta}(8, 4). \tag{2.15}$$

This posterior distribution reflects our updated belief about the coin's bias, favoring probabilities close to $\frac{8}{12} \approx 0.667$.

Definition 2.4 (The Beta Distribution)**.** *The probability density function (PDF) of the Beta distribution is expressed as:*

$$f_X(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}, \quad 0 \le x \le 1, \tag{2.16}$$

where

- $\alpha > 0$ *and* $\beta > 0$ *are the shape parameters,*

- $\mathrm{B}(\alpha, \beta)$ *is the Beta function, defined as:*

$$\mathrm{B}(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{2.17}$$

- $\Gamma(\cdot)$ *is the Gamma function, which generalizes the factorial function to real and complex numbers.*

The cumulative distribution function (CDF) of the Beta distribution is computed using the regularized incomplete Beta function:

$$F_X(x) = I_x(\alpha, \beta) = \frac{B(x; \alpha, \beta)}{B(1; \alpha, \beta)}, \tag{2.18}$$

where $B(x; \alpha, \beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\, dt$ is the incomplete Beta function.

The first four moments and the excess kurtosis of the Beta distribution are summarized in Table 2.7.

Table 2.7: Moments and Excess Kurtosis of the Beta Distribution

Property	**Symbol**	**Alternative Notation**	**Mathematical Expression**
Mean	μ	$\mathbb{E}[X]$	$\frac{\alpha}{\alpha+\beta}$
Variance	σ^2	$\mathrm{Var}(X)$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Skewness	γ_1	$\mathrm{Skew}(X)$	$\frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$
Kurtosis	β_2	$\mathrm{Kurt}(X)$	$\frac{6\big((\alpha-\beta)^2(\alpha+\beta+1)-\alpha\beta(\alpha+\beta+2)\big)}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)} + 3$
Excess Kurtosis	γ_2	—	$\beta_2 - 3$

Chapter 3

DEFINING AND MEASURING FAIRNESS IN AI: OBJECTIVE FAIRNESS INDEX (OFI)

Leveraging current legal standards, we define bias through the lens of marginal benefits and objective testing with the novel metric "Objective Fairness Index". This index combines the contextual nuances of objective testing with metric stability, providing a legally consistent and reliable measure. Utilizing the Objective Fairness Index, we provide fresh insights into sensitive machine learning applications, such as COMPAS (recidivism prediction), and bias mitigation analyses, highlighting the metric's practical and theoretical significance. The Objective Fairness Index allows one to differentiate between discriminatory tests and systemic disparities.

3.1 Introduction

In the 2010s, machine learning took major leaps forward in everyday life. However, we soon realized that bias is rampant in machine learning models. One of the most prominent examples regarding COMPAS, which deals with Americans' civil rights in the United States justice system (Brennan et al., 2009; Dieterich et al., 2016). Researchers are actively working to mitigate bias in general in a myriad of ways (Dai et al., 2022; Damak et al., 2022; Wang et al., 2022; Mehrabi et al., 2021; Caton and Haas, 2020; Corbett-Davies and Goel, 2018; Zhang et al., 2020; Hong et al., 2021; Kwon et al., 2023; Russo and Tonia, 2023; Zhang et al., 2023; Cornacchia et al., 2023; Gao et al., 2023; Le and Deng, 2023; Briscoe et al., 2021). Despite being the cornerstone of many proposals, definitions of bias remain inconsistent (Caton and Haas, 2020; Verma and Rubin, 2018). Due to various philosophies and situations, there is no consensus on a formal definition. This work focuses on the legal context, namely the disparate impact metric (DI). DI is useful as a flag or reparative measure, however it is insufficient in objective-testing laws.

To address this gap, we introduce the Objective Fairness Index (OFI), extending our work in (Briscoe et al., 2024). We formalize OFI by defining bias as the difference between marginal

benefits, derived via the lens of objective testing. In law, objective testing is the philosophy that correct assessments found via rational, related tests are not discriminatory. Additionally, we provide applications to two critical bias scenarios: COMPAS and Folktables' adult employment dataset.

The principle of objective testing in non-discrimination laws is recognized not only in the US but worldwide. According to Chopin and Germaine (2017), 34 out of 35 surveyed countries (including Spain, Germany, Switzerland, Turkey, and Romania) have such laws in their national legislation. Going beyond the precedents found in the U.S. and the 34 other countries, we find that South Africa (Employment Equity Act), Canada (Public Service Employee Relations Commission v British Columbia Government Service Employees' Union), and Australia (Fair Work Act) uphold similar legal standards. Our work with the Objective Fairness Index aims to establish an agreeable and naturally interpretable definition of bias, grounded in these comprehensive legal frameworks and precedents.

We make three key contributions in this chapter:

- We introduce the Objective Fairness Index (OFI) and substantially support it with insights from the legal literature, highlighting the gap of bias metrics in the literature.
- We bring novel insights into two popular ML fairness applications: COMPAS (predicts recidivism) and Folktables' Adult Employment (predicts employment).
- Using a different dataset, (CASAS' smart home data) we demonstrate how OFI is used to assess bias and bias mitigation. Bias mitigation is achieved via a custom GAN: Conditional HydraGAN.

3.2 Limitations of Disparate Impact

This section gives context to the founding of the Objective Fairness Index by reviewing the Supreme Court of the United States' decision that spurred disparate impact and a following precedent. Under objective-testing laws, we show that OFI is better than DI in identifying discriminatory tests.

3.2.1 Legal Foundations of Disparate Impact

Disparate impact's significance comes from United States labor laws regarding the biased treatment of protected groups. In Griggs v. Duke Power Co., 401 U.S. 424 (1971), the Supreme

Court unanimously agreed that employers must prove that any test or assessment used must be related to job performance (objective). Written by Chief Justice Burger, the decision says that a lower admittance rate of a protected class warrants further investigation—even if such discrimination may be unintentional. Since this precedent was passed from a lawsuit under Title VII of the Civil Rights Act of 1964, the Court's ruling extends to decision-making whenever protected classes are concerned.

To summarize this ruling, employers must have objective and fair tests (neutral in form) and should not inadvertently advantage a protected class (neutral in impact). Should there be a disparate impact, the employer is subject to scrutiny and bears the burden of proof that the test is objective and "consistent with business necessity". The disparate impact bias metric is defined in Definition 3.1. Where $P = \text{FN} + \text{TP}$ denotes the amount of positive labels, $\hat{P} = \text{FP} + \text{TP}$ denotes the amount of positive predictions, $N = \text{FP} + \text{TN}$ denotes the amount of negative labels, and $\hat{N} = \text{FN} + \text{TN}$ denotes the amount of negative predictions.

Definition 3.1. *Disparate impact compares the rates of positive outcomes between two groups using ratios.*

$$DI \triangleq \hat{P}_i/n_i \div \hat{P}_j/n_j \tag{3.1}$$

The implication of Equation 3.1 is that $DI = 1$ suggests no bias. However, the real world can have small and non-representative sampling, causing unbiased decisions to be $DI \neq 1$. As such, a Four-Fifths Rule (80% Rule) has been established in law as an initial screening tool to identify potential discrimination (Biddle, 2017). This rule is indoctrinated in 29 CFR § 1607.4 - Information on impact by the U.S. Equal Employment Opportunity Commission (1978). This rule suggests that if $DI > 5/4$ then group i has positive bias, if $DI < 4/5$ then there is positive bias for j, otherwise DI finds no bias.

In Griggs v. Duke Power Co., the Court goes on to note that Congress does not intend to require businesses to give jobs to unqualified candidates, and that context should still be considered, emphasizing that truly objective assessments are not discriminatory. However, because disparate impact lacks situational context, it cannot capture this legal nuance.

Disparate impact has been influential enough for employers to discard assessments that fail the heuristic to prevent any federal investigation or social uprising, regardless of the assessments' objectivity. Ricci v. DeStefano, 557 U.S. 557 (2009) is the first of these cases to be argued in the Supreme Court when New Haven, Connecticut dismissed objective assessments that benefited the white and Hispanic firefighters over other protected classes. The Court ruled

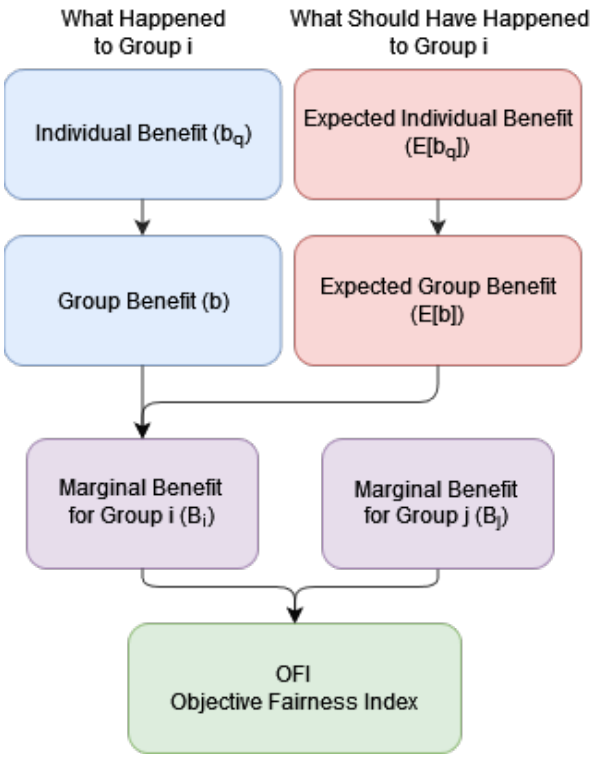


Figure 3.1: Objective Fairness Index's conceptual framework.

in the firefighters' favor—that omitting an objective performance test solely due to differing acceptance rates (disparate impact) is a violation of the rights of those who correctly benefit from such a test. This provides us with a use case where equal positive prediction rates (which became zero after the test's omission) are discriminatory.

3.2.2 Defining Benefit and Bias

We define bias as a function of what happened to the group (benefit) and what should have happened to the group (expected benefit). An objective test will show that benefit equals expected benefit. Consistent with common understanding and legal precedents, we define benefit as an advantage that is gained from a decision or action, such as hiring from a machine-learning prediction. We achieve formal definitions of these terms using binary confusion matrices, allowing generalization to situations where information such as specific features cannot be used due to privacy, legal protection, or information loss.

We define an individual benefit in Definition 3.2 to help describe a group's benefit in Definition 3.3. Let us assume the positive prediction is the beneficial prediction in all definitions.

Definition 3.2. *An individual q's benefit* b_q *is the individual's prediction.* $b_q = 1$ *if predicted*

positive, otherwise 0.

Definition 3.3. *A group's benefit is the arithmetic mean of all individual benefits within the group.*

$$b \triangleq \frac{1}{n}\sum_{\forall b_q} b_q = \frac{1\cdot\hat{P}+0\cdot\hat{N}}{n} = \frac{\mathrm{TP}+\mathrm{FP}}{n} \tag{3.2}$$

The Objective Fairness Index also integrates the expected benefit of an individual (Definition 3.4) and the group's expected benefit (Definition 3.5) by using the actual labels.

Definition 3.4. *An individual's expected benefit* $\mathbb{E}[b_q]$ *is their actual label.* $\mathbb{E}[b_q] = 1$ *if labeled positive, otherwise 0.*

Definition 3.5. *A group's expected benefit is the arithmetic mean of all individual expected benefits within the group.*

$$\mathbb{E}[b] \triangleq \frac{1}{n}\sum_{\forall b_q} \mathbb{E}[b_q] = \frac{1\cdot P - 0\cdot N}{n} = \frac{\mathrm{TP}+\mathrm{FN}}{n} \tag{3.3}$$

Formally defining benefit and expected benefit allows a counterfactual test of "what happened" versus "what should have happened". As mentioned, this is of paramount importance in law. In Ricci v. DeStefano (2009), the city of New Haven omitted their test solely based on the ratio of benefits to be obtained by each protected class of firefighters. However, this discriminates against the firefighters qualified by this objective test. Pursuant to the Supreme Court's ruling, We present a solution by scoring this group's treatment as the "marginal benefit" ($\mathcal{B}$ in Definition 3.6).

Definition 3.6. *Marginal benefit (*$\mathcal{B}$*) is the difference between a group's benefit and expected benefit.*

$$\mathcal{B} \triangleq b - \mathbb{E}[b] = (\mathrm{FP}-\mathrm{FN})/n \tag{3.4}$$

This marginal benefit score suggests: $\mathcal{B} < 0$ an unjust lack of benefit; $\mathcal{B} = 0$ an appropriate amount of benefit; $\mathcal{B} > 0$ an unjust surplus of benefit. With marginal benefit defined, We now formalize the definition of the Objective Fairness Index in Definition 3.7. We provide a graphical representation in Figure 3.1.

Definition 3.7. *The Objective Fairness Index is the difference between two groups' marginal benefits.*

$$\mathrm{OFI} \triangleq \mathcal{B}_i - \mathcal{B}_j = \frac{\mathrm{FP}_i - \mathrm{FN}_i}{n_i} - \frac{\mathrm{FP}_j - \mathrm{FN}_j}{n_j} \in [-2, 2] \tag{3.5}$$

3.2.3 Improvement over Disparate Impact

To illustrate the Objective Fairness Index's significance, consider the instances in Table 3.1, similar to the case of New Haven and its firefighters. The same objective test is given in all scenarios. In Scenario A, the employer fulfills their promise and promotes (OFI = −0.06, DI = 0.38). In Scenario B, the test is discarded (OFI = 0.22, DI = 1). Note Scenario B's DI score is undefined. However, since the philosophy of DI suggests equal positive prediction rates indicate no bias, we use DI = 1.

Table 3.1: Comparison of OFI, DI, and Law (Ricci v. DeStefano) Across Different Scenarios

Group	Scenario	TP	FN	FP	TN	n	b	$\mathbb{E}[b]$	$\mathcal{B}$	OFI: $\mathcal{B}_i - \mathcal{B}_j$	DI: b_i/b_j	Ricci v. DeStefano
i	A	1	0	0	5	6	1/6	1/6	0	−0.06	0.38	Likely no bias
j	A	7	0	1	10	18	8/18	7/18	1/18			
i	B	0	1	0	5	6	0/6	1/6	−1/6	0.22	NaN or 1	Bias for i
j	B	0	7	0	11	18	0/18	7/18	−7/18			
i	α	1	1	0	5	7	1/7	2/7	−1/7	0.23	2.71	Bias for i
j	α	1	7	0	11	19	1/19	8/19	−7/19			

With OFI's absolute value being greater in Scenario B than in Scenario A, OFI is consistent with the ruling of Ricci v. DeStefano (2009). However, DI disagrees with this precedent, suggesting for strong bias in Scenario B and none in Scenario A.

In Scenario α, we conduct a thought experiment with a smoothing technique. Scenario α adds one qualified individual to each group that benefits (e.g., is correctly promoted) to Scenario B. Adding this single correct prediction changes DI's score from 1 to 2.71, suggesting a transition from no bias to heavy bias for group i. However, OFI adjusts from 0.22 to 0.23 in the same Scenario α, illustrating its robustness.

Unlike DI, the Objective Fairness Index indicates the correct directional bias in both scenarios. Additionally, OFI's defined-everywhere property and robustness against small changes showcase its stability. However, note that DI is complementary to OFI in systemic disparity detection. With OFI, we can identify if an algorithm is at fault (OFI $\not\approx$ 0) or if a systematic disparity exists outside of the algorithm (OFI $\approx$ 0 and DI $\not\approx$ 1).

Table 3.2: Popular confusion-matrix-based bias metrics and checks for satisfying desirable properties. Each metric contrasts group i with group j. For all metrics, a greater score indicates more bias toward group i. Abbreviations: Satisfies Objective Testing (OBJ), Real-Valued ($\mathbb{R}$), Directed (DIR), Symmetric (SYM), Bounded in Range (BND), Defined Everywhere (ALL).

Metric	Formula	OBJ	$\mathbb{R}$	DIR	SYM	BND	ALL
Accuracy Difference	$ACCD \triangleq \frac{\mathrm{TP}_i+\mathrm{TN}_i}{n_i} - \frac{\mathrm{TP}_j+\mathrm{TN}_j}{n_j}$	×	✓	✓	✓	✓	✓
MCC Difference	$MCCD \triangleq MCC_i - MCC_j$	×	✓	✓	✓	✓	×
Disparate Impact	$DI \triangleq \frac{\hat{P}_i}{n_i} \div \frac{\hat{P}_j}{n_j}$	×	✓	✓	×	×	×
Predictive Parity	$PP \triangleq \frac{\mathrm{TP}_i}{P_i} - \frac{\mathrm{TP}_j}{P_j}$	×	✓	✓	✓	✓	×
Treatment Equality	$TE \triangleq \frac{\mathrm{FN}_i}{\mathrm{FP}_i} - \frac{\mathrm{FN}_j}{\mathrm{FP}_j}$	×	✓	✓	✓	×	×
Equalized Odds	$\mathrm{EO} \triangleq (\mathrm{FPR}_i = \mathrm{FPR}_j) \wedge (\mathrm{TPR}_i = \mathrm{TPR}_j)$	×	×	×	✓	✓	×
Statistical Parity	$SP \triangleq \hat{P}_i = \hat{P}_j$	×	×	×	✓	✓	✓
Average Absolute Odds Difference	$AAOD \triangleq \frac{1}{2}\left(\left\lvert\frac{\mathrm{FP}_i}{N_i} - \frac{\mathrm{FP}_j}{N_j}\right\rvert + \left\lvert\frac{\mathrm{TP}_i}{P_i} - \frac{\mathrm{TP}_j}{P_j}\right\rvert\right)$	×	✓	×	✓	✓	×
Difference in Conditional Acceptance	$DCA \triangleq \frac{P_i}{\hat{P}_i} - \frac{P_j}{\hat{P}_j}$	×	✓	✓	✓	×	×
Difference in Conditional Rejection	$DCR \triangleq \frac{N_j}{\hat{N}_j} - \frac{N_i}{\hat{N}_i}$	×	✓	✓	✓	×	×
Difference in Positive Proportion & Labels	$DPPL \triangleq \frac{\hat{P}_i}{n_i} - \frac{\hat{P}_j}{n_j}$	×	✓	✓	✓	✓	✓
Objective Fairness Index	$OFI \triangleq \frac{\mathrm{FP}_i-\mathrm{FN}_i}{n_i} - \frac{\mathrm{FP}_j-\mathrm{FN}_j}{n_j}$	✓	✓	✓	✓	✓	✓

3.3 Desirable Metric Properties

Ideally, metrics should be explainable and have compelling reasons for their existence such as legal precedents. We present and define five desirable traits for all metrics: real-valued, directed, symmetric, bounded, and defined everywhere. Next, we define how a metric satisfies objective testing. Finally, we present a table comparing many popular bias metrics in Table 3.2. We introduce two new symbols: s for the result of a bias metric M, and p to represent a scalar pivot. A pivot only occurs in directional metrics and is the pinpoint of bias direction. The subscript ϕ is a reminder that the symbol represents a property.

3.3.1 Real-Valued

Real-valued metrics provide a continuous scale to quantify the bias difference between groups, offering a nuanced understanding of bias magnitude. The advantage of real-valued metrics over Boolean or categorical metrics is their ability to capture a wide range of outcomes and degrees of bias, rather than merely indicating the presence or absence of bias. This property is crucial for fine-grained analysis and comparison across different models or systems. For instance, Equalized Odds is introduced as being binary in its basic form, yet adapting the metric into a real-valued measure by considering the differences in false positive rates (FPR) and true positive rates (TPR) between groups highlights degrees of bias and direction, allowing for more detailed analyses.

Real-valued metrics are essential for creating a detailed and actionable understanding of bias, allowing law enforcement and other stakeholders to prioritize issues based on the magnitude of disparity. However, the interpretation of these values requires careful consideration, as the scale and range can vary significantly between different metrics, potentially leading to confusion or misinterpretation. We formally define real-valued properties in Definition 3.8.

Definition 3.8. *The real-valued property R_ϕ is satisfied when a metric can give any real value within some non-empty range given by (a, b) where a and b are defined by the metric's range and are in $(-\infty, \infty)$.*

$$\mathbb{R}_\phi \triangleq s \in \mathbb{R} \cap (a, b) \tag{3.6}$$

3.3.2 Direction

Direction is necessary to understand which group is favored or disadvantaged by a model's predictions, providing essential insights beyond the mere presence or amount of bias. This property is vital for diagnosing the nature of bias and guiding corrective actions. By identifying the direction of bias, practitioners can tailor interventions to support the disadvantaged group, whether through re-balancing training data, adjusting model parameters, or applying post-processing fairness techniques.

However, the utility of directional information depends on the clarity and consistency of its definition across metrics. Ambiguities in how direction is calculated or interpreted can undermine its effectiveness in fairness assessments. Therefore, a standardized approach to determining and reporting direction can enhance the utility of this property in bias metrics.

In the literature, we find that this approach often comes with a pivot where the pivot indicates the absence of bias. If the bias score is less than the pivot, group j holds the advantage. On the other hand, if the score is greater than the pivot, group i holds the advantage.

Upon all reviewed directional bias metrics, we find that they are split into two categories: ratios and Manhattan distance. For ratios such as disparate impact, one is the pivot. Otherwise, Manhattan-distance directional metrics (including OFI) have zero as the pivot. We formally define the directional property in Definition 3.9.

Definition 3.9. *The directional property D_ϕ is satisfied when a metric indicates which group holds the advantage using a pivot p and bias score s.*

$$D_\phi \triangleq s \begin{cases} \textit{suggests bias for group } i \textit{ and against group } j & \textit{if } s > p; \\ \textit{suggests no bias} & \textit{if } s = p; \\ \textit{suggests bias against group } j \textit{ and for group } i & \textit{if } s < p. \end{cases} \tag{3.7}$$

3.3.3 Symmetry

Symmetry ensures that the metric yields the same magnitude of bias, regardless of which group is denoted as "group i" or "group j". This property simplifies the analysis by eliminating the need to consider group order in the calculation, thereby enhancing the interpretability and comparability of results.

The challenge with ensuring symmetry lies in the design of the metric itself. Metrics that naturally lend themselves to symmetry, such as differences or ratios, are preferable. Disparate impact is a ratio of ratios and is therefore asymmetric. However, OFI and others are symmetric. We define symmetry in Definition 3.10.

Definition 3.10. *The symmetric property S_ϕ of a metric M yields equal magnitudes of bias that are suitable for direct comparison.*

$$S_\phi \triangleq |M(i,j)| = |M(j,i)| \tag{3.8}$$

3.3.4 Bounded

Bounded metrics have a fixed range, making them easier to interpret and compare across different contexts. A bounded range, such as $[0,1]$ or $[-1,1]$, provides a clear reference point

for evaluating the severity of bias, with the bounds representing the absence or extremity of bias. This property is particularly useful for non-experts or stakeholders who need to understand bias metrics without delving into complex statistical details. Disparate impact is unbounded while OFI is bounded.

The primary challenge with bounded metrics lies in their score distribution and ensuring that it accurately reflects the nuances of bias across the entire range. Metrics that are too coarse (e.g., binary scores) or that saturate easily (e.g., sigmoidal function) may not capture subtle variations in bias, while those that are too sensitive might fluctuate wildly for minor changes in the underlying data or model behavior.

We satisfy checks for coarseness with the real-value property. Additionally, none of the metrics we review inherently saturate easily. Nonetheless, some situations may cause metrics to saturate. Thus, checking for context-dependent saturation requires more information such as the model's hypothesis distribution and its data. We define the bounded properly in Definition 3.11.

Definition 3.11. *The bounded property B_ϕ is satisfied if the metric M yields scores bounded by finite scalars a and b, or if the set of possible scores is finite.*

$$B_\phi \triangleq s \begin{cases} s \in [a, b] & \text{if } M \models R_\phi; \\ s \text{ is an element of a finite set} & \text{otherwise.} \end{cases} \tag{3.9}$$

3.3.5 Universal Applicability

Metrics defined for any possible confusion matrix (CM) and population size ($n > 0$) are robust and versatile, capable of handling a wide variety of scenarios without becoming undefined or losing meaning. The universal-applicability property ensures that the metric can be consistently applied across different datasets and models, regardless of the size or distribution of the groups being compared. This is especially important when considering groups with a low sample size.

Achieving this property requires careful mathematical formulation to avoid divisions by zero, undefined logarithms, or other issues that can arise in edge cases. Metrics that are not universally applicable may require additional handling or assumptions, complicating their use and interpretation. We formally define the universal-applicability property in Definition 3.12.

Definition 3.12. *The universal-applicability property U_ϕ is satisfied for any metric being defined*

everywhere for any non-empty group ($n > 0$). $\mathbb{S}$ describes the set of all valid scores from M.

$$U_\phi \triangleq M(\text{CM}) \mapsto \mathbb{S},\ \forall \text{CM}\ |_{n>0} \tag{3.10}$$

3.3.6 Objective Testing

Objective testing, a concept rooted deeply in legal precedents such as Griggs v. Duke Power Co. (1971) et seq., evaluates bias through a lens that compares what happened (actual benefit) against what should have happened (expected benefit) under a non-discriminatory system. This approach is encapsulated in the Objective Fairness Index (OFI) metric, which uniquely satisfies this property by employing a counterfactual analysis that incorporates both actual and potential outcomes.

An objective-bias metric performs a counterfactual test against benefit b with expected benefit $\mathbb{E}[b]$, like $\mathcal{B}$ does. We define the objective-testing property in Definition 3.13. This is the only property highlighting the intra-group case.

Definition 3.13. *The intra-objective-testing property ($\mathcal{B}_\phi$) is satisfied when some metric M increases or decreases as b does, and does the opposite for $\mathbb{E}[b]$. Furthermore, it recognizes context by having correct predictions (TP and TN) be evidence for objectivity, lowering the bias score respectively. We informally summarize this in Equation 3.11 below. We assume that only one sample is added or removed in each scenario.*

$$\mathcal{B}_\phi \triangleq \begin{cases} M(b, \mathbb{E}[b]) \uparrow & \text{as } b \uparrow \text{ or } \mathbb{E}[b] \downarrow \\ M(b, \mathbb{E}[b]) \downarrow & \text{as } b \downarrow \text{ or } \mathbb{E}[b] \uparrow \\ M(b, \mathbb{E}[b]) \text{ goes toward } 0 & \text{as correct values} \uparrow \\ M(b, \mathbb{E}[b]) \text{ moves away from } 0 & \text{as correct values} \downarrow \end{cases} \tag{3.11}$$

OFI compares two expressions satisfying $\mathcal{B}_\phi$, finding the disparity of marginal benefit. As such, OFI satisfies the inter-objective-testing property as defined in Definition 3.14. The inter-objective-testing property may be shortened to "objective-testing property" or O_ϕ.

Definition 3.14. *The objective-testing property (O_ϕ) is satisfied when the metric $M(B_i, B_j)$ takes one parameter for each group, both satisfying $\mathcal{B}_\phi$. Additionally, O_ϕ finds bias for group i (as*

decided from threshold p) iff B_i is greater than B_j.

$$O_\phi \triangleq M(B_i, B_j) > p \iff B_i > B_j \; s.t. \; B \models \mathcal{B}_\phi \tag{3.12}$$

From Table 3.2, we see that OFI is the only objective-testing metric and the only metric to satisfy all desirable properties and is the only one to satisfy the objective-testing property.

Nuances

In this section, we further illustrate how OFI complies with O_ϕ unlike two deceptively similar metrics: treatment equality (TE) and difference in conditional acceptance (DCA). In Table 3.3, we give Scenarios C and D for groups i and j, the scores obtained by each metric, and the threshold for when that metric determines bias against group i. We define confusion matrices in the form [TP, FN, FP, TN].

Table 3.3: Comparative Analysis of Deceptively Similar Bias Metrics, Highlighting OFI's Unique Consideration of Objective Testing

Description	Scenario C	Scenario D	Threshold
Group i's CM	$[1, 2, 1, 1]$	$[1, 2, 1, 1]$	-
Group j's CM	$[1, 1, 2, 2]$	$[1, 1, 2, 4]$	-
TE = $\mathrm{FN}_i/\mathrm{FP}_i - \mathrm{FN}_j/\mathrm{FP}_j$	1.5	1.5	TE > 0
DCA = $P_i/\hat{P}_i - P_j/\hat{P}_j$	0.83	0.83	DCA > 0
OFI = $\mathcal{B}_i - \mathcal{B}_j$	**-0.37**	**-0.33**	OFI < 0

Table 3.3 compares a scenario where Group i's CM is stagnant and Group j gains three more TNs in Scenario D. This highlights how neither TE nor DCA recognizes the information of three more correct predictions. However, OFI captures the subtlety by calculating the marginal benefit for each group and scenario. We only claim that OFI uniquely satisfies the objective-testing property, and stress that reparative metrics such as TE or DI be considered as necessary.

OFI stands out by not only quantifying the discrepancy between groups' marginal benefits but also providing a means to objectively test for bias across different groups, making it invaluable for legal consistency. Importantly, OFI is not designed to be reparative but rather to identify and quantify bias, aligning closely with the objective testing sought in legal contexts. When appropriate, we recommend using a reparative metric such as DI. For example, South Africa's Employment Equity Act (Section 6.2) requires considering affirmative action and objective

testing.

3.4 Applying OFI: Empirical Case Studies

We explore the application of the Objective Fairness Index in practical scenarios: COMPAS (predicts recidivism), and Folktables' Adult Employment dataset (predicts employment).

For COMPAS, we obtain the confusion matrix from actual COMPAS predictions and cases of recidivism occurring within two years (Angwin et al., 2016a; Gursoy and Kakadiaris, 2022). Since this work assumes that the positive label is beneficial, we flip the COMPAS positive label to be "not a recidivist".

Figure 3.2 shows bias scores for each ethnicity pair in the COMPAS dataset. Here, we see that OFI affirms DI's findings. This is an important revelation, as some argue that the disparate prediction rates are correct. In return, some argue that the labeled dataset suffers from systematic bias. However, this case illustrates that OFI can sidestep the need to prove incorrect ground labels, and shows that even if one assumes correct ground labels, there is algorithmic bias against certain races in COMPAS (OFI $\not\approx$ 0). We believe this revelation will encourage more people to take corrective measures.

For the Folktables dataset, we analyze data from the Georgia census (393,236 observations) to predict employment status using a Random Forest classifier (RF) and Naïve Bayes (NB). In RF, OFI corroborates the findings from DI. However, OFI reveals differences to DI in the NB experiment (see Figure 3.3). Here, DI suggests a positive bias for Pacific Islanders in 7/8 cases. In contrast, OFI shows positive bias in 2/8 cases.

With the lens of objective testing, we find that Pacific Islanders suffer from algorithmic bias, despite having high disparate impact scores. These studies demonstrate OFI confirming suspicions of algorithmic bias (COMPAS), and when protected classes suffer from algorithmic bias, despite a high positive prediction rate.

3.5 Evaluating Conditional HydraGAN for Bias Mitigation with the Objective Fairness Index

In this section, we evaluate Conditional HydraGAN's effectiveness in mitigating bias in the CASAS dataset using the Objective Fairness Index (OFI) and other bias metrics. Our analysis

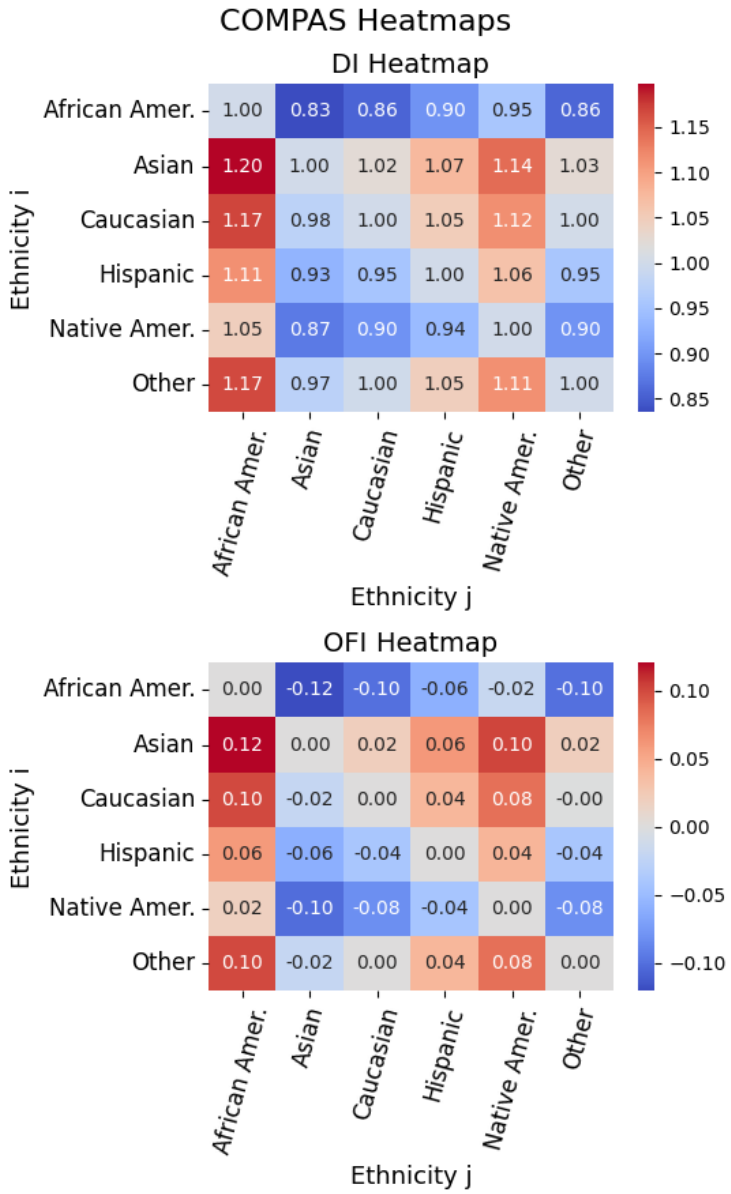


Figure 3.2: OFI shows that COMPAS manifests algorithmic bias in addition to manifesting disparate impact (DI). We use the predictions and labels published in Angwin et al. (2016a).

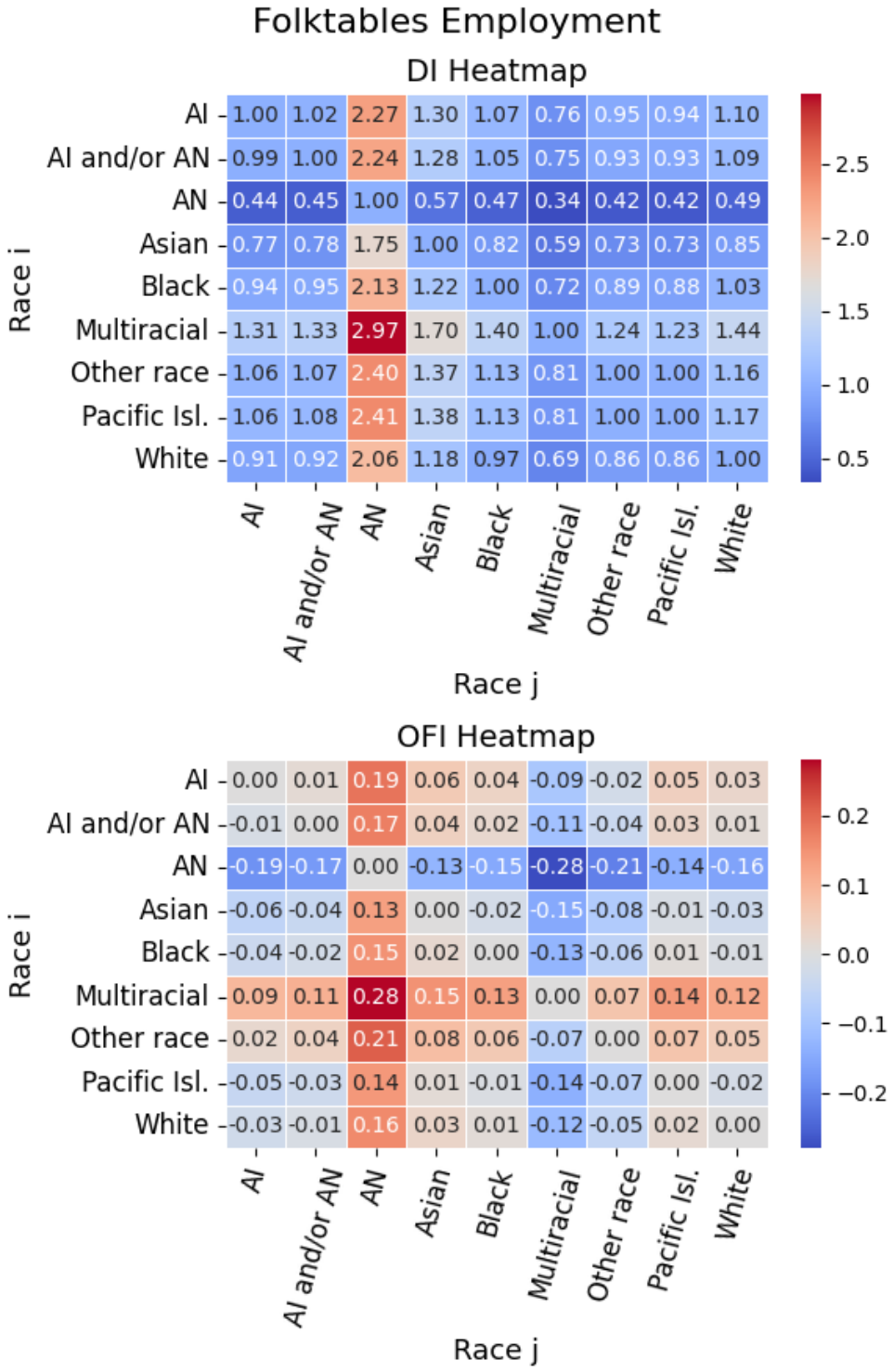


Figure 3.3: DI conveys that Pacific Islanders have positive bias for 7/8 comparisons. However, OFI shows that Pacific Islanders suffer from algorithmic bias in 6/8 cases. **Key:** AI is American Indian, AN is Alaska Native, Pacific Isl. is Pacific Islander.

demonstrates the utility of OFI in evaluating bias mitigation strategies and promoting fairness in healthcare outcomes.

3.5.1 Background

In 2014, Goodfellow et al. introduced the Generative Adversarial Network (GAN), which became highly influential for its capability to produce realistic synthetic data. A GAN consists of two neural networks, a generator and a discriminator, which engage in a competitive, two-player, zero-sum game. The generator takes a random latent vector and generates synthetic data, aiming to deceive the discriminator into classifying this data as real. The discriminator, trained using binary cross-entropy, evaluates both real and synthetic data, attempting to distinguish between them.

Shortly after, Mirza and Osindero (2014) introduced the Conditional GAN (CGAN), which extends GANs by allowing users to control the output distribution. CGAN accomplishes this by concatenating a conditional vector to the generator's latent vector, enabling users to specify attributes for the generated output.

In subsequent years, Arjovsky et al. (2017) refined the GAN architecture by introducing the Wasserstein GAN (WGAN), which reframes the generator/discriminator setup as an actor/critic model. In WGAN, the critic employs the Wasserstein loss function to score the generator's output, aiming for improved stability and gradient flow. However, Arjovsky et al. found that large neural weights could destabilize training, leading them to implement weight clipping as a control measure. Later, Gulrajani et al. (2017) addressed the limitations of weight clipping by introducing a gradient penalty (WGAN-GP), which imposes a softer constraint on weight sizes and enhances stability by penalizing large weights in the loss function.

We integrate the principles of CGAN and WGAN-GP into a Conditional Wasserstein GAN (CWGAN). Building on this, we develop Conditional HydraGAN, an extension that incorporates DeSmet and Cook's (2024) HydraGAN—a multi-objective GAN with multiple critics/discriminators, each specialized for different aspects of data quality and diversity.

3.5.2 Introduction to Conditional HydraGAN

In the pursuit of equitable healthcare outcomes, particularly within geriatric populations, addressing algorithmic bias is essential. Machine learning models deployed in clinical settings

face the challenge of biases that, if uncorrected, can propagate healthcare disparities. This section presents *Conditional HydraGAN*, a multi-objective generative adversarial network (GAN) developed specifically for mitigating biases in geriatric clinical datasets. Unlike DeSmet and Cook's original HydraGAN model (2024), which functions as a general-purpose data augmenter, the Conditional HydraGAN introduces conditional attributes to generate synthetic data, enforcing an arbitrary output distribution of sensitive features, such as age and gender. By enabling customized data generation, Conditional HydraGAN enhances the representativeness of training data, promoting fairness in health outcome predictions.

We use the CASAS' dataset to illustrate OFI's effectiveness outside of traditional bias datasets. Our conditional HydraGAN serves as a valuable instance where OFI can objectively evaluate bias and its mitigation across subgroups. Utilizing OFI, we analyze the effectiveness of Conditional HydraGAN by measuring bias both pre- and post-synthetic augmentation. This approach ensures that any reduction in bias through Conditional HydraGAN is rigorously and transparently evaluated.

3.5.3 Methodology: Conditional HydraGAN Design and Implementation

The Conditional HydraGAN employs a multi-agent architecture where a generator is balanced by multiple critics, each responsible for different objectives, such as data realism and fairness. To achieve targeted bias mitigation, we introduce conditional vectors representing sensitive attributes into the generator's input space, as done in Briscoe et al. (2022). This design allows the network to learn variations specific to underrepresented subgroups within the geriatric dataset. During training, each critic provides feedback not only on the data's realism but also on the extent to which bias is minimized with respect to specific protected attributes.

For the CASAS dataset, Conditional HydraGAN is guided by three core objectives: individual diversity, activity diversity, and data realism. In the original HydraGAN model, these goals would typically be addressed through three separate critics. However, by incorporating conditional vectors, Conditional HydraGAN inherently fulfills the individual diversity requirement, eliminating the need for a dedicated critic on this aspect.

Activity diversity, while essential, is given a more flexible role by being treated through a loss function rather than as a strict conditional constraint. This design choice allows the generator to prioritize individual diversity while permitting more natural variability in the types of activities represented. By relaxing activity diversity into a loss function, the model gains

the flexibility to balance this objective against the overarching goal of data realism, ensuring that synthetic data better reflect the authentic, often unpredictable patterns of geriatric health behaviors. Treating activity diversity as a loss function minimizes the risk of producing overly rigid activity distributions that may lack representational richness, while still guiding the model toward generating a broad range of activity types. The activity diversity critic minimizes the divergence-based diversity loss. The divergence-based diversity loss measures the MSE between the actual proportion of activities and the desired proportion for each activity label.

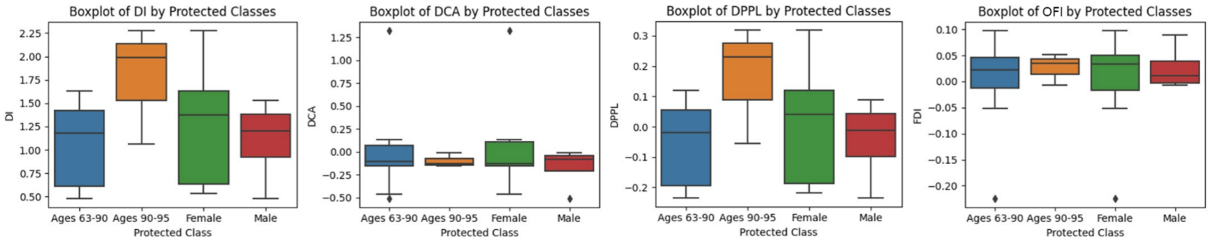
Data realism remains independent of the conditional attributes, as it relies on the generator's ability to produce samples that reflect the broader distribution of the geriatric dataset without being tightly constrained to specific activity patterns. By decoupling data realism from conditional constraints, Conditional HydraGAN focuses on generating data that is both demographically aligned and realistic, capturing the complex and naturally varied patterns found within geriatric populations. This configuration allows Conditional HydraGAN to strike a balance between maintaining essential diversity criteria and achieving realistic, nuanced data that can better support bias mitigation in healthcare applications. The data realism critic minimizes the Wasserstein + Gradient Penalty (WGAN+GP) loss (Gulrajani et al., 2017).

3.5.4 Results: Evaluating Bias Mitigation with OFI

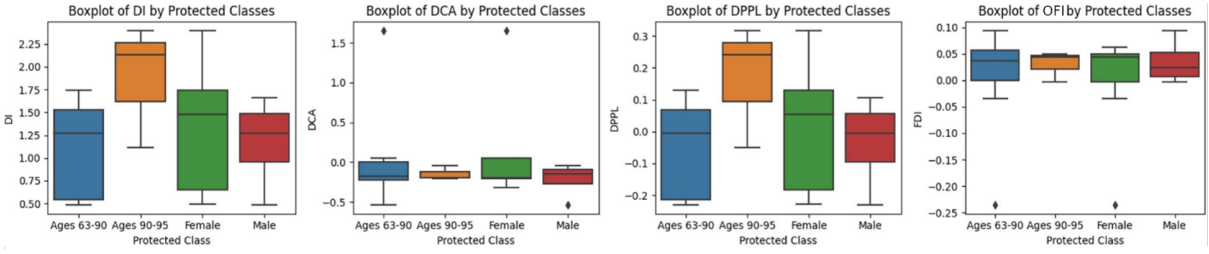
We focus on two sensitive attributes: age and gender. In the case of gender, we assess bias for the traditional male and female groups. In the case of age, we assess bias for the older 25% of the sample in comparison with the group containing the younger 75% of the participants. Rather than evaluate bias for all of the metrics listed in Table 3.2, we select a subset including Disparate Impact (DI), Difference in Conditional Acceptance (DCA), Difference in Proportionate Positives and Labels (DPPI), and Fairness Disparity Index (OFI). This is a representative set: the remaining metrics yield very similar results to these.

The selected bias metrics focus on a task, in this case activity recognition. To simplify analyses, we aggregate activity categories into two classes: *active* behavior (bed-toilet transition, cook, eat, enter home, leave home, hygiene, wash dishes) and *sedentary* behavior (relax, sleep, work, other). For DCA, DPPL, and OFI, a no-bias score is zero. For DI, the no-bias score is one. A closer value to the no-bias score indicates less bias. A positive value indicates a bias toward the sedentary categories, while a negative value indicates a bias toward the active categories.

While Aminikhanghahi and Cook (2019) employ a random forest for CASAS human activity recognition, this work uses predictions from a multilayer perceptron (MLP) classifier. We find that the MLP is better able to infer classifications of the real data using the synthetic data. This is likely because the synthetic data's sensors and activities are processed by a softmax activation function while the real data are one-hot encoded. We do not use argmax to one-hot encode the generated data because some information is lost in the process. Empirically, the MLP better transfers knowledge between the synthetic softmax and real one-hot encoded probability density functions.



(a) Evaluating the original dataset with DI, DCA, DPPL, and OFI.



(b) Evaluating DI, DCA, DPPL, and OFI on the synthetic dataset generated by Conditional HydraGAN.

Figure 3.4: Comparison of algorithmic bias using original and synthetic datasets, illustrating bias metrics before and after synthetic data generation using the Conditional HydraGAN approach.

3.5.5 Algorithmic Bias

We compare the algorithmic bias between the original and expanded datasets in Figure 3.4. We see that Conditional HydraGAN yields improvements in OFI, DI, and DCA. DPPL shows about no difference, however this metric assesses the proportion of positive labels (is sedentary),

Using the Original Dataset

Figure 3.4a depicts boxplots of quantified bias on the original dataset. Since the Ages 90-95 class receives high scores in DI and DPPL, we see they are more likely to be predicted as sedentary. Referencing OFI, which gives context for correct predictions, we see that these positive predictions are largely correct as the Ages 90-95 OFI score is near zero. This makes sense as the data reflects that these older patients tend to be sedentary more often than the younger class. However, since OFI is positive, we see that there is still a slight bias for predicting the older age group as sedentary more often than they should be.

Using the Expanded Dataset

We train the model using synthetic data generated by HydraGAN then quantify the bias by testing on real data. We summarize the bias results in Figure 3.4b. Here, individual diversity is enforced by querying the generator with each individual's label. This individual diversity also improves the protected class's diversities. In total, 3,072,000 synthetic points are created to be realistic and improve diversity for the underrepresented groups.

We see in Table 3.4 that the synthetic data mitigates algorithmic bias overall (the bold cells indicate improvement). For OFI, DPPL, and DCA, a score of zero indicates no bias. For DI, the ideal score is 1. Over many trials, these scores have a negligible standard deviation. Some improvement results seem off due to rounding.

Table 3.4: Comparing synthetic and real data for each protected class using bias metrics.

Metric	Data	Ages 63-90	Ages 90-95	Female	Male
OFI	Synthetic	0.002	0.026	-0.001	0.026
	Real	0.004	0.030	-0.001	0.033
	Improves	**0.002**	**0.005**	-0.000	**0.007**
DPPL	Synthetic	-0.056	0.165	0.012	-0.043
	Real	-0.055	0.170	0.012	-0.036
	Improves	-0.002	**0.005**	**0.000**	-0.007
DI	Synthetic	1.060	1.778	1.280	1.103
	Real	1.109	1.891	1.340	1.174
	Improves	**0.049**	**0.113**	**0.060**	**0.071**
DCA	Synthetic	0.003	-0.100	0.046	-0.172
	Real	0.012	-0.151	0.061	-0.219
	Improves	**0.009**	**0.051**	**0.014**	**0.048**

Table 3.5: Comparison of synthetic data and real data to the target uniform distribution.

	Activities			Individuals		
Metric	**Synthetic**	**Real**	**Improvement**	**Synthetic**	**Real**	**Improvement**
KS statistic	0.863	0.900	**4.11%**	0.000	0.733	**100%**
KL Divergence	1.822	2.302	**20.9%**	0.000	0.252	**100%**
JS Distance	0.652	0.725	**10.1%**	0.000	0.250	**100%**

3.5.6 Data Bias

We see that our infinitely strong conditional drastically improves representation for all individuals in Table 3.5. The softer constraint (done by a critic) improves the diversity of activities between 4% and 20%, depending on the metric considered.

3.5.7 Conclusion

The Conditional HydraGAN demonstrates that targeted synthetic data generation can effectively address specific biases in clinical data, particularly in the context of geriatric health monitoring. By applying OFI, we not only ensure that Conditional HydraGAN's bias mitigation is quantitatively assessed but also highlight the broader applicability of OFI in evaluating the fairness of diverse machine learning interventions. Conditional HydraGAN thus represents a practical tool for advancing bias mitigation strategies, with the OFI providing an objective lens for assessing fairness improvements in real-world clinical applications.

3.6 Conclusions and Future Directions

In this work, we introduce the Objective Fairness Index (OFI) to evaluate bias in a legally grounded and context-aware perspective. By leveraging legal frameworks and precedents, particularly those stemming from objective testing and disparate impact theory, OFI captures the nuanced understanding of bias that encompasses both what occurred and what ought to have occurred.

In Section 3.2, we illustrate that the Objective Fairness Index is legally grounded, even where the disparate impact metric fails. Specifically, OFI finds no bias in objective testing while DI may or may not, and OFI can correctly detect bias where DI does not. The analyses include legal-case-driven situations and popular bias problems, including COMPAS (risk of

recidivism).

The Objective Fairness Index addresses a significant gap in the machine learning and AI fairness literature by offering a metric that not only aligns with legal standards but also provides novel insights into binary classification problems. Our findings highlight the limitations of existing metrics, such as disparate impact, which lack context and may lead to misinterpretation or oversight of critical bias factors. OFI offers a solution by integrating objective testing directly into the evaluation of bias.

For future work, we envision OFI in more complex scenarios, including multi-class classification problems, regression tasks, and recommender systems. Another promising avenue is the exploration of OFI's role in the development of debiasing techniques.

Additionally, corrective procedures should be investigated. As an objective measure, OFI identifies discriminatory selections, whereas reparative measures such as DI are for identifying social disparities. Thus, if the OFI fails, correction at the algorithmic or employer level is appropriate. However, if OFI shows no bias but reparative measures do, this suggests systematic disparity and can rationalize social programs. We recommend that all measures and corrective procedures are routinely scrutinized regarding ethics and law.

There are some inherent limitations to applying concepts in law to binary classification. Examples include the "good faith" of the test-maker (Canada's Meiorin test), proportional responses (UK's Equality Act 2010, suggesting a non-binary decision), and expectations of "reasonable accommodations" for certain situations (Americans with Disabilities Act). We hope to explore solutions for these in more complex scenarios where they are addressable.

This work underscores the importance of interdisciplinary approaches in tackling bias in AI. By grounding OFI in well-established legal precedents and demonstrating its applicability to contemporary machine learning challenges, we contribute a tool for researchers striving for fairness and transparency in AI systems.

Chapter 4

ADDRESSING EVALUATION BIASES IN CLASSIFICATION: SAMPLE-SIZE-INDUCED BIAS

Evaluating machine learning models is crucial not only for determining their technical accuracy but also for assessing their potential societal implications. While the potential for sample size-induced bias in algorithms is well known, we demonstrate that many widely used classification metrics also exhibit this bias. This revelation challenges the efficacy of these metrics in assessing bias with high resolution, especially when comparing groups of disparate sizes, which frequently arise in social applications. We provide analyses of the bias that appears in several commonly applied metrics and propose a correction technique. Additionally, we explore the often-overlooked issue of undefined behaviors in metric calculations, which can lead to ambiguous or misleading evaluations. This work illuminates the previously unrecognized challenge of jaggedness in standard evaluation practices, hoping to advance the community's approach for performing equitable and trustworthy classification methods.

4.1 Introduction

Classification metrics derived from confusion matrices are fundamental tools in evaluating machine learning models, particularly in binary classification tasks. Metrics such as accuracy, precision, recall, and Matthews Correlation Coefficient (MCC) provide critical insights into model performance. However, an often overlooked issue is the impact of sample size on these metrics, especially when comparing performance across different subgroups or datasets of varying sizes.

Small sample sizes introduce significant variability and jaggedness in classification metrics due to the discrete and combinatorial nature of confusion matrices. This variability can lead to misleading interpretations of a model's performance and fairness. For instance, minor changes

in the counts of true positives or false negatives can cause disproportionate shifts in metric values, complicating comparisons across groups and potentially obscuring biases or disparities.

Moreover, certain metrics become undefined under specific conditions, such as divisions by zero. These undefined cases, or "holes", in the metric space further challenge the reliability of performance assessments, especially in subgroups with limited data.

Despite the widespread use of confusion-matrix metrics, there has been limited attention to the systematic biases and inconsistencies induced by sample size variations. Existing literature often assumes large sample sizes or overlooks the combinatorial complexities that small samples introduce, leaving a gap in understanding how to accurately assess model performance.

We address these challenges by providing a comprehensive analysis of sample-size-induced biases in confusion-matrix metrics. This chapter's contributions include:

- **Demonstrating Metric Variability**: We present both theoretical and empirical evidence of the jaggedness and variability in classification metrics caused by small sample sizes. By illustrating how these effects distort performance evaluations, we underscore the need for careful interpretation of metrics in small-sample scenarios.
- **Quantifying Undefined Cases**: We systematically count the number of "holes" (situations where metrics are undefined) in several commonly used classification metrics. This quantification reveals the extent to which certain metrics may be unreliable or misleading under specific sample configurations.
- **Metric Alignment Trial for Checking Homogeneity (MATCH) Test**: We introduce a statistical test that assesses the significance of a subgroup's metric score by comparing it to the distribution of scores from a reference group. This approach allows us to determine whether observed differences are due to genuine performance disparities or merely artifacts of sample size variability.
- **Cross-Prior Smoothing (CPS)**: We propose a smoothing technique that incorporates prior information from other groups to enhance metric reliability. By adjusting confusion matrix counts using cross-group priors, CPS uniformly reduces variability and improves the stability of metric estimates.

This work sheds light on a critical issue in the use of confusion-matrix metrics and offers practical solutions to improve their reliability. By addressing the biases introduced by sample size, we aim to enhance the fairness and accuracy of model evaluations, particularly in applications where subgroup comparisons are essential.

4.2 Related Work

As machine learning models are increasingly used in high-stakes domains, ensuring fairness across subgroups is critical. However, the impact of sample size on classification metrics such as accuracy, precision, and recall remains underexplored. This study examines how these metrics evolve with changing sample sizes, revealing potential biases that may exacerbate disparities between subgroups. Understanding these distributional shifts can help develop more equitable AI systems, particularly in settings where sample sizes differ significantly. Metric definitions are provided in Section 2.1.2.

Confusion matrices are fundamental tools for assessing classification performance, providing the basis for scalar metrics like accuracy and MCC. While these metrics are widely used, they can exhibit biases related to sample size—a phenomenon we refer to as *score distribution shift*. This bias affects common fairness metrics, including disparate impact (Feldman et al., 2015), equalized odds (Hardt et al., 2016), and predictive parity (Das et al., 2021b). By highlighting this issue, we aim to enhance the robustness of metric comparisons across datasets with varying sample sizes.

In social data analysis, accurate measurement of fairness is crucial, particularly in legal contexts where metrics like disparate impact assess potential discrimination. For instance, in employment or housing discrimination cases (Dennis J. Aigner and Wiles, 2024; Gastwirth and Miao, 2009) and recidivism prediction (Dressel and Farid, 2018; Moore et al., 2023), datasets often feature small or imbalanced sample sizes. Recognizing how distributional shifts in metrics affect evaluations is vital for ensuring fair outcomes. The recently introduced Objective Fairness Index (Briscoe and Gebremedhin, 2024) is also susceptible to such shifts.

Recent work by Feng et al. (2024) introduces a method to assess model calibration across subgroups, ensuring that models are reliable for different populations, particularly in high-dimensional settings with limited data.

Statistical methods designed for small sample sizes are relevant to this approach. Techniques like Laplacian smoothing, commonly used in natural language processing to handle unobserved events (Eisenstein, 2019), can be adapted to confusion matrices where certain classes may not be represented. Discussions around the use of non-informative or weakly informative priors in Bayesian estimation for small samples (He et al., 2021; Lemoine, 2019) inform this methodology, particularly in developing methods robust under data limitations.

Sokolova and Lapalme (2009) provides a comprehensive analysis of various performance measures used in classification tasks, focusing on their invariance properties with respect to changes in the confusion matrix. Their work introduces the concept of "measure invariance", which refers to a metric's ability to maintain consistent evaluations despite transformations in the underlying data distribution.

Additionally, Goutte and Gaussier (2005) introduces a probabilistic framework that interprets precision, recall, and the F_1-score by using a symmetric beta distribution and Monte Carlo sampling techniques. This approach moves beyond simple point estimates by estimating the likelihood that one algorithm would outperform another based on these scores. Part of this work extends this idea by generalizing precision and recall to a broader set of metrics, grouped as *Joint Ratio Metrics*.

The issue of metric unreliability in small sample sizes is often overlooked, with practitioners sometimes attributing it to the need for more data. Chicco and Jurman (2020) specifically addresses this issue along with imbalanced data by comparing MCC against the more widely used accuracy and F_1-score. They argue that MCC provides a more reliable assessment of classifier performance in binary tasks, particularly when there is class imbalance, because MCC incorporates all four confusion matrix categories. In contrast, accuracy and F_1 score can give misleadingly high values in imbalanced datasets, failing to capture poor performance in one class.

In another work, Rudner et al. (2024) adjusts model parameters with Group-Aware Priors to improve model robustness under subpopulation shifts. Aghbalou et al. (2024) examines imbalanced classification and derive sharp error bounds under class imbalance. Their focus is on improving classification performance when one class is underrepresented, offering new theoretical insights on error rates when class probabilities approach zero.

This work builds on these foundations by examining the effect of small sample sizes on a broader range of classification metrics. We demonstrate that metrics like accuracy, precision, recall, and even MCC not only exhibit increased variability, but also have fluctuating distributions (jaggedness) as the sample size scales, potentially leading to misleading evaluations. For instance, as Chicco and Jurman present, while MCC is robust in imbalanced scenarios, it can still fluctuate in extreme cases of imbalance or small samples, an issue the Cross-Prior Smoothing (CPS) technique aims to mitigate by using prior information from reference groups. Additionally, the introduction of the MATCH Test allows for a rigorous assessment of whether

observed metric scores in small subgroups are meaningful or merely artifacts of sampling variability.

By addressing these gaps, we contribute to the growing literature on improving the robustness and fairness of classification metrics. This approach offers both theoretical insights and practical tools to ensure more reliable evaluations across diverse subgroups, particularly when sample sizes are limited or imbalanced.

4.3 Discrete Distribution Shifts

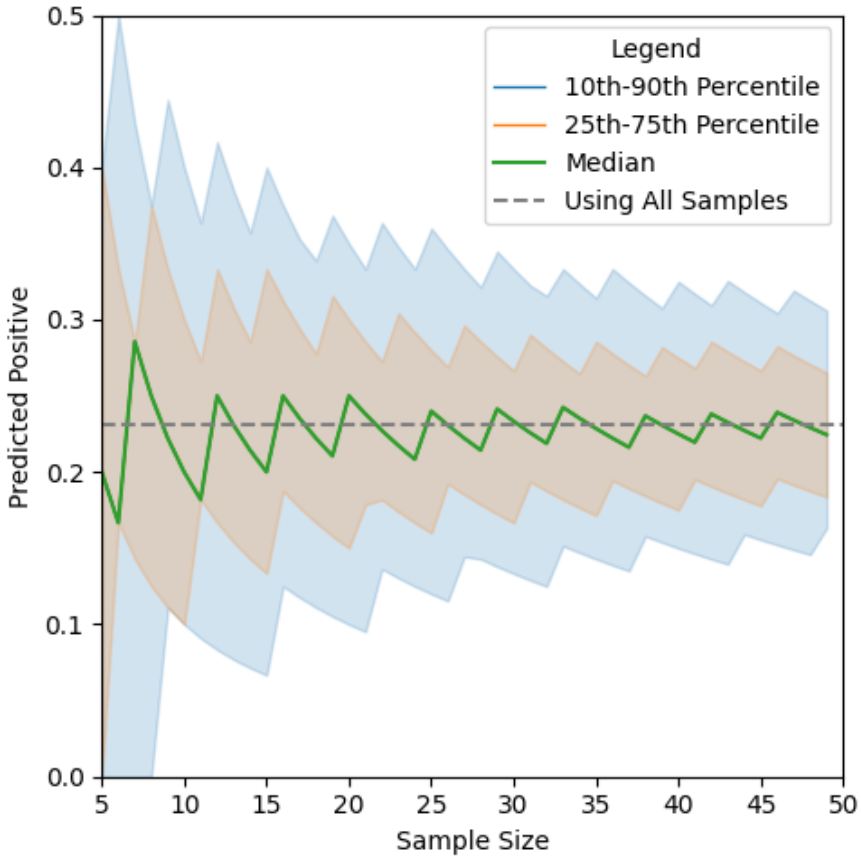


Figure 4.1: Variability of Positive Predictive Rate for Wealth Classification Among Multiracial Individuals: A Monte Carlo Simulation Study Based on Sample Size.

The variability and jaggedness in the predicted positive rates, as illustrated in Figure 4.1, is a direct consequence of both the probabilistic nature of small sample sizes and the combinatorial explosion of the confusion matrix space. This section explains the underlying phenomena contributing to these effects through a probabilistic lens.

To summarize Definition 2.1, the binary confusion matrix, CM, is a 2×2 matrix of the non-negative counts of true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN). Each of these counts is referred to as a *cell* c. The total number of samples n is partitioned among these four categories, such that $n = \text{TP} + \text{FN} + \text{FP} + \text{TN}$. For a given sample size n, we denote the set of all possible configurations (where a configuration is a unique confusion matrix) in Theorem 4.1. Next, we show how to calculate the number of times a given count x occurs for any cell in Theorem 4.2.

Theorem 4.1. *The cardinality of $\mathcal{M}(n)$ (the space of all possible confusion matrices given size n) is $\mathcal{N}(n) = \binom{n+3}{3} = (n+1)(n+2)(n+3)/6$.*

Proof. We prove this using the stars and bars method, which is defined as:

$$\binom{n^* + k - 1}{k - 1}, \tag{4.1}$$

where n^* represents the number of samples, and k denotes the number of buckets.

For our case, we have $n^* = n$ samples and $k = 4$ buckets, which corresponds to $k - 1 = 3$ bars. Applying the formula, we obtain:

$$|\mathcal{M}(n)| = \mathcal{N}(n) = \binom{n + 4 - 1}{4 - 1} = \frac{(n+1)(n+2)(n+3)}{6}. \tag{4.2}$$

□

Corollary 4.1.1. *From Equation 4.2, the space complexity of $\mathcal{M}(n)$ is $O(n^3)$, and more precisely, $\Theta(n^3)$.*

Next, we move to count the number of times $c = x \big| n$. We begin by visualizing the sets $\mathcal{M}(1)$ in Equation 4.3 and $\mathcal{M}(2)$ in Equation 4.4.

$$\mathcal{M}(1) = \{[1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]\} \tag{4.3}$$

$$\mathcal{M}(2) = \begin{cases} [2,0,0,0], [0,2,0,0], [0,0,2,0], [0,0,0,2], \\ [1,0,0,1], [1,0,1,0], [1,1,0,0], [0,1,0,1], \\ [0,1,1,0], [0,0,1,1]. \end{cases} \tag{4.4}$$

As we progress from $\mathcal{M}(1)$ to $\mathcal{M}(2)$, certain counting patterns start to emerge. Let $C(x; n)$ represent the frequency of the value x appearing in any given cell across n samples. For instance,

in $\mathcal{M}(1)$, we observe that $C(1;1) = 1$ and $C(0;1) = 3$. Continuing this examination reveals a connection to triangular numbers, allowing us to generalize the pattern. Notably, 1 and 3 correspond to the first two triangular numbers. We adopt Knuth's notation for the w^{th} triangular number, as shown in Equation 4.5 (Knuth, 1968).

$$w? = \sum_{j=1}^{w} j \tag{4.5}$$

We see that $\mathcal{M}(2)$ also follows this pattern:

$$(C(2;2), C(1;2), C(0;2)) = (1?, 2?, 3?) = (1, 3, 6)\,. \tag{4.6}$$

Theorem 4.2. *Given the number of samples, n, the count of value $x \in [0, n]$, denoted as $C(x;n)$, appearing in $\mathcal{M}(n)$ is found by:*

$$C(x;n) = 0.5(n-x+1)(n-x+2) = (n-x+1)? \tag{4.7}$$

Proof. First, we prove that C follows the triangular sequence. Since all cells are interdependent and required to sum to n, we use the stars and bars method for n^* stars and $k-1$ bars:

$$\binom{n^* + k - 1}{k-1}. \tag{4.8}$$

If cell c has value x, then the other three cells must comprise of three non-negative integers summing to $n - x$. In the stars and bars terminology, there are $n - x$ stars and $k = 3$ other buckets/cells to choose from.

$$C(x;n) = \binom{n-x+3-1}{3-1} = \binom{n-x+2}{2} \tag{4.9}$$

Rewriting $C(x;n)$ from the binomial coefficient, we create Equation 4.10.

$$C(x;n) = \frac{(n-x+2)!}{2!(n-x)!} = 0.5(n-x+1)(n-x+2) \tag{4.10}$$

Now we prove that $C(x;n)$ is a triangular number:

$$0.5(n-x+1)(n-x+2) = (n-x+1)?. \tag{4.11}$$

It is well established that the w^{th} triangular number is found by

$$w? = 0.5w(w + 1). \tag{4.12}$$

So by substituting $w = n - x + 1$, we see that Equation 4.11 follows the formula for $w?$. □

Thus, the narrowing of the distribution observed in Figure 4.1 can be attributed to the combinatorial expansion of $\mathcal{M}(n)$ with increasing n. For small sample sizes, the number of possible confusion matrices is limited, causing each configuration to represent a significant fraction of the total space, including those yielding extreme metric values. As n grows, the number of possible confusion matrices increases rapidly, and $\mathcal{M}(n)$ becomes densely populated.

The jaggedness in metric distributions is further related to the counts of configurations corresponding to specific cell values in the confusion matrix. These counts are associated with triangular numbers, as proven in Theorem 4.2.

Additionally, these configurations are influenced by the underlying probability distribution of classification outcomes, denoted as $(p_{\text{TP}}, p_{\text{TP}}, p_{\text{FP}}, p_{\text{TN}})$. Using k_i as the integer count of i, we use the multinomial distribution definition to define the probability of any CM configuration:

$$P(\text{CM}) = \frac{n!}{k_{\text{TP}}!\, k_{\text{FP}}!\, k_{\text{TP}}!\, k_{\text{TN}}!}\, p_{\text{TP}}^{k_{\text{TP}}}\, p_{\text{FP}}^{k_{\text{FP}}}\, p_{\text{TP}}^{k_{\text{TP}}}\, p_{\text{TN}}^{k_{\text{TN}}}. \tag{4.13}$$

Due to the factorial terms in the multinomial coefficient from Equation 4.13, even a single-unit increase in n can significantly disrupt the probability mass distribution for small sample sizes, as depicted in Figure 4.1.

The score distribution shifts are not solely a probabilistic artifact. As n increases, the discrete space of possible confusion matrices becomes more finely granulated, and the number of configurations corresponding to certain metric scores increases at different rates. Figure 4.2 shows how the distribution of metric scores changes with varying n, as all possible confusion matrix configurations are considered for each group size.

As the sample size increases, the probability masses of metric scores tend to converge around their expected values. In Figure 4.1, the expected value of the predicted positive rate converges to $1365/5956 \approx 0.23$. For any metric M, the expected value is $\mathbb{E}[M(p_{\text{TP}}, p_{\text{TP}}, p_{\text{FP}}, p_{\text{TN}})]$.

For example, Figure 4.2 uses all configurations in $\mathcal{M}(n)$, so all probabilities are 0.25. We see that accuracy and predicted positive converge to $(0.25 + 0.25)/1 = 0.5$; MCC (with or

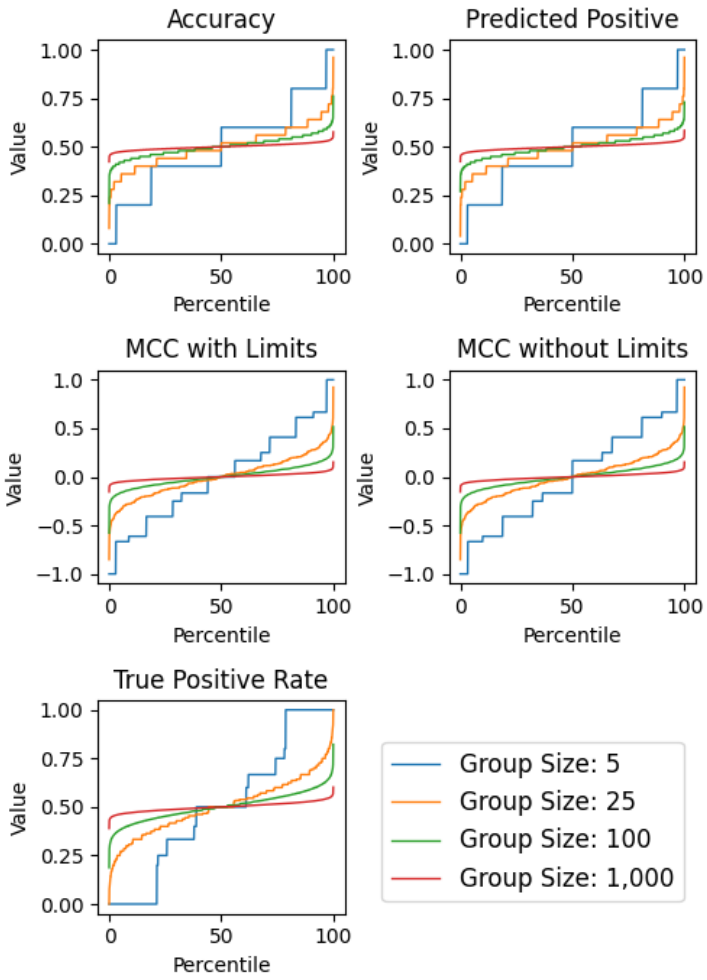


Figure 4.2: Cumulative distribution functions of common classification metrics as sample size increases.

without limits) converges to $\frac{0.25 \cdot 0.25 - 0.25 \cdot 0.25}{\sqrt{(0.25+0.25)\cdot(0.25+0.25)\cdot(0.25+0.25)\cdot(0.25+0.25)}} = 0$; and TPR goes to $0.25/(0.25 + 0.25) = 0.5$.

By understanding discrete distribution shifts, we highlight the importance of considering sample size effects when interpreting classification metrics, particularly where small subgroups are compared. Ignoring these effects can lead to misleading conclusions about model performance and fairness across different populations.

4.4 Edge Cases

Classification metrics derived from confusion matrices can exhibit undefined behaviors, or "holes", under certain conditions—particularly when denominators in their formulas become

zero. Understanding and quantifying these edge cases is crucial for metric selection, especially with small sample sizes.

We generalize metrics of the form $M = \frac{c_i}{c_i + c_j}$, as *Joint Ratio Metrics* (JRMs), which include common measures like true positive rate (TPR), false positive rate (FPR), precision (PPV), and others.

Theorem 4.3. *For a population size $n \geq 1$ and group sizes n_i such that $\sum \forall n_i = n$, the count of holes for the following metrics are:*

Metric	***Hole Count***	***Asymptotic Limits***
Binomial Metrics	0	0
F_1 Score	1	$\Theta(1)$
Joint Ratio Metrics (JRM)	$n + 1$	$\Theta(n)$
MCC	$4n$	$\Theta(n)$
Prevalence Threshold (PT)	$\geq 2n + 2$ & $< e^{\gamma} n \log\log n + \frac{0.6483n}{\log\log n}$	$\Omega(n)$ & $O(n \log\log n)$
Treatment Equality	$\binom{n_1+2}{2} + \binom{n_2+2}{2} - 1$	$\Theta(n^2)$

Note on Prevalence Threshold: *$e^{\gamma} \approx 1.781$ and this upper bound is for $n \geq 3$.*

Proof. Let c_i be some cell in a confusion matrix.

Binomial Metrics: With form $(c_i + c_j)/n$, these metrics are defined for all $n \geq 1$.

F_1 Score: Introduced as the harmonic mean of precision and recall, modern ML benchmarks often use the simplified version: $2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$. Here, F_1 is only undefined when $2\text{TP} + \text{FP} + \text{FN} = 0$, giving one hole for $\text{TN} = n$.

Joint Ratio Metrics (JRMs): JRMs are only undefined when $c_i + c_j = 0$ which occurs when $c_k + c_l = n$. There are $n + 1$ ways to partition n into c_k and c_l, resulting in $n + 1$ holes.

Matthews Correlation Coefficient (MCC): Defined as $\frac{\text{TP}\times\text{TN} - \text{FP}\times\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$, MCC is only undefined when any term in the denominator's product is zero (e.g., $\text{TP} + \text{FN} = 0$). From the JRM proof, there are $n + 1$ configurations for $c_i + c_j = 0$. However, there are four overlapping cases, all when some $c_i = n$. Hence, there are $4(n + 1) - 4 = 4n$ holes.

Prevalence Threshold (PT): Defined as $\frac{\sqrt{\text{TPR}\cdot\text{FPR}} - \text{FPR}}{\text{TPR} - \text{FPR}}$, PT is undefined if $\text{TP} + \text{FN} = 0$ (TPR), or $\text{FP} + \text{TN} = 0$ (FPR), or $\text{TPR} - \text{FPR} = 0$. TPR and FPR are JRM metrics, hence have

$n + 1$ holes each. We now move to $\text{TPR} - \text{FPR} = \text{TP}/(\text{TP} + \text{FN}) - \text{FP}/(\text{FP} + \text{TN}) = 0$.

$$\text{TP}/(\text{TP} + \text{FN}) - \text{FP}/(\text{FP} + \text{TN}) = 0 \tag{4.14}$$

$$\iff \quad \text{TP}(\text{FP} + \text{TN}) - \text{FP}(\text{TP} + \text{FN}) = 0 \tag{4.15}$$

$$\iff \quad \text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN} = 0 \tag{4.16}$$

Equation 4.14 only occurs if one row or column is a multiple of the other. For a matrix with a zero row or column, this is the set of MCC holes described above. Otherwise, there is at most one multiple of that column such that the sum of the entries is n. For example, if the predicted-negative column is an integer (k) multiple of the predicted-positive column, then we have $k \cdot (\text{TP} + \text{FN}) = n$ and hence $(\text{TP} + \text{FN})|n$. Thus, we can bound the number of holes using the sum of divisor function. Then, by Robin's inequality (Robin, 1984), there are less than $e^{\gamma} n \log\log n + \frac{0.6483n}{\log\log n}$ holes for $n \geq 3$ where $\gamma \approx 0.577$ is the Euler-Mascheroni constant and $e = 2.718\ldots$ is Euler's Number. Hence, $e^{\gamma} \approx 1.781$.

Furthermore, since the total holes from FPR and TPR equals $2n + 2$, the lower bound is $\Omega(n)$.

Treatment Equality (TE): TE is defined as $\text{FN}_1/\text{FP}_1 - \text{FN}_2/\text{FP}_2$ for two subgroups partitioned by n: $\text{CM}_1 + \text{CM}_2 = \text{CM}$. TE is only undefined if either $\text{FP}_1 = 0$ or $\text{FP}_2 = 0$. Using the stars and bars, the number of 3-tuples that sum to n is given by $\binom{n_i+3-1}{3-1} = \binom{n_i+2}{2}$ for $i \in \{1, 2\}$. Removing the double counting of the case where $\text{FP}_1 = \text{FP}_2 = 0$, the number of holes is $\binom{n_1+2}{2} + \binom{n_2+2}{2} - 1$. The asymptotic complexity is $\Theta(n_1^2 + n_2^2)$. Since $n_1 + n_2 = n$, the complexity is $\Theta(n^2)$. □

Corollary 4.3.1. *Comparing these holes with the cardinality of possible configurations, $\Theta(n^3)$ from Corollary 4.1.1, we see that the frequencies of holes diminish toward zero as n increases.*

By analyzing the limiting behavior of metrics when they are undefined, we begin to understand their implications. Next, we briefly discuss the limits of JRMs and MCC.

JRMs are indeterminate when $c_i + c_j = 0$. This occurs because as $c_i + c_j \to 0$, the ratio $M = \frac{c_i}{c_i + c_j}$ approaches an indeterminate form $0/0$. Depending on how c_i and c_j approach zero, M can take any value in $(-\infty, \infty)$. We formally prove this in Theorem 4.4.

Theorem 4.4. *Joint Ratio Metrics (JRMs) are indeterminate when undefined.*

Proof. Defined by $c_i/(c_i + c_j)$, a JRM is undefined when $c_i + c_j = 0$. We manipulate this

equation with the ratio: $c_i = rc_j$. Now, $JRM \triangleq rc_j/(rc_j + c_j) = r/(r + 1)$.

We consider $r \to -1$. As $r = -1^-$, $c_i \to -1^+c_j$, JRM $\to -\infty$. Conversely, if $r = -1^+$, $c_i \to -1^-c_j$, JRM $\to \infty$. As such, a JRM is indeterminate for $c_i + c_j = 0$.

However, one may argue that it is impossible to have a negative ratio of counts. We consider the problem of defining JRM = 0/0 using FPR = FP/(FP + TN). On one hand, FP equals the count of negatives N = FP + TN, suggesting $N/N = 0/0 = 1$. However, the algorithm also made no incorrect predictions in the negative set, suggesting $0/N = 0/0 = 0$. Finally, we consider the naïve prior: $\text{FPR}_\epsilon = \frac{\text{FP}+1}{N+2}$. Here, $c_i + c_j = 0 \implies 0.5$. With similar interpretations yielding different results, we can conclude that a JRM is indeterminate for $c_i + c_j = 0$. □

For $c_i + c_j = 0$, Chicco and Jurman (2020) finds that MCC tends to zero. However, they neglect the four cases of $c = n$. We find that all four cases are indeterminate since perturbations (relocating one instance to another cell) either remains a hole (tending to zero) if a single entry in the same row or column is increased, or a valid value of ±1 if the increased cell does not share a row or column with c.

4.5 MATCH Test

The inherent variability and jaggedness in classification metrics due to small sample sizes complicate the assessment of model performance across groups of different sizes. Small variations in the sample size n can lead to disproportionately large changes in metric values, making direct comparisons potentially misleading. This issue is particularly acute in fairness evaluations, where comparing metrics across subgroups (e.g., different demographic groups) is essential.

To address this challenge, we propose a *Metric Alignment Trial for Checking Homogeneity (MATCH) Test* that quantifies the likelihood of observing a given metric score under a reference distribution. Specifically, we assess whether an observed metric score, $M(\text{CM}_i)$, computed from a confusion matrix CM_i for subgroup i, is consistent with what would be expected if this subgroup followed the same performance distribution as a reference group. By comparing the observed score from subgroup i against the cumulative distribution function (CDF) derived from the reference group, We determine the percentile rank of the observed score within this distribution. This method allows us to evaluate the statistical significance of observed metrics, facilitating more informed interpretations when sample size variations could skew

direct comparisons.

We focus on several classification metrics, including accuracy, prevalence, and predicted positive rate (collectively referred to as *Binomial Metrics*), as well as the *marginal benefit* $\mathcal{B}$, and JRMs. For each metric, we derive methods to compute the cumulative probability of an observed score S_{obs} under the assumption that it is drawn from the reference distribution.

To prepare, we now define several terms. Let $k_i \geq 0$ denote the integer count of c_i, p_i is the probability of c_i, $q_i = 1 - p_i$ is the probability of not c_i, S_{obs} is the observed score of a metric M, and $P(S \leq S_{\text{obs}}) \in [0, 1]$ is the probability that $S \leq S_{\text{obs}}$. The subscript $i + j$ means i or j.

4.5.1 Binomial Metrics

We find distributions for all Binomial Metrics that have the form $(c_i + c_j)/n$. This includes accuracy, prevalence, and predicted positive rate, inaccuracy, negative prevalence, and predicted negative rate. We define and describe these metrics in Table 2.2.

Theorem 4.5. *For Binomial Metrics, the cumulative probability is given by the binomial cumulative distribution function (CDF):*

$$P(S \leq S_{\text{obs}}) = \sum_{f=0}^{k_{i+j}} \binom{n}{f} p^f q^{n-f} \tag{4.17}$$

Proof. The total count k_{i+j} corresponds to the number of successes in n trials, each with success probability p. Thus, k_{i+j} follows a binomial distribution with parameters n and p. □

When $np \geq 5$ and $nq \geq 5$, we can approximate $P(S \leq S_{\text{obs}})$ using the normal distribution due to the Central Limit Theorem (Ye et al., 2024; Freund et al., 2014). This is done in constant time. By definition, the mean $\mu = np_{i+j}$ and standard deviation $\sigma = \sqrt{np_{i+j}q_{i+j}}$.

$$\text{Standardize: } z = \frac{k_{i+j} + 0.5 - \mu}{\sigma} \tag{4.18}$$

The +0.5 is the continuity correction as we approximate this discrete distribution with a continuous one. Then, we use the standard normal CDF with z: $\Phi(z)$.

For smaller n, exact computation using the binomial CDF is feasible in $O(n^2)$ time. Otherwise, if the approximation criteria does not hold, one can use efficient approximations such as the Lanczos approximation for $O(n \log n)$ time (Lanczos, 1964).

Example. Consider $n = 100$, $p = 0.75$, and observed score $S_{\text{obs}} = 0.80$. Then $k_{i+j} = nS_{\text{obs}} = 80$, $\mu = 75$, $\sigma \approx 4.33$, and $z \approx 1.27$. Thus, $P(S \leq 0.80) \approx \Phi(1.27) \approx 0.90$, so the observed score is higher than approximately 90% of the expected outcomes under the reference distribution.

4.5.2 Marginal Benefit

The marginal benefit is defined as $\mathcal{B} = (\text{FP} - \text{FN})/n$, and is designed to measure the net benefit or cost to a group (Briscoe and Gebremedhin, 2024). Unlike Binomial Metrics, $\mathcal{B}$ involves the difference of counts, making its distribution symmetric around zero when $p_{\text{FP}} = p_{\text{FN}}$. In this section, we derive the probability mass function (PMF) and CDF of $\mathcal{B}$. Understanding this distribution allows us to compute the exact probability of observing a given value of $\mathcal{B}$ under the reference distribution.

Theorem 4.6. *We find the PMF of $\mathcal{B} = \frac{k}{n}$ by summing over the probabilities of all counts FP and FN such that* $\text{FP} - \text{FN} = k$.

$$P\left(\mathcal{B} = \frac{k}{n}\right) = \sum_{n_{-1}=n_{-1}^{\min}}^{n_{-1}^{\max}} \frac{n!}{(k+n_{-1})!\, n_{-1}!\, (n-k-2n_{-1})!} \times p_{\text{FP}}^{k+n_{-1}}\, p_{\text{TP}}^{n_{-1}}\, p_0^{n-k-2n_{-1}} \tag{4.19}$$

where:

$$n_{-1}^{\min} = \max(0, -k), \quad n_{-1}^{\max} = \left\lfloor \frac{n-k}{2} \right\rfloor$$

and k ranges over the integers $-n$ and n since $k = \text{FP} - \text{FN} \in [-n, n]$.

Proof. $\mathcal{B}$ can be expressed as the average of n i.i.d. random variables X_i: $\frac{1}{n}\sum_{i=1}^{n} X_i$. For

$$X_i = \begin{cases} +1 & \text{with probability } p_+ = p_{\text{FP}} \\ -1 & \text{with probability } p_- = p_{\text{FN}} \\ 0 & \text{with probability } p_0 = p_{\text{TN+TP}}. \end{cases} \tag{4.20}$$

We proceed with the proof in two parts. First, we determine the bounds of the PMF, and then we derive the corresponding multinomial probability.

The Bounds of the PMF: Let n_{+1} be the number of times $X_i = +1$ (number of false positives), n_{-1} be the number of times $X_i = -1$ (number of false negatives), and n_0 be the number of times

$X_i = 0$ (number of true positives and true negatives). Then:

$$n = n_{+1} + n_{-1} + n_0$$
$$S = k = n_{+1} \cdot (+1) + n_{-1} \cdot (-1) + n_0 \cdot 0 = n_{+1} - n_{-1}$$

It follows that $n_{+1} = k + n_{-1}$. We use this to redefine n_0 in terms of n, k, and n_{-1}:

$$n_0 = n - (k + n_{-1}) - n_{-1} = n - k - 2n_{-1}.$$

Our bounds are constrained by all cells being non-negative. As such, the counts n_{+1}, n_{-1}, and n_0 must satisfy:

$$n_{+1} \geq 0 \implies k + n_{-1} \geq 0 \implies n_{-1} \geq -k \tag{4.21}$$
$$n_{-1} \geq 0 \tag{4.22}$$
$$n_0 \geq 0 \implies n - k - 2n_{-1} \geq 0 \implies n_{-1} \leq \frac{n-k}{2} \tag{4.23}$$

Hence, the valid range for n_{-1} is:

$$n_{-1}^{\min} = \max(0, -k), \quad n_{-1}^{\max} = \left\lfloor \frac{n-k}{2} \right\rfloor \tag{4.24}$$

The Multinomial Probability: By definition, the multinomial probability of observing counts (n_{+1}, n_{-1}, n_0) is:

$$P(n_{+1}, n_{-1}, n_0) = \frac{n!}{n_{+1}!\, n_{-1}!\, n_0!} \times p_{\text{FP}}^{n_{+1}}\, p_{\text{TP}}^{n_{-1}}\, p_0^{n_0} \tag{4.25}$$

Substituting $n_{+1} = k + n_{-1}$ and $n_0 = n - k - 2n_{-1}$, we have:

$$P(S = k) = \sum_{n_{-1}=n_{-1}^{\min}}^{n_{-1}^{\max}} \frac{n!}{(k+n_{-1})!\, n_{-1}!\, (n-k-2n_{-1})!} \times p_{\text{FP}}^{k+n_{-1}}\, p_{\text{TP}}^{n_{-1}}\, p_0^{n-k-2n_{-1}} \tag{4.26}$$

Thus, the PMF of $\mathcal{B} = \dfrac{k}{n}$ is given by $P\left(\mathcal{B} = \dfrac{k}{n}\right) = P(S = k)$. □

Using the PMF, we can compute the cumulative probability of $\mathcal{B}$ for a given b using the following theorem.

Theorem 4.7. *The cumulative distribution function of $\mathcal{B}$ is:*

$$\begin{aligned} P(\mathcal{B} \leq b) &= P(S \leq nb) \\ &= \sum_{k=-n}^{\lfloor nb \rfloor} \sum_{n_{-1}=n_{-1}^{\min}(k)}^{n_{-1}^{\max}(k)} \frac{n!}{(k+n_{-1})!\, n_{-1}!\, (n-k-2n_{-1})!} \times p_{\text{FP}}^{k+n_{-1}}\, p_{\text{TP}}^{n_{-1}}\, p_0^{n-k-2n_{-1}} \end{aligned} \tag{4.27}$$

where b is a real number ranging from -1 to 1 and k is an integer ranging from $-n$ to $\lfloor nb \rfloor$.

Proof. We start by noting that $\mathcal{B} = \dfrac{S}{n}$, where $S = \text{FP} - \text{FN} = \sum_{i=1}^{n} X_i$ is the sum of independent random variables X_i defined in Equation 4.20.

Next, we reframe $P(\mathcal{B} \leq b)$ as

$$P(\mathcal{B} \leq b) = P\left(\frac{S}{n} \leq b\right) = P(S \leq nb)\,. \tag{4.28}$$

Since $S = \text{FP} - \text{FN}$ is an integer ranging from $-n$ to n, the possible values of $\mathcal{B}$ are $\dfrac{k}{n}$ for integer $k \in [-n, n]$. Thus, the CDF is:

$$P(\mathcal{B} \leq b) = \sum_{k=-n}^{\lfloor nb \rfloor} P(S = k)$$

□

Computing the CDF from Theorem 4.7 takes $O(n^2)$ time. However, for large n, we can approximate the distribution of $\mathcal{B}$ in constant time using the normal distribution (Theorem 4.8).

Theorem 4.8. *The cumulative probability for $\mathcal{B}$ can be approximated using the normal distribution:*

$$F(S \leq S_{\text{obs}}) \approx \Phi\left(\frac{k_{\text{FP}} - k_{\text{FN}} - (p_+ - p_-)}{\sqrt{(p_+ + p_-) - (p_+ - p_-)^2}}\right) \tag{4.29}$$

Proof. Recall that $\mathcal{B}$ can be expressed as the average of n i.i.d. random variables X_i: $\frac{1}{n}\sum_{i=1}^{n} X_i$. For

$$X_i = \begin{cases} +1 & \text{with probability } p_+ = p_{\text{FP}} \\ -1 & \text{with probability } p_- = p_{\text{FN}} \\ 0 & \text{with probability } p_0 = p_{\text{TN+TP}}. \end{cases}$$

Now we calculate the mean and standard deviation of $\mathcal{B}$.

$$\mu = \mathbb{E}[X_i] = p_+ - p_- + 0 \cdot p_0 = p_+ - p_- \tag{4.30}$$

The variance is defined by

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left[(X_i - \mathbb{E}[X_i])^2\right] \\
&= \mathbb{E}\left[X_i^2 - 2X_i\mathbb{E}[X_i] + \mathbb{E}[X_i]^2\right] \\
&= \mathbb{E}[X_i^2] - 2\mathbb{E}[X_i]\mathbb{E}[X_i] + \mathbb{E}[X_i]^2 \\
&= \mathbb{E}[X_i^2] - 2\mathbb{E}[X_i]^2 + \mathbb{E}[X_i]^2 \\
&= \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 \\
&= \mathbb{E}[X_i^2] - \mu^2 \quad \text{(the expanded definition of variance).}
\end{aligned} \tag{4.31}$$

Applying this to $\mathcal{B}$, we find that:

$$\mathbb{E}[X_i^2] = (+1)^2 p_+ + (-1)^2 p_- + (0)^2 p_0. \tag{4.32}$$

Thus,

$$\sigma^2 = (p_+ + p_-) - (p_+ - p_-)^2 \tag{4.33}$$

$$\implies \sigma = \sqrt{p_+ + p_- - (p_+ - p_-)^2}. \tag{4.34}$$

Then we calculate the z-score and use the normal CDF Φ. Note that the continuity correction δ cancels out:

$$X_i = k_{\text{FP}} - k_{\text{FN}} \tag{4.35}$$

$$\implies z = \frac{(k_{\text{FP}} + \delta) - (k_{\text{FN}} + \delta) - \mu}{\sqrt{p_+ + p_- - (p_+ - p_-)^2}}. \tag{4.36}$$

□

4.5.3 Joint Ratio Metrics

Joint Ratio Metrics (JRMs) are expressed as $c_i/(c_i + c_j)$. Computing the cumulative probability for JRMs presents additional challenges due to the ratio of random variables involved. Theorem 4.9 extends the work of Goutte and Gaussier (2005) by introducing a generalized framework

for these metrics, along with an analysis of computational complexity. For simplicity in the following discussion, we define $k_{i+j} = k$ and $p_{i+j} = p$.

Theorem 4.9. *The cumulative probability for a JRM score S_{obs} is given by:*

$$\begin{aligned} P(S \leq S_{\text{obs}}) &= \sum_{k=1}^{n} P(c_{i+j} = k) P\left(\frac{k_i}{k} \leq S_{\text{obs}} \middle| c_{i+j} = k\right) \\ &= \sum_{k=1}^{n} \left(\binom{n}{k} p^k (1-p)^{n-k} \cdot \sum_{k_i=0}^{k_i^{\max}} \binom{k}{k_i} \theta^{k_i} (1-\theta)^{k-k_i} \right) \end{aligned} \tag{4.37}$$

Proof. From our analyses in binomial metrics, we have c_{i+j} and c_i following binomial distributions:

$$c_{i+j} \sim Binomial(n, p_{i+j}), \quad c_i \sim Binomial(n, p_i).$$

Since S is only defined for $k_i + k_j = k > 0$, we focus on $k \geq 1$. We begin with the simplest solution in Equation 4.38 which sums all probabilities for scores lower than S_{obs}.

$$P(S \leq S_{\text{obs}}) = \sum_{k=1}^{n} \sum_{k_i=0}^{n} \begin{cases} P(k_i \big| c_{i+j} = k) & \text{if } \frac{k_i}{k} \leq S_{\text{obs}} \\ 0 & \text{otherwise.} \end{cases} \tag{4.38}$$

By definition, if c_i is sampled, then c_{i+j} is also sampled, implying that $P(c_{i+j} \mid c_i) = 1$. Using Bayes' rule, we express $P(c_i \mid c_{i+j})$ as:

$$\theta = P(c_i \mid c_{i+j}) = \frac{P(c_{i+j} \mid c_i) \cdot P(c_i)}{P(c_{i+j})} = \frac{p_i}{p_{i+j}}. \tag{4.39}$$

Consequently, we model the conditional sampling from a JRM M as:

$$c_i \mid c_{i+j} \sim \text{Binomial}(c_{i+j}, \theta).$$

We now proceed to derive the PMFs to construct the CDFs. Since the events are binomially distributed, the PMFs are given by:

$$P(c_{i+j}) = \binom{n}{c_{i+j}} p_{i+j}^{c_{i+j}} (1 - p_{i+j})^{n - c_{i+j}} \tag{4.40}$$

$$\text{and} \quad P(c_i | c_{i+j}) = \binom{c_{i+j}}{c_i} \theta^{c_i} (1-\theta)^{c_{i+j} - c_i}. \tag{4.41}$$

Then, the CDF of Binomial(c_{i+j}, θ) can be expressed as:

$$\begin{aligned} P\left(\frac{k_i}{k} \leq S_{obs} \middle| c_{i+j} = k\right) &= P\left(k_i \leq S_{obs} \cdot k \middle| c_{i+j} = k\right) \\ &= \sum_{k_i=0}^{k_i^{\max}} \binom{k}{k_i} \theta^{k_i} (1-\theta)^{k-k_i} \end{aligned} \tag{4.42}$$

where $k_i^{\max} = \lfloor S_{obs} \cdot c_{i+j} \rfloor$.

By expanding the PMF of c_{i+j} and combining it with the CDF of $c_i \mid c_{i+j}$, we obtain a calculable CDF of any JRM:

$$\begin{aligned} P(S \leq S_{\text{obs}}) &= \sum_{k=1}^{n} P(c_{i+j} = k) P\left(\frac{k_i}{k} \leq S_{\text{obs}} \middle| c_{i+j} = k\right) \\ &= \sum_{k=1}^{n} \left(\binom{n}{k} p^k (1-p)^{n-k} \cdot \sum_{k_i=0}^{k_i^{\max}} \binom{k}{k_i} \theta^{k_i} (1-\theta)^{k-k_i} \right) \end{aligned} \tag{4.43}$$

□

Computing this sum directly has computational complexity $O(n^2)$. However, we can improve efficiency by approximating the distribution of the JRM using the beta distribution. When applying a Bayesian approach with a uniform prior (equivalent to adding pseudo-counts $\lambda = 1$), the posterior distribution of the JRM is a beta distribution with parameters $k_i + \lambda$ and $k_j + \lambda$.

Using the beta distribution, the cumulative probability is approximated as:

$$P(S \leq S_{\text{obs}}) \approx I_{S_{\text{obs}}}(k_i + \lambda, k - k_i + \lambda) \tag{4.44}$$

where I is the regularized incomplete beta function. This improves the time complexity to $O(n)$ since I integrates once over an independent variable. Beta tables allow constant time if they are available. Otherwise, efficient continued fraction and recursive approximations, or quadrature approximations of integrals may allow further improvements as the integrand is well-behaved.

4.5.4 More on the Normal Approximation

The normal approximation is a powerful tool for approximating the distribution of Binomial Metrics and $\mathcal{B}$. However, the approximation's precision is limited by the sample size n. In this section, we showcase the normal approximation's validity and precision as n scales, and provide

ways to improve the approximation's precision.

Justifying the Normal Approximation

To justify the use of the normal approximation in our MATCH Test for binomial metrics and the marginal benefit metric $\mathcal{B}$, we perform empirical validations. Figure 4.3 illustrates the probability density functions (PDFs) of these metrics as the sample size n increases, demonstrating their convergence towards the normal distribution.

In our simulations, we consider non-uniform probabilities to reflect realistic scenarios where class distributions are imbalanced. Specifically, we set $p_{\text{TP}} = 0.3$, $p_{\text{TP}} = 0.2$, $p_{\text{FP}} = 0.1$, and $p_{\text{TN}} = 0.4$. We generate one million samples for each sample size n to observe how the distribution of the metrics evolves.

As shown in Figure 4.3, both the accuracy and marginal benefit metrics exhibit decreasing skewness and increasing kurtosis as n grows. This behavior indicates that the distributions are becoming more symmetric and peaked, aligning with the characteristics of the normal distribution. The convergence towards normality supports the validity of using the normal approximation for these metrics in our MATCH Test, particularly for larger sample sizes.

Assessing and Improving the Normal Approximation's Precision

While the normal approximation is generally acceptable for large sample sizes, it may not provide sufficient accuracy for small n. To address this, we examine theoretical bounds on the approximation error and explore more precise approximations applicable to smaller sample sizes.

The Berry-Esseen theorem provides an upper bound of $O\left(n^{-1/2}\right)$ on the error of the normal approximation to the distribution of a sum of independent random variables (Berry, 1941; Esseen, 1942). Specifically, it states that the maximum difference between the cumulative distribution function $F_n(x)$ of the standardized sum and the standard normal cumulative distribution function $\Phi(x)$ is bounded by:

$$|F_n(x) - \Phi(x)| \leq \frac{C_0 \cdot \gamma}{\sigma^3 \sqrt{n}} \tag{4.45}$$

where C_0 is a constant between 0.4097 and 0.5600 (gustav Esseen, 1956; Shevtsova, 2010), γ is the third absolute central moment of the summands, σ^2 is the variance of the summands, and n is the number of summands.

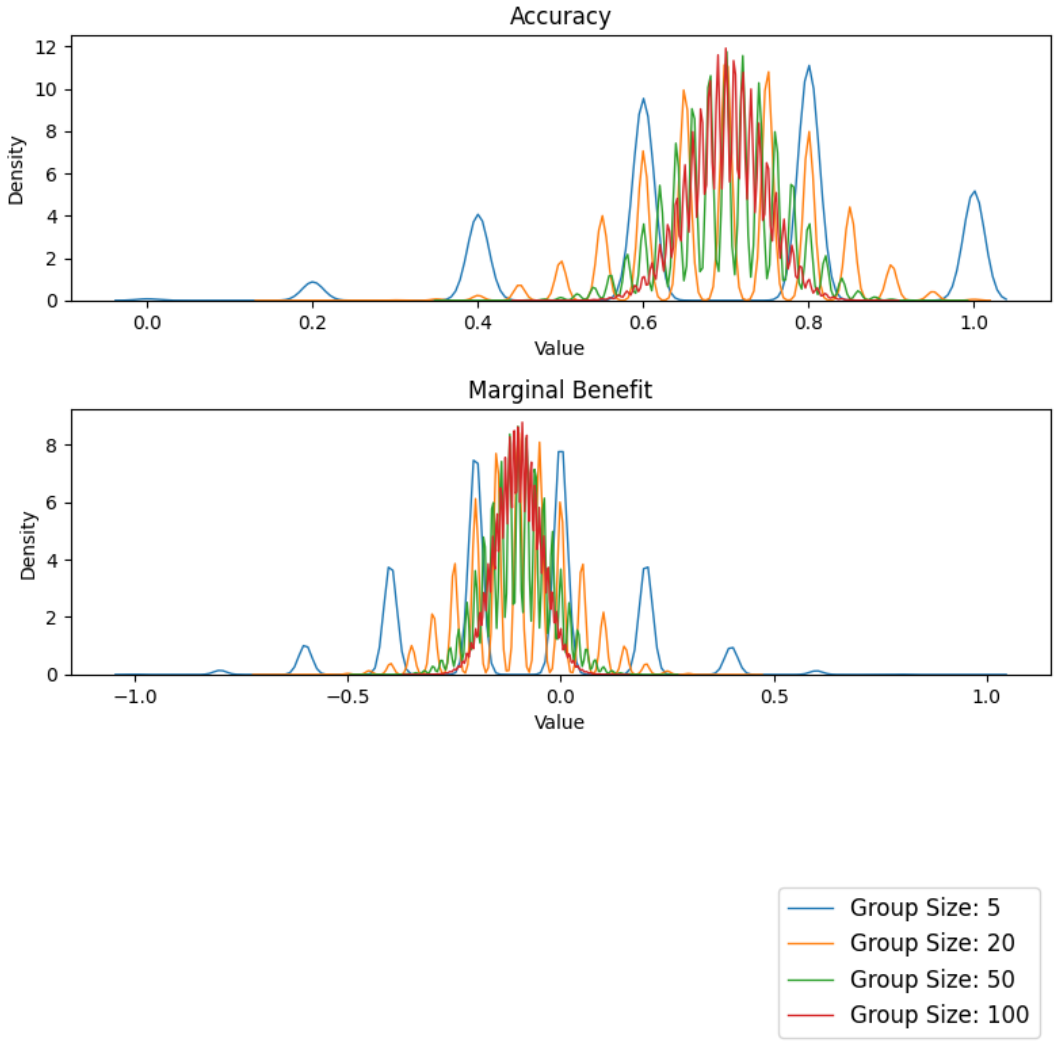


Figure 4.3: Probability density functions of accuracy and marginal benefit metrics for varying sample sizes n. As n increases, both metrics converge towards the normal distribution.

To reduce approximation error, we can use Peizer and Pratt's (1968) normal approximation for the binomial distribution. This is far more complicated, but the error becomes less than 0.1% for $\min(x+1, n-x) \geq 2$ (Johnson et al., 2005).

By having binomial metrics and $\mathcal{B}$ analogous to binomial distributions, we improve the bounds of what can be approximated within 0.1% by a better z-score approximation. Peizer and Pratt recommend approximating the z-score as:

$$\frac{x+\frac{2}{3}-\left(n+\frac{1}{3}\right)p}{\sqrt{\left(n+\frac{1}{6}\right)pq}}\times\frac{\sqrt{npq}}{\left(x+\frac{1}{2}-np\right)}\times\left\{2\left[\left(x+\frac{1}{2}\right)\ln\left(\frac{x+\frac{1}{2}}{np}\right)+\left(n-x-\frac{1}{2}\right)\ln\left(\frac{n-x-\frac{1}{2}}{nq}\right)\right]\right\}^{1/2}. \tag{4.46}$$

This gives good results, but for marginal improvement, add

$$\frac{1}{50}\left[(x+1)^{-1}q-(n-x)^{-1}p+(n+1)^{-1}\left(q-\frac{1}{2}\right)\right] \tag{4.47}$$

to the expression $x+\frac{2}{3}-(n+\frac{1}{3})p$. Altogether, we obtain the precision of strictly less than 0.1% for all x, n such that $\min(x+1, n-x) \geq 2$.

The proper parameters for z_p for binomial metrics and $\mathcal{B}$ are given in Table 4.1.

Table 4.1: Proper Parameters for Peizer and Pratt's z-score Approximation.

Metrics	x	p	q	n
Binomial metrics	k_{i+j}	p	q	n
$\mathcal{B}$	$k_{\text{FP}}-k_{\text{FN}}$	p_+-p_-	p_0	n

4.6 Cross-Prior Smoothing

In this section, we introduce *Cross-Prior Smoothing* (CPS), a novel correction method designed to enhance the reliability of CM metrics, particularly when comparing groups of varying sizes or distributions. CPS addresses the inherent variability and instability in classification metrics that arise due to sample size disparities, while preserving the consistency of bias assessments across different subgroups. Additionally, CPS drastically reduces the likelihood of encountering zero counts in corresponding cells of CM_i, which can lead to undefined scores. Furthermore, we empirically demonstrate the effectiveness of CPS in improving the reliability of classification metrics.

4.6.1 Methodology

CPS leverages information from a reference group's confusion matrix, CM_{ref}, to inform and smooth the metrics derived from a target group's confusion matrix, CM_i. This approach is especially pertinent in fairness analysis, where both CM_i and CM_{ref} are generated by the same underlying algorithm but may represent different demographic groups.

Let CM_{total} denote the confusion matrix derived from the entire dataset, and $\text{CM}_i \subset \text{CM}_{\text{total}}$ be the confusion matrix for subgroup i. In our experiments, we define the reference confusion matrix, CM_{ref}, as the portion of CM_{total} excluding the data from group i: $\text{CM}_{\text{ref}} = \text{CM}_{\text{total}} \setminus \text{CM}_i$.

Assumption 4.1. *The reference confusion matrix* CM_{ref} *provides a sufficiently informative prior for correcting the metrics derived from* CM_i.

As we demonstrate in the following experiments, this assumption is natural in practice, particularly given that once the parameters are fixed for a given model, the multinomial parameters are simple to estimate.

We define CPS in Algorithm 1. We normalize CM_{ref} to prevent it from dominating the target group that has small sample sizes. Furthermore, the choice of λ is crucial for balancing the contribution of the reference group's data. We set $\lambda = 5$ for all experiments.

Algorithm 1: Cross-Prior Smoothing (CPS)

Input: Confusion matrix for group i, CM_i; Normalized reference confusion matrix, $\hat{\text{CM}}_{\text{ref}}$; Smoothing constant λ
Output: Smoothed confusion matrix CM_{smooth}
Create the concentrations (α) using the prior;
foreach $c \in \text{CM}_i$, $c' \in \hat{\text{CM}}_{\text{ref}}$ **do**
    $\alpha_c = c + \lambda \cdot c'$;
foreach $c_{\text{smooth}} \in \text{CM}_{\text{smooth}}$ **do**
    Normalize the posterior distribution;
    $c_{\text{smooth}} = \alpha_c / (\sum_{\forall \alpha} \alpha)$;
    Scale back for size-dependent metrics;
    $c_{\text{smooth}} \leftarrow c_{\text{smooth}} \cdot \mid CM_i \mid$;
return Smoothed confusion matrix CM_{smooth};

This algorithm is inspired by the Dirichlet distribution, which is a conjugate prior to the multinomial distribution. However, this approach extends beyond a basic Bayesian estimate with a non-informative prior. By incorporating a reference group's data, the CPS technique provides a more robust estimate, particularly when the sample size n is small. This smoothing

method significantly improves the stability and reliability of metric estimates by reducing the variability and jaggedness that arises in small-sample scenarios.

4.6.2 Experiments and Results

We conduct extensive experiments across fifteen commonly used classification metrics, including the eight JRMs, accuracy, prevalence, predicted positive rate, marginal benefit, Matthews Correlation Coefficient, F_1 score, and prevalence threshold. The experiments are performed on two datasets: the COMPAS dataset and the Folktables income dataset.

COMPAS Dataset: The COMPAS dataset comprises risk assessments used in the U.S. judicial system, which have been scrutinized for potential bias (Angwin et al., 2016b; Larson et al., 2016; ProPublica, 2016). We utilize the confusion matrices reported in ProPublica to evaluate the effectiveness of CPS in a real-world fairness context.

Folktables' Income Dataset: We also employ the Folktables income dataset, training a random forest classifier to predict whether individuals earn more than $50k per year based on features such as marital status, race, and education. For bias assessment, we partitioned the data into eight racial groups and computed the corresponding confusion matrices.

For each metric, we consider various subgroup sample sizes, ranging from $n = 5$ to $n = 150$. To ensure statistical robustness, we conduct one million random samples for each sample size and metric combination. We compare the performance of the original metrics, metrics with additive smoothing ("bashful" smoothing with $\varepsilon = 1 \times 10^{-10}$ and the non-informative prior $\varepsilon = 1$), and metrics smoothed using CPS with the priori scales $\lambda = 5$, $\lambda = 10$, and $\lambda = 20$.

To obtain real-valued results, our experimental analyses excludes any undefined values. However, this could make direct comparisons between CPS and the original metrics (which often include such undefined cases) potentially unfair. As such, we include "bashful" smoothing ($\varepsilon = 1e - 10$) to maintain scores that closely resemble the original metrics. We improve over all original metrics. In Figures 4.4 and 4.5, we zoom into four different metrics with $\lambda \in [5, 20]$ and $\varepsilon \in [0, 1e - 10]$. The improvement in MCC is particularly encouraging, as this metric is well-recognized for its strong informational value (Chicco and Jurman, 2023).

We also explore the effect of a the more common non-informative prior, $\varepsilon = 1$, with prevalence and FNR in Figure 4.6. While additive smoothing with $\varepsilon = 1$ reduces errors for some metrics, it leads to inconsistent and less predictable results. In some cases, smoothing with $\varepsilon = 1$ performs worse than no smoothing (e.g., False Omission Rate (FOR), Negative Predictive

Value (NPV), and prevalence). In other cases, it performs better than CPS with $\lambda = 10$ for certain metrics (e.g., False Negative Rate (FNR) and TPR). However, CPS with higher λ values still outperforms additive smoothing with $\varepsilon = 1$ overall.

With a 95% confidence interval, the error bands in our experiments confirm statistical significance of CPS' improvement. The error bands are narrow and often indistinguishable, indicating high consistency across datasets and subgroups. This consistency is crucial for ensuring that the bias assessments remain reliable and robust across different groups.

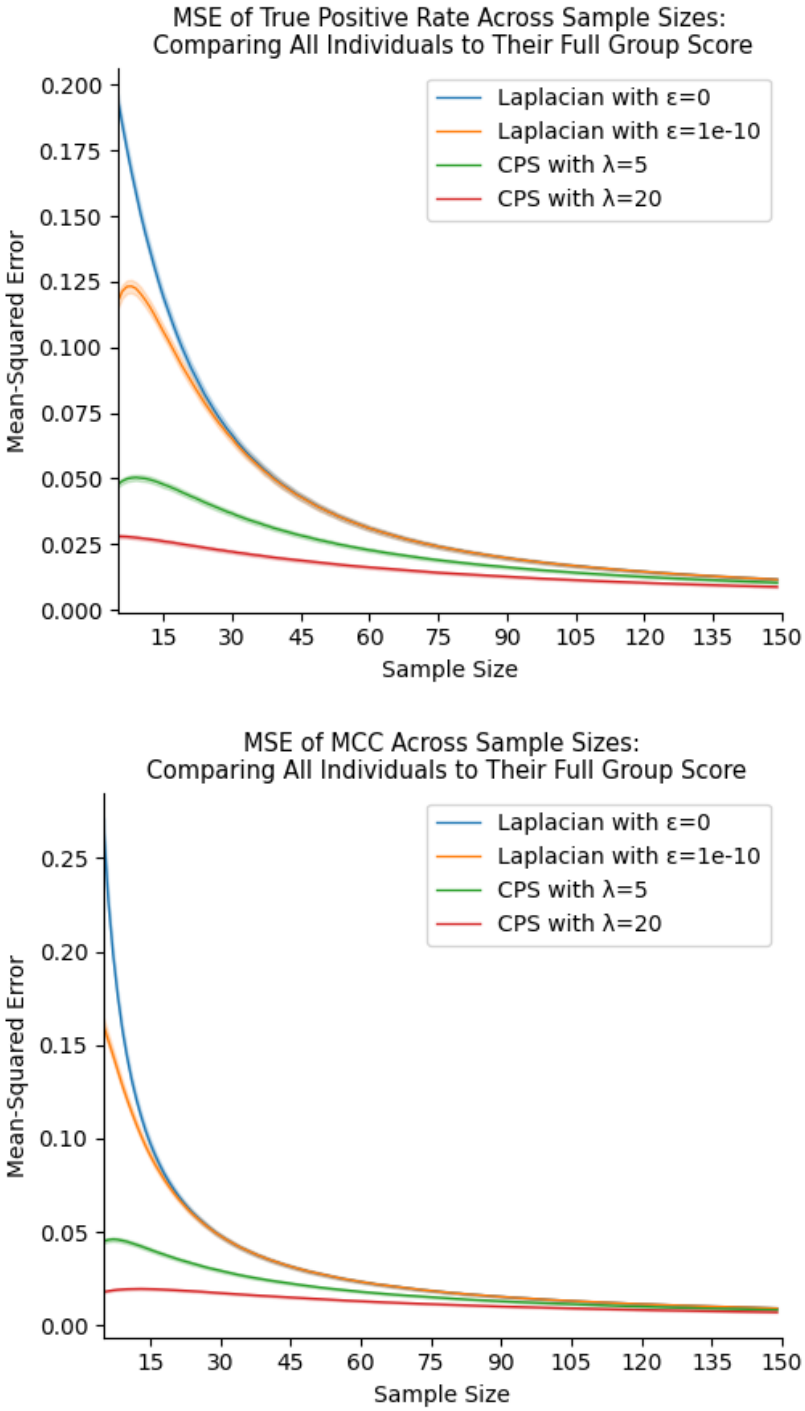


Figure 4.4: Effect of smoothing techniques on metrics with undefined values. Bashful smoothing ($\varepsilon = 1e^{-10}$) reduces errors compared to no smoothing ($\varepsilon = 0$). Cross-Prior Smoothing (CPS) offers further improvements, with the stronger prior ($\lambda = 20$) outperforming the weaker prior ($\lambda = 5$), indicating that CPS uses sufficiently informative priors.

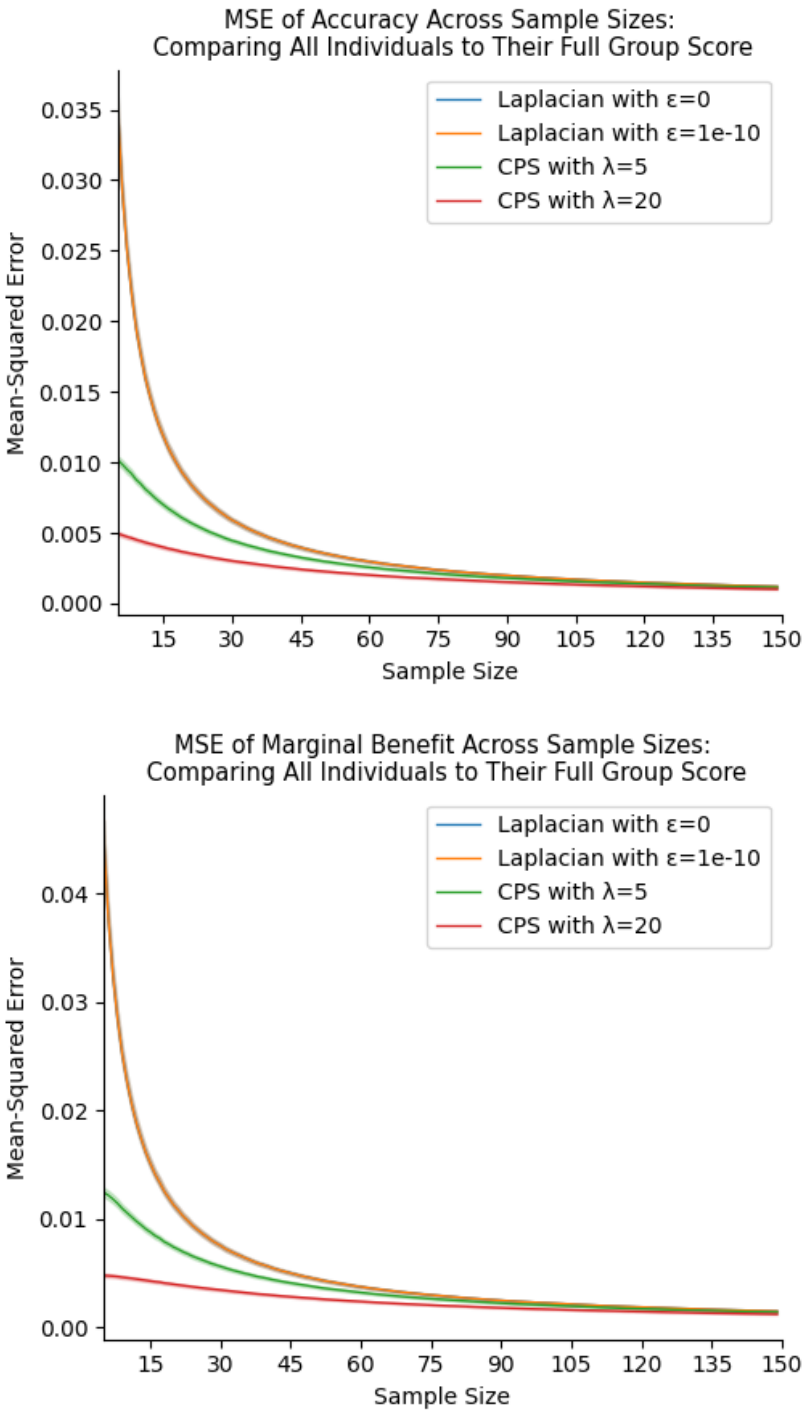


Figure 4.5: Comparison of smoothing effects on metrics without holes. Results show that applying a small smoothing factor ($\varepsilon = 1e^{-10}$) has minimal impact compared to no smoothing. CPS continues to reduce errors, following the same trend observed in Figure 4.4.

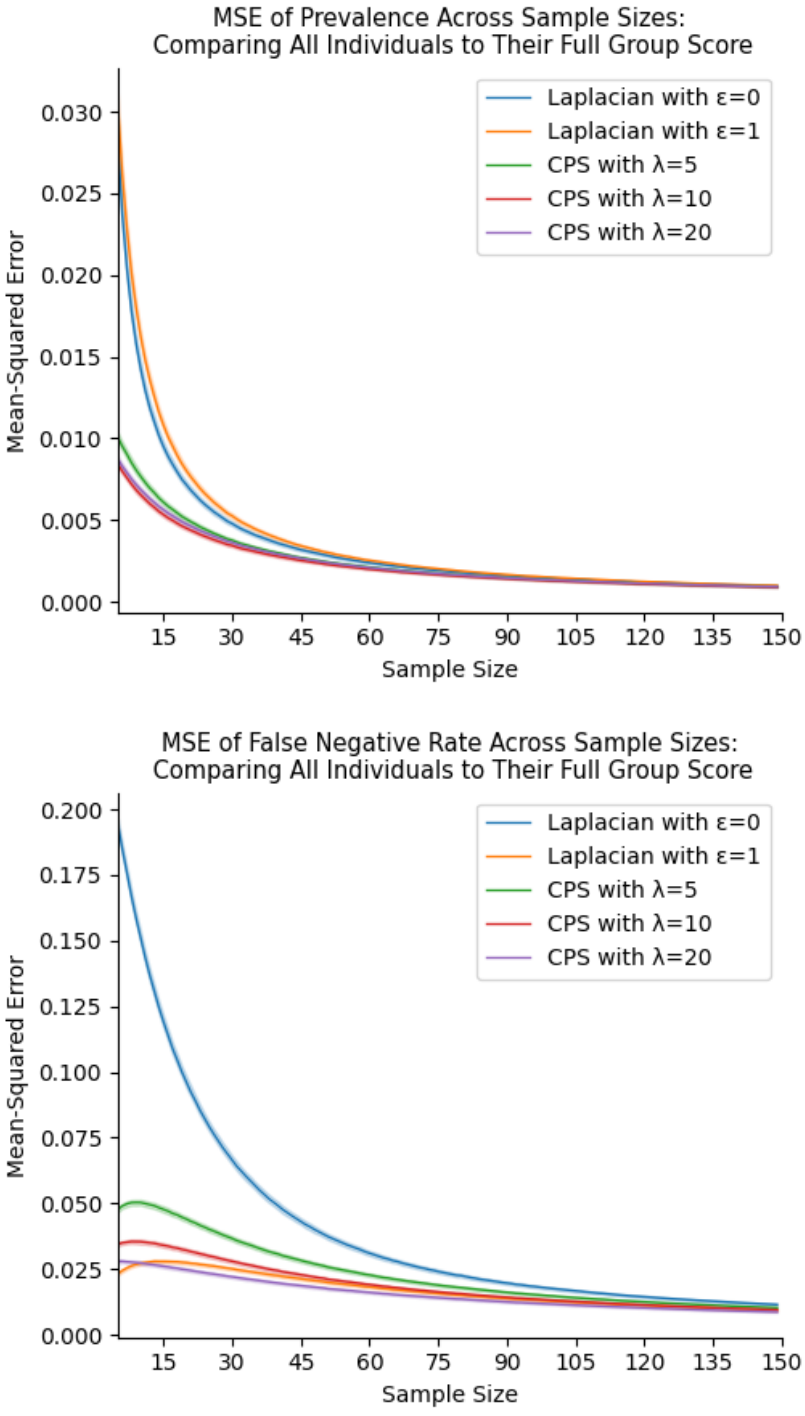


Figure 4.6: Smoothing with $\varepsilon = 1$ yields inconsistent results. In some metrics (left), it performs worse than no smoothing, while in others (right), it outperforms CPS with $\lambda = 10$.

4.6.3 Discussion

The experimental results validate Assumption 4.1, confirming that the reference confusion matrix provides a sufficiently informative prior for correcting the metrics derived from small subgroups. CPS consistently improves the reliability of classification metrics, especially in scenarios with small sample sizes where traditional metrics are prone to high variability and instability.

By leveraging information from a larger reference group, CPS effectively reduces the MSE of the metrics, leading to more accurate and stable performance assessments. The method also mitigates issues arising from undefined metric values due to zero counts, which are common in small-sample scenarios.

While additive smoothing can partially address undefined values, CPS offers superior performance by incorporating a data-driven prior. Increasing the smoothing constant λ enhances the benefits of CPS, as long as the reference group remains a valid representation of the target subgroup. Furthermore, CPS offers a practical and effective solution for correcting classification metrics in the presence of sample-size-induced bias.

It is important to note that the success of CPS hinges on the validity of Assumption 4.1. The reference group must provide a meaningful prior that is representative of the target group's distribution. In cases where this assumption may not hold, the effectiveness of CPS could be diminished.

4.7 Conclusions

In this chapter, we present a comprehensive analysis of sample-size-induced bias in confusion-matrix metrics, highlighting significant implications for evaluating classification performance and fairness across groups of varying sizes. The theoretical exploration revealed that small sample sizes lead to increased variability and jaggedness in metric scores due to the discrete and combinatorial nature of confusion matrices.

To address these challenges, we propose two novel approaches. First, we introduce the *MATCH Test* that assesses the statistical significance of an observed metric score relative to a reference distribution. This test enables more informed comparisons across groups by accounting for the variability introduced by different sample sizes. Second, we develop *Cross-Prior Smoothing* (CPS), a method that leverages prior information from a reference group's confusion

matrix to correct and stabilize metrics derived from smaller subgroups. CPS consistently improved the reliability of all 15 metrics tested, as evidenced by reductions in mean-absolute error and mean-squared error in the experiments on the COMPAS and Folktables income datasets.

Our findings underscore the importance of accounting for sample-size-induced bias when interpreting classification metrics, especially in fairness assessments where subgroup comparisons are critical. By revealing that sample size can be manipulated to present misleadingly favorable results, we highlight a vulnerability in current evaluation practices.

Future work could explore optimizing sample sizes to minimize metric variability and potential exploitation. Investigating the most advantageous configurations that could be misused and refining the MATCH Test by incorporating uncertainty in reference probabilities are also promising directions. Additionally, extending the theoretical foundations of CPS and examining its applicability to a broader range of metrics and settings (such as multiclass classification) would further enhance its utility.

By providing both theoretical insights and practical solutions, this work contributes to the development of more robust and equitable evaluation methods in machine learning, promoting fairer and more accurate assessments across diverse populations.

Chapter 5

ROBUST TIME-SERIES FORECASTING ACROSS DOMAINS: PROBABILISTIC PARALLEL TIME NETWORKS (PTN)

Time series forecasting remains a complex challenge, especially in intricate domains such as chaotic systems. This paper introduces "Probabilistic Parallel Time Networks", a novel deep-neural-network hypermodel designed to generalize across diverse domains in complex, probabilistic time series forecasting. The learned multi-modal model takes three distinct inputs: static data, agent history, and forecasts, and subsequently predict probability distributions over time. This innovative approach offers a more interpretable and unified framework for complex forecasting tasks. The utilization of a genetic hyperparameter optimization algorithm intertwined with the Differentiable Architecture Search (DARTS) neural architecture search, and the careful minimization of the neural architecture space, ensures a tractable and efficient solution. Through empirical studies in weather forecasting for microclimates and the chaotic Lorenz system, we demonstrate the model's applicability, robustness, and potential. The Probabilistic Parallel Time Network model offers a promising advancement in time series forecasting, with potential implications for research and practical applications in diverse areas, including healthcare, finance, and environmental sciences.

5.1 Introduction

In an era where data-driven decision-making is paramount, forecasting complex, temporal phenomena is crucial across various domains. From healthcare and financial markets to transportation and energy management, accurate forecasting is the cornerstone of strategic planning and risk mitigation. The nonlinearity and complexity of many real-world systems necessitate models that capture intricate temporal dynamics, offer robust generalizability, and interpretability. Furthermore, with the large successes found in meta-learning, we strive to move past the

proposals of neural architectures and propose novel model spaces or hypermodels. In this work, we introduce the Probabilistic Parallel Time Network, a novel hypermodel designed to address these challenges.

Probabilistic Parallel Time Network (PTN) is a hypermodel yielding a multi-modal neural network with three inputs: static, history, and forecasts. The static input gives context about the agent or state, the history gives temporal context, and the forecasts are given by a given deterministic forecaster. As a hypermodel, PTNs generalize across various use cases, as the algorithm learns the network's architecture to the given task. In this work, we highlight farm microclimates and chaotic systems. Furthermore, we demonstrate that Probabilistic Parallel Time Networks have improvement over Google's TiDE (Das et al., 2023) and PTN's given deterministic forecaster. Additionally, PTN's scalability is promising with its successful in the few-shot learning of new farm microclimates, assisted by each farm's static data.

The real-life application of weather forecasting for the microclimates of farmland is of paramount importance. Many farmers in the western United States operate in unique microclimates that publich weather forecasts do not accurately represent. This bias creates hardships for farmers who rely on forecasts for crop scheduling and preventative planning. PTN's showcasing of success in learning farm microclimates suggests promising applicability to other temporal domains with underrepresented groups. Examples include, but are not limited to: environmental conservation in small environments, often containing less common or endangered animals; niche markets in financial datasets, providing better risk management and investment decisions; microcosms in society, allowing one to predict human behavior in social groups that deviate from the norm; and energy conservation in microgrids.

Next, PTN successfully gives probabilistic forecasts for one of the most difficult deterministic systems, the chaotic Lorenz system. This application serves as a key demonstration of PTN's strong capabilities, showcasing PTN's generalizability to capture complex temporal dynamics. Furthermore, we highlight Probabilistic Parallel Time Network's interdisciplinary relevance as its temporal distributions allow a unique perspective to explore chaos theory by quantifying what sequences are often uncertain. PTNs successfully capturing chaotic systems suggest success for probabilistic forecasting in epidemic modeling (Mangiarotti et al., 2020), human behavior modeling (Guess and Sailor, 1993; Ward and West, 1998), robotic behavior modeling (Zang et al., 2016), and other complex temporal problems.

Our contributions in this chapter are as follows:

- **Introduction of Probabilistic Parallel Time Networks:** We propose a novel approach to provide probabilistic forecasts for complex time-series. This probabilistic nature enhances user trust, as the model communicates its uncertainty effectively.
- **Concept of "Parallel Time".** We incorporate multiple forecasts over time into the framework of Probabilistic Parallel Time Networks, enabling richer temporal insights.
- **Demonstration of PTN's Effectiveness:** We validate the success of PTN in forecasting estimated probability-density functions for two challenging scenarios: microclimate-weather data and the chaotic Lorenz system.

5.2 Related Work

We discuss related works related to PTNs that we have built upon, including meta-learning techniques such as neural architecture search algorithms (NAS) and hyperparameter optimizations (HPO), multi-modal networks, and model stacking.

Over the past several years, meta-learning has shifted the machine-learning paradigm (Verma et al., 2020; White et al., 2021; Bischl et al., 2023). We find that there are broadly eight categories: grid search, random search, evolutionary algorithms, gradient-based methods, Bayesian optimization, reinforcement learning, one-shot NAS, and hybrid methods.

Grid searches are exhaustive and iterate throughout the entire model space. Bergstra and Bengio finds that random searches, which randomly and naïvely traverse the model space often outperform grid searches (2012). Evolutionary NAS algorithms test each model in a generation with a fitness function; the model space is then traversed by using crossovers and mutations while considering the fitness score (Liu et al., 2020). Gradient-based methods shift the paradigm from testing discrete model subspaces by relaxing it into a continuous space, DARTS is such an algorithm (Liu et al., 2019). Additionally, Bayesian optimization builds a probabilistic model of the objective function to predict the most promising architectures (Zhou et al., 2019). Next, reinforcement-learning-based NAS algorithms use a controller with a learned policy to predict model candidates (Zoph and Le, 2016). Another approach is the one-shot NAS algorithm which trains a single large network (supernet) that includes all possible architectures as subnetworks. The supernet then estimates each architecture's performance (Bender et al., 2018). Finally, hybrid methods combine two or more approaches. Our PTN is a hybrid method combining evolutionary hyperparameter optimization with a gradient-based NAS.

Multi-modal networks allow the network to reason from different representations. For example, CLIP and GPT-4 allows image and text inputs to allow a greater flexibility of communication. Additionally, having different modalities has been found to increase the model's understanding (OpenAI, 2023).

Finally, stacking models is when one pipes a model's output into another model's input. Model stacking has been shown to improve forecasting results (Pavlyshenko, 2018). While other stacking methods only include the most recent forecast, we introduce parallel time by including the sequence of forecasts in a two dimensional framework with channels: (forecast number, time within the forecast, channels).

In weather forecasting, most deep learning techniques for weather forecasting employ spatiotemporal methods (Yuan et al., 2022; Bojesomo et al., 2021; Diehl et al., 2015), these techniques also assume the presence of numerous neighboring stations. Our PTN incorporates this by stacking the National Blend of Model's (NBM) forecasts (Craven et al., 2020). However, the challenge comes from the farm terrain being substantially different from most of the forecasted grid.

5.3 Data Preparation

In this section, we discuss our data preparation and processing techniques for the two scenarios we study: weather forecasting and analysis of chaotic Lorenz systems.

5.3.1 Weather Data

For weather forecasting, our learned model input includes static information about farms (such as elevation), temporal weather history sampled within the farms, and NBM's forecasts. The static data includes the latitude, longitude, elevation, slope, aspect, and exposure of the farms. Slope is the angle of the terrain, aspect is the orientation the terrain faces, and exposure measures how sheltered an area is.

The farm history includes observations of air temperature, atmospheric pressure, vapor pressure, wind speed, wind direction, wind gusts, solar radiation, and dew point. Each measurement is taken from two meters above the ground every fifteen minutes. Solar radiation is the intensity of the sun upon the given terrain. The dew point is the temperature at which the water vapor will begin to condense. We calculate dew point using MetPy (May et al., 2022).

Our hourly NBM forecasts include temperature, dewpoint, solar radiation, wind direction, wind speed, and total cloud cover for 36 hours. We forecast up to 40 hours out with fifteen minute intervals.

Except for the time features, we normalize all features between zero and one. For time, we properly capture its cyclical measure before scaling. We convert each date into a numerical day of year (DoY) and each time of day into "seconds from midnight" (SFM). Then, we normalize DoY and SFM to $[-2\pi, 2\pi)$ before casting both time features into two pairs of sine and cosine. After the trigonometry conversions, we normalize the time data between zero and one.

5.3.2 The Chaotic Lorenz System

Chaotic systems are deterministic dynamical systems that exhibit highly sensitive dependence on initial conditions. The sensitivity is severe enough that it appears random and unpredictable—a minute change from an input of (10,10,10) to (10.00000000001,10,10) yields a substantial change over time in the output. This is known as the "butterfly effect". Additionally, the systems have "topological mixing" where points in the system's phase space eventually become close to each other. Furthermore, chaotic systems have dense periodic orbits, where every point in the space is approached arbitrarily close by periodic orbits. We focus on the Lorenz system (Lorenz, 1963).

The Lorenz system is a simplified mathematical model for atmospheric convections. This chaotic system's phase space is a 3D Cartesian coordinate system with three dependent variables: (x, y, z) and three parameters (σ, ρ, β). Formally, we name it $\mathcal{L}(x, y, z; \sigma, \rho, \beta)$ or $\mathcal{L}$ for short. $\mathcal{L}$ is defined by an ordinary differential equation for each of its three axes:

$$\frac{dx}{dt} = \sigma(y - x), \tag{5.1}$$

$$\frac{dy}{dt} = x(\rho - z) - y, \tag{5.2}$$

$$\frac{dz}{dt} = xy - \beta z. \tag{5.3}$$

For our experiments, we sample our initial dependent variables from uniform distributions:

$$(x_0, y_0, z_0) \sim \big(U(-20, 20), U(-20, 20), U(0, 40)\big) \tag{5.4}$$

$\mathcal{L}$ is parameterized by σ, ρ, and β. The typical values are references from Lorenz's original

work.

- σ: The Prandtl number represents the ratio of momentum diffusivity (how momentum is transferred through a fluid due to its viscosity) to thermal diffusivity (how quickly heat is transferred through a material). Typically, $\sigma = 10$ and straying too far from 10 may have uninteresting dynamics. Our experiments sample σ from a uniform distribution bounded by 8 and 12: $\sigma \sim U(8, 12)$.
- ρ: The Rayleigh number represents the temperature difference between the top and bottom of the fluid layer. Typically, $\rho = 28$ and increasing ρ often leads to more chaos. For our experiments, $\rho \sim U(27.9, 28.1)$.
- β: A geometry factor that is related to the aspect ratio of the physical system, specifically the ratio of the height to the width of the convective rolls. Large deviations from $8/3$ might lead to unphysical behaviors. Typically, $\beta = 8/3$. We use $\beta \sim U(7.9/3, 8.1/3)$ in our experiments.

$\mathcal{L}$ partly achieved such notoriety due to its strange attractor as visualized in Figure 5.1. This is a subset of the phase space with a fractal structure where trajectories from a wide region are attracted to. Once trajectories enter the attractor, they stay there, wandering chaotically with no repetition. Due to $\mathcal{L}'s$ significant numerical instability, machine learning is an attractive method to predict several steps in the future when exact parameters or coordinates are unknown. Furthermore, an approximate probability density function is further appealing since we can see possible paths and their likelihoods. This is as opposed to the deterministic forecast, which may take the center of two trajectories in times of high uncertainty even if the center is impossible. For example, if a bit is randomly assigned 0 or 1 with a uniform distribution, its optimal for a deterministic forecast to predict 0.5. However, a distribution can correctly predict 0 or 1, both with 50% probability.

In our Lorenz system case study, our three inputs follow: the static data is the parameters (σ, ρ, β), the temporal history are the coordinates (x, y, z) over sixteen time steps, and ten iterations of forecasts with sixteen time steps each are derived from our trained deterministic forecaster, a multilayer perceptron based on the nonlinear vector autoregressive model (NVAR). We choose this structure since Shahi et al. (2022) found the NVAR model to outperform other machine learning techniques. We then forecast the x coordinate's approximate probability density function for sixteen time steps using 111 quantiles. All of our inputs are normalized between zero and one.

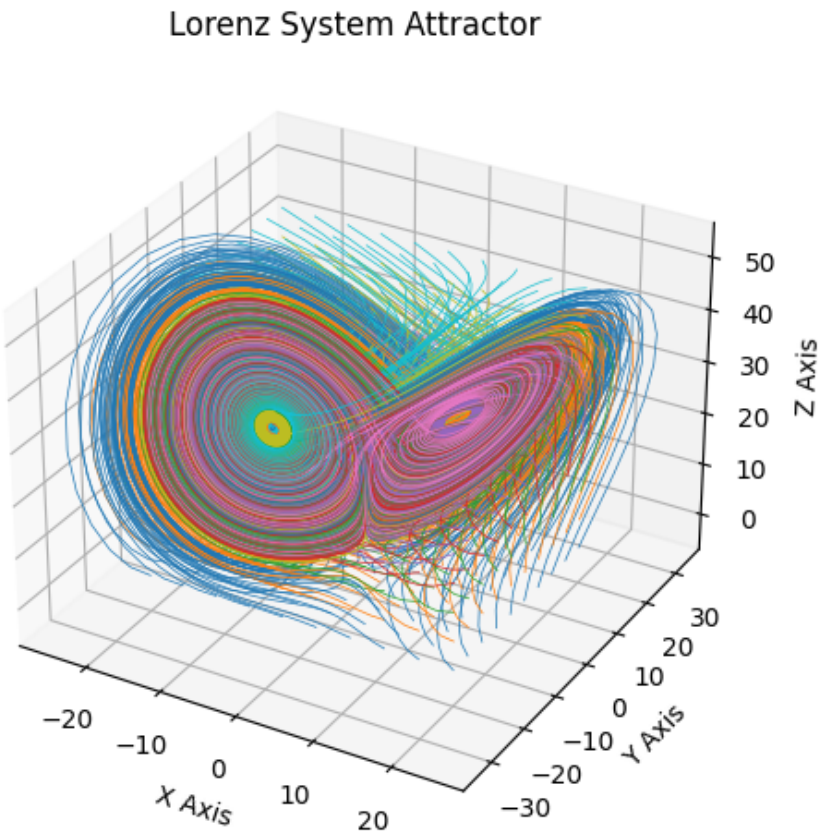


Figure 5.1: The Lorenz System's strange attractor. Every trajectory is chaotic but once they are near enough to the attractor space, they never leave the space nor visit the same position again.

5.4 Probabilistic Parallel Time Networks

PTN is a hypermodel and is generalizable to any time series applications such as weather, video game actions, finances, and epidemics. The PTN hypermodel takes three inputs: agent history, agent static data, and a series of forecasts (the parallel time aspect). Then, the PTN outputs a multi-step probabilistic forecast by using n quantiles. Notably, we are able to produce probabilistic forecasts given deterministic targets. To illustrate the effectiveness and robustness of our framework, we present two use cases: weather forecasting and the chaotic Lorenz system. First, we describe the model space, as illustrated in Figure 5.2.

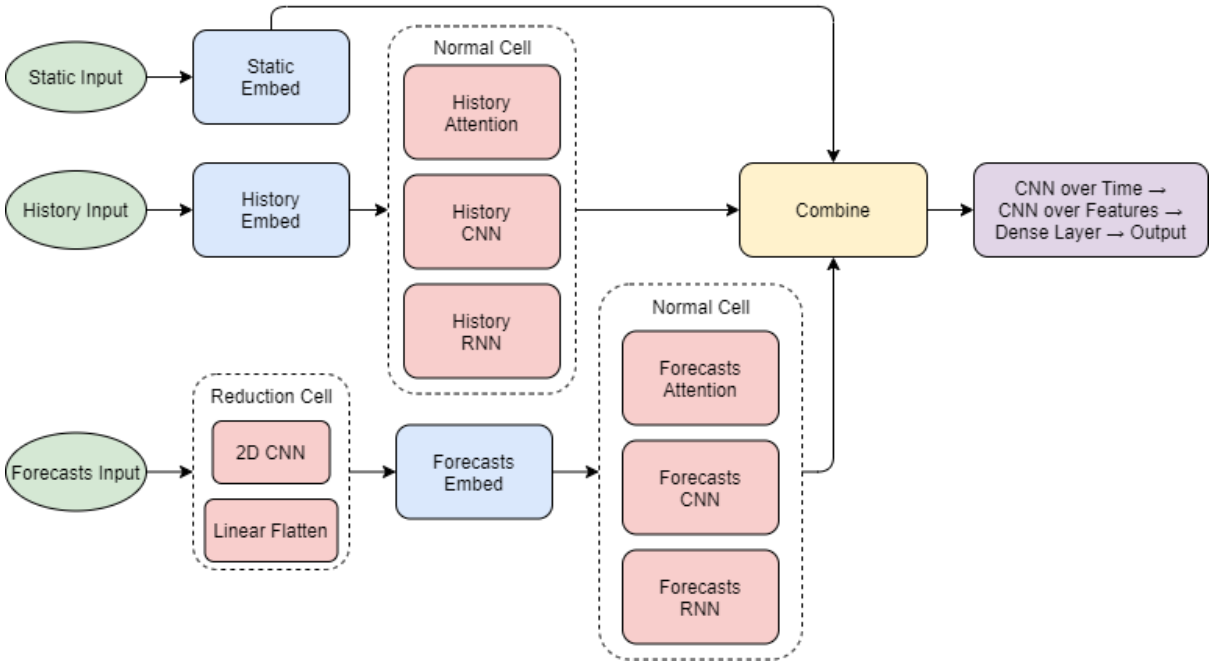


Figure 5.2: Probabilistic Parallel Time Network's architecture space. The cells represent choices the algorithm can make and select from one to all of the sub-architectures.

5.4.1 Model Space

To make the combination of genetic hyperparameter optimization and DARTS tractable, we restrict our model space. The DARTS algorithm contains two optimizers: one for the architecture and the other for the other parameters. The architecture parameters are the weights assigned to each neural layer and pass through a softmax function to learn the layer's importance for later pruning. The other parameters are trained with one data set and the architecture parameters are trained a separate data set. The architecture parameters are trained for five steps in each epoch

with batch sizes of 64. The other parameters are trained for ten steps per epoch. We test each model candidate for 30 epochs each throughout our hyperparameter and architecture search.

We use a separate AdamW optimizer (Loshchilov and Hutter, 2017) for both sets of parameters. Both optimizers have their own "reduce on plateau" learning-rate scheduler with a patience of one. A patience of p reduces the learning rate by ten times if the loss is greater than the smallest seen loss for $p + 1$ consecutive epochs.

Each hypermodel subspace mentioned is followed by a set of architecture parameters with one parameter for each layer that signal the given layer's importance. This is part of the DARTS algorithm (Liu et al., 2019).

Architecture Space

We embed each of our three inputs (static, history, and forecasts) into a $d \times d$ space. Static data is a vector of c_s real valued features.

History is a matrix with a temporal τ_h and channel c_h dimensions. Finally, forecasts is represented by a $\mathbb{R}^{f_f \times \tau_f \times c_f}$ tensor

where f_f is the number of the most recent forecasts, τ_f is the number of time steps predicted by each forecast, and c_f is the quantity of channels forecasted. Recall that our forecast input is given by an arbitrary, deterministic forecaster. The hyperparameters are drawn from the hyperparameter space in Section 5.4.1. For example, the Leaky ReLU's (LReLU's) negative slope is learned from the range $[0.0, 0.3]$. All of our 1D-CNN layers use causal padding.

The static data's embedding block begins with Gaussian noise regularization, a fully connected layer transforming the dimensions $c_s \mapsto d$, batch normalization over the channel dimension, and a LReLU activation. Next, the d dimensions are repeated along the temporal dimension ($1 \times d \mapsto d \times d$), followed by the first dropout layer, a dense layer, and the LReLU activation. We choose to implement Gaussian noise in the static data to allow generalizability for new systems and mitigate memorization.

The history embedding block is a dense layer ($\tau_h \times c_h \mapsto \tau_h \times d$) followed by an LReLU activation and a dropout layer. Next, the temporal dimension is changed by using a 1D-CNN layer over the temporal dimension ($\tau_h \times d \mapsto d \times d$) followed by the LReLU activation. Next, the data travels through three layers with learned architecture parameters: multihead attention, 1D-CNN, and a bi-directional GRU RNN. These three layers are part of the hypermodel subspace and pruned if the parameters find them unsubstantial.

The forecasts embedding block begins with a hypermodel subspace with two possible layers: a 2D-CNN (named Conv2D Collapse) and a dense layer. Both of which transform the data from $f_f \times \tau_f \times c_f$ to $f_f \times \tau_f$. Next, the forecasts are cast to $d \times d$ space in the same manner as the history, by a dense layer, then an LReLU activation, which is followed by a 1D-CNN over the temporal dimension. Following the forecasts embedding, we use another hypermodel subspace of three layers: a multihead attention layer, a 1D-CNN, and a bi-directional GRU RNN.

To help with interpretibility, we also multiply all three embedding spaces with a unique learnable parameter before their fusion by summation. Although, we do not prune any of them.

After fusion, we apply the learned normalization layer, use a 1D-CNN over the temporal dimension changing the space from $d \times d$ to $\tau_t \times d$, implement layer normalization and an LReLU activation, implement a 1D-CNN over the channel dimensions with an LReLU ($\tau_t \times c_t$), then our output layer is a dense layer with no activation. This output is now c_t quantiles over τ_t timesteps. The unbounded output allows the quantiles to go beyond the range of the trained data.

Hyperparameter Space

Our hyperparameters have two types: choice and range. If choices are unordered, the crossover elicits one of the parents' current choices with probability proportional to the parents' fitnesses. Otherwise, the child receives the interpolated choice, weighted by the parents' fitnesses. Similar to ordered choices, the range types elicit weighted interpolated values in their crossovers. If the range is logarithmic, the interpolation happens in the natural log space.

In Table 5.1, we provide PTN's hyperparameter space. The hidden size d is sampled from the set of all integers between 72 and 255 that are divisible by the least common multiple of the number of MHA choices (24). This is due to the restrictions of the multihead attention block. We omit the attention head count of one since this value often performs worse in our preliminary experiments and as implied in Vaswani et al. (2017). Additionally, we cap the hidden size to 255 since we find an increase to 256 doubles the training time in our environment and often gives less favorable results. The first-layer dropout rate is learned separately from the dropout rate in other layers in case too much information would be lost on the first dropout. The kernel sizes are odd to prevent the copying overhead of even kernel sizes as implemented in PyTorch.

Algorithm 2: Probabilistic Parallel Time Network Optimization

Randomly initialize current population A_c;
Track all individuals in the complete set A;
for $generation = 1$ **to** $num\_generations$ **do**
    // Evaluate the fitness of the current population.
    foreach $a \in A_c$ **do**
        Encode the model, then train using DARTS;
        Update fitness and validation fitness for a;
    $A \leftarrow A \cup A_c$ // Add current population to the global set
    // Check for early stopping based on patience threshold p.
    if *No improvement in minimum* $f_v \in A$ *for* p *generations* **then**
        break;
    // Every e_f rounds, inject the top n_e agents, decided by f_v.
    if $generation \mod e_f = 0$ **then**
        $A_c \leftarrow \text{sort}(A_c; f_v)[1, 2, \ldots, n - n_e]$;
        $A_c \leftarrow A_c \cup \text{sort}(A \setminus A_c; f_v)[1, 2, \ldots, n_e]$;
    // Generate new population using tournament selection.
    $A_c \leftarrow \emptyset$;
    foreach *Parent pair* **do**
        Create child c using proportional crossover of parent traits;
        Estimate child's fitness as the average f_v of both parents;
        Mutate c with probability $\propto$ expected fitness in $[0.05, 0.6]$;
        $A_c \leftarrow A_c \cup \{c\}$;
// Sort agents by their combined training and validation fitness.
return $\text{sort}(A; f^2 + f_v^2)$;

5.4.2 Genetic Hyperparameter Optimization with DARTS

We simultaneously traverse the neural architecture and hyperparameter search spaces by implementing DARTS within the genetic HPO.

Our evolutionary algorithm is explained in Algorithm 2. This algorithm implements state-of-the-art techniques such as proportional crossovers, a dynamic mutation rate, elite rounds, the DARTS NAS algorithm, and a patience threshold for the population's evolution. We then select the x best agents by the smallest sum of squares for the fitness and fitness validation scores. Our data is partitioned five ways: neural training and neural validation for the DARTS training, genetic training and genetic validation for measuring the fitness and fitness validation scores, and a test data set.

The variable names for Algorithm 2 are defined as follows:

- (f, f_v): the fitness and validation fitness scores.
- e_f: the frequency of elitism rounds.
- n_e: number of elite agents injected into the population.
- $sort(S; x)$: sort S by x in descending order.
- p: the patience generation count.
- n: number of agents within a living population.

We present our hyperparameter space in Table 5.1. The curly brackets denote a set, and the square brackets denote a range.

5.4.3 Loss and Fitness Functions

To minimize the loss when predicting the approximate probability distribution from deterministic target data, we minimize the approximate continuous ranked probability score (CRPS) using an estimated cumulative density function during run time (see Equation 5.7). Our approximation uses summations as opposed to the integrals used for probability density functions since our predicted quantiles are discrete. The fitness functions and evaluation metrics use CRPS as described in Equation 5.7 while our neural network optimizers use CRPS plus the ordering penalty as shown in 5.9.

Our notations are as follows:

- $\hat{y} \in \tau_t \times c_t$: the predicted quantiles over time.

Table 5.1: The Hyperparameter Space for PTN.

Name	Values	Type	Is Logarithmic	Is Ordered
Hidden Size d	$\{72, 96, \ldots, 72 + 24n, \ldots, 240\}$	Choice	False	True
Dropout Rate	$[0.0, 0.5]$	Range	False	N/A
First Layer Dropout Rate	$[0.0, 0.5]$	Range	False	N/A
Leaky ReLU Negative Slope α	$[0.0, 0.3]$	Range	False	N/A
History MHA Heads	$\{2, 3, 4, 6, 8, 12\}$	Choice	False	True
Forecasts MHA Heads	$\{2, 3, 4, 6, 8, 12\}$	Choice	False	True
Gaussian Noise's Standard Deviation	$[0.0, 0.25]$	Range	False	N/A
Architecture Parameters' AdamW's LR	$[0.00001, 0.001]$	Range	True	N/A
Architecture Parameters' AdamW's β_1	$[0.5, 0.99]$	Range	False	N/A
Architecture Parameters' AdamW's β_2	$[0.95, 0.999]$	Range	False	N/A
Other Parameters' AdamW's LR	$[0.00001, 0.001]$	Range	True	N/A
Other Parameters' AdamW's β_1	$[0.5, 0.99]$	Range	False	N/A
Other Parameters' AdamW's β_2	$[0.95, 0.999]$	Range	False	N/A
Normalization Layer	{Layer Norm, Batch Norm, None}	Choice	N/A	False
History Conv1D Embed Kernel Size	$\{1, 3, \ldots, 2n + 1, \ldots, 15\}$	Choice	False	True
History Conv1D Kernel Size	$\{1, 3, \ldots, 2n + 1, \ldots, 15\}$	Choice	False	True
Forecasts Conv2D Collapse Kernel Width	$\{1, 3, \ldots, 2n + 1, \ldots, 15\}$	Choice	False	True
Forecasts Conv2D Collapse Kernel Height	$\{1, 3, \ldots, 2n + 1, \ldots, 15\}$	Choice	False	True
Forecasts Conv1D Embed Kernel Size	$\{1, 3, \ldots, 2n + 1, \ldots, 9\}$	Choice	False	True
Forecasts Conv1D Kernel Size	$\{1, 3, \ldots, 2n + 1, \ldots, 15\}$	Choice	False	True
Output Kernel Temporal Size	$\{1, 3, \ldots, 2n + 1, \ldots, 15\}$	Choice	False	True
Output Kernel Features Size	$\{1, 3, \ldots, 2n + 1, \ldots, 15\}$	Choice	False	True

- $y \in \tau_t \times 1$: the deterministic target over time.
- $t \in [1, 2, \ldots, \tau_t]$: the time step.
- $T_1 \in \tau_t \times 1$: the MAE for each quantile by subtracting each corresponding deterministic target in time, element-wise. We formalize T_1 in Equation 5.5.
- $T_2 \in \tau_t \times 1$: measures the spread of $\hat{y}$ by calculating the distance between each pair and giving a summarized scalar. We formalize T_2 in Equation 5.6.

$$T_1 = \frac{1}{\tau_t} \sum_{i=1}^{\tau_t} \left(\frac{1}{c_t} \sum_{j=1}^{c_t} \left| \hat{y}_{i,j} - y_i \right| \right) \tag{5.5}$$

$$T_2 = \frac{1}{\tau_t} \sum_{i=1}^{\tau_t} \left(\frac{1}{c_t^2} \sum_{j=1}^{c_t} \sum_{k=1}^{c_t} \left| \hat{y}_{i,j} - \hat{y}_{i,k} \right| \right) \tag{5.6}$$

$$\text{CRPS} = T_1 - 0.5 T_2 \tag{5.7}$$

Following Takeuchi et al.'s results (2006), we add an ordering penalty O to the quantiles. Recall our predictions are in the space $\tau_t \times c_t$ for c_t quantiles over τ_t time steps. Our ordering penalty loss occurs iff the values are not monotonically non-decreasing, and is formalized in Equation 5.8.

$$O = \frac{1}{\tau_t \cdot (c_t - 1)} \sum_{i=1}^{\tau_t} \sum_{j=1}^{c_t - 1} \max(0, \hat{y}_{i,j} - \hat{y}_{i,j+1}) \tag{5.8}$$

Combining the ordering penalty with weight ω, we describe our loss function, ℓ in Equation 5.9.

$$\ell = \text{CRPS} + \omega O \tag{5.9}$$

We use $\omega = 0.0075$.

5.5 Experiments and Results

To our knowledge, no multi-modal, probabilistic forecaster like PTN exists. This causes difficulty for baseline comparisons. We overcome this and create two baselines. Our first baseline converts the deterministic forecasts to PDFs over time by assuming a normal distribution with the mean as their actual prediction and the standard deviation as their mean-absolute error.

We search the literature for a precedent to compare, and we find that Google's TiDE is the closest (Das et al., 2023). With its incorporation of static data, history, and dynamic covariate inputs, TiDE's input space is conceptually similar. However, TiDE does not include previous forecasts, nor does it output probabilistic forecasts. To accommodate, we update TiDE's GitHub implementation to feed TiDE forecasts in addition to the static data and history. Furthermore, we have TiDE produce estimated PDFs by changing their loss function from mean-squared error to ordered CRPS (Equation 5.9).

Our experiments are able to run on a local desktop with this hardware: RTX 4090 GPU, i9-13900K CPU, and 96 GB of DDR5 RAM. For weather forecasting, PTN takes about 8.5 hours. For the chaotic system, the PTN takes about 6.5 hours. This discrepancy is likely due to the large amount of data (about 60 GB) that we use in the weather forecasting opposed to about two gigabytes for the Lorenz system. Training the PTN's learned neural network takes between thirty and sixty seconds. For our analyses, we train the network on the data used in the DARTS training and validation data, then implement a warm start by training on a fraction of the left out data before testing on the rest of the left out data. For the weather data, we partition data by sites. For $\mathcal{L}$, we generate 50 systems with randomly initialized parameters and 125 different, random initial conditions for each data partition. For each system and instance, $\mathcal{L}$ has 1000 evenly spaced times between 0 and 100 at which the solution to the initial value problem is evaluated.

We use skill scores as defined in Equation 5.10 to compare PTN with each baseline. A skill score of one indicates a perfect forecast, between zero and one means the model outperforms the baseline, zero means the forecasts are equal, and a skill score below zero means that the baseline outperforms the given model. These results aggregate seven separate runs to ensure a representative sample. Please see all results in Table 5.2.

$$\text{Skill score} = 1 - \frac{\text{performance of model}}{\text{performance of baseline}} \in (-\infty, 1] \tag{5.10}$$

Since the skill score is susceptible to heavy left skews, we find that the mean is not useful information. As such, we collect the CRPS for every prediction for every model, then take the quartiles of each CRPS set. We compare PTN to the baselines by using the skill score with these quartiles. For further insight, we find the percentage of forecasts that PTN improves by locating the percentile of a zero score (where the forecasts are equally accurate) then subtract it from one. For example, a zero score located at the percentile of 6 means that 6% of the scores

Table 5.2: Probabilistic Parallel Time Network's Skill Scores

Dataset	Chaos		Microclimate	
Model	Det.	TiDE	NBM	TiDE
First Quartile	0.40	0.82	0.42	0.21
Second Quartile	0.57	0.73	0.59	0.18
Third Quartile	0.68	0.56	0.70	0.16
Percent of Forecasts Outperformed by PTN	94.3%	99.8%	96.4%	77.5%

are below zero so 94% are greater than or equal to zero. We find that PTN dominates baselines for forecasting chaotic systems, also dominates NBM for the microclimate, and does better in 77.5% of all forecasts than TiDE for microclimate weather forecasting.

For weather forecasting, we estimate 111 quantiles over forty hours with fifteen minute resolution. Thus, PTN produces 160 estimated probability density functions, one for each time step. This estimated PDF is illustrated in Figure 5.3 where we see that the actual temperature is entirely within the bounds of our quantiles. Additionally, we provide an illustration of TiDE's probabilistic forecasting is provided in Figure 5.4.

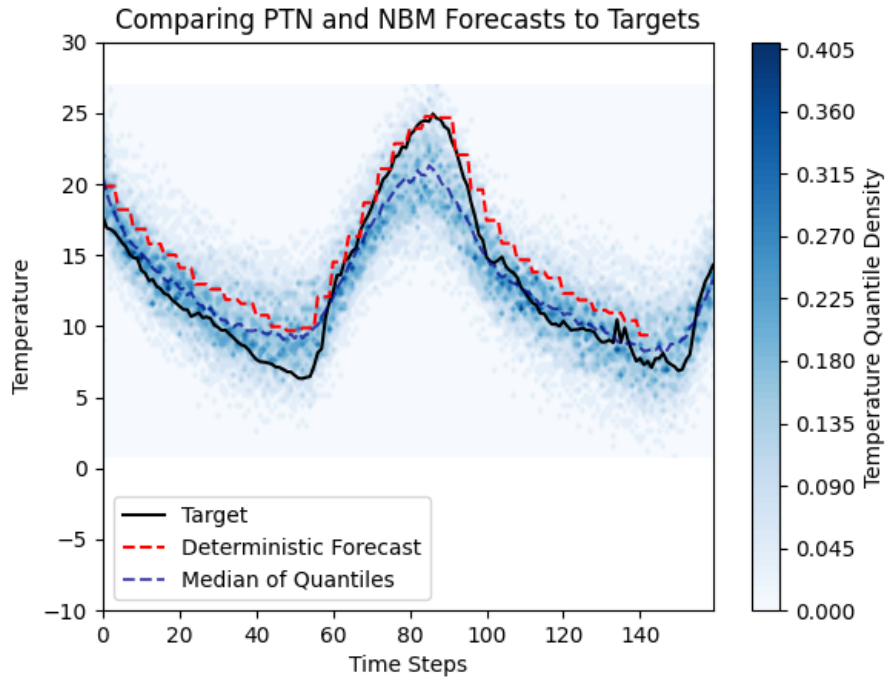


Figure 5.3: Here, we compare PTN's estimated PDFs (blue shading) over time to NBM's deterministic forecast (red line) for the microclimate's temperature (black line).

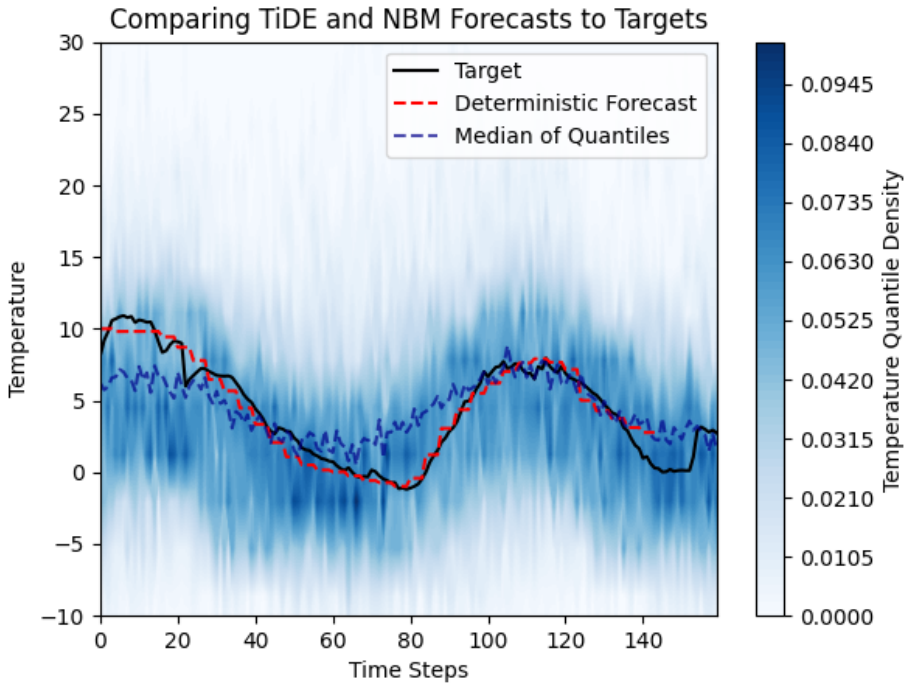


Figure 5.4: An example of TiDE's probabilistic forecasting abilities. We notice a wide spread (vertically) in the probability distribution, indicating an uncertain model.

Figure 5.5 illustrates a confident forecast (one with little vertical spread) and Figure 5.6 illustrates a trimodal probability distribution between time steps two and six. We see that our deterministic forecast is consistent with one of three clusters of quantiles, but is substantially off from the target data. Our median is in the center cluster, and our third cluster contains the target data. This illustrates the importance of probabilistic forecasts in understandability and risk assessment. Should we have a deterministic forecast that takes the middle, we may falsely that assume the distribution is nearly Gaussian and be surprised at the frequency of large errors in certain states. Note that Figure 5.6 is not representative of the deterministic forecast; it is used as an example for our former statement.

5.6 Conclusions

We presented Probabilistic Parallel Time Networks (PTNs), a hypermodel that approximates a probability distribution for each of multiple time steps. By searching an architecture and a hyperparameter space, our PTNs are designed to be effective, transferrable, and robust. PTNs are designed to handle multi-modal input for additional context. The inputs are static data, historical temporal data, and a series of forecasts from a deterministic forecasting model (the

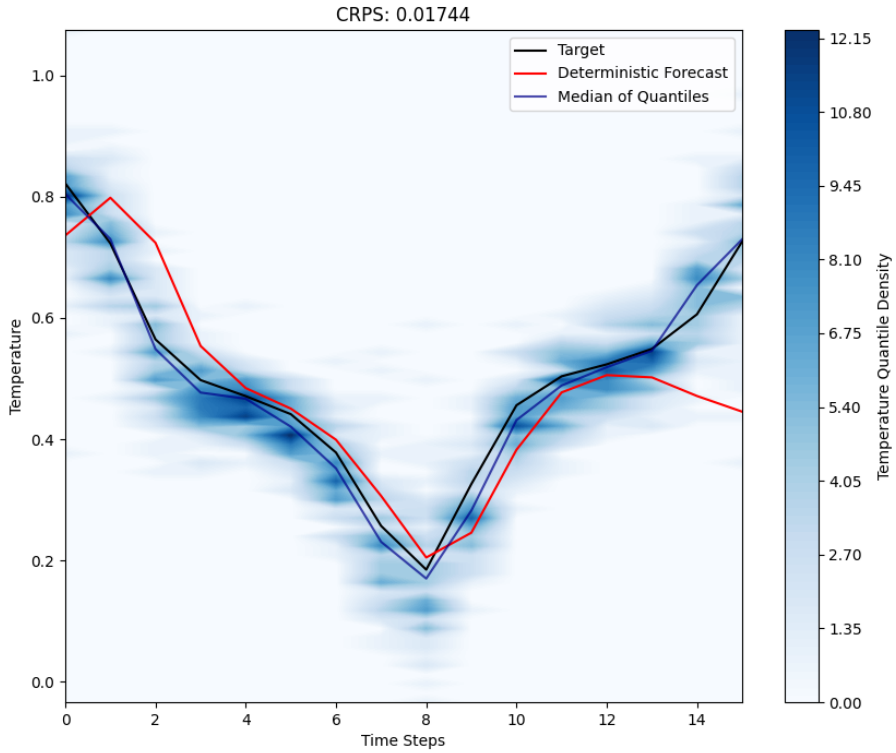


Figure 5.5: A confident prediction by PTN for the chaotic Lorenz system. A confident prediction has less spread in its quantiles.

parallel time aspect).

We illustrated how the PTN's probabilistic forecasts are more interpretable and informative than deterministic forecasts in our case studies of weather forecasting in microclimates and difficult systems such as the chaotic Lorenz system. In real life applications such as economic preparations and weather forecasting, this understanding model uncertainty is critical for the quality of human life. Additionally, these use cases demonstrate PTN's effectiveness and robustness.

Future research may explore the application of PTNs to other complex systems such as forecasting a double pendulum forecasting, electricity demand forecasting, epidemic predictions, full general relativity, and other environmental sciences. Furthermore, forecasts that often go to one extreme or another greatly benefit from probabilistic forecasts, such as volatile financial markets, green energy reliant on wind speed, flooding in certain terrains, and election forecasting in polarized political climates. Additionally, other combinations of NAS and hyperparameter optimizations may be explored to optimize the hypermodel exploration.

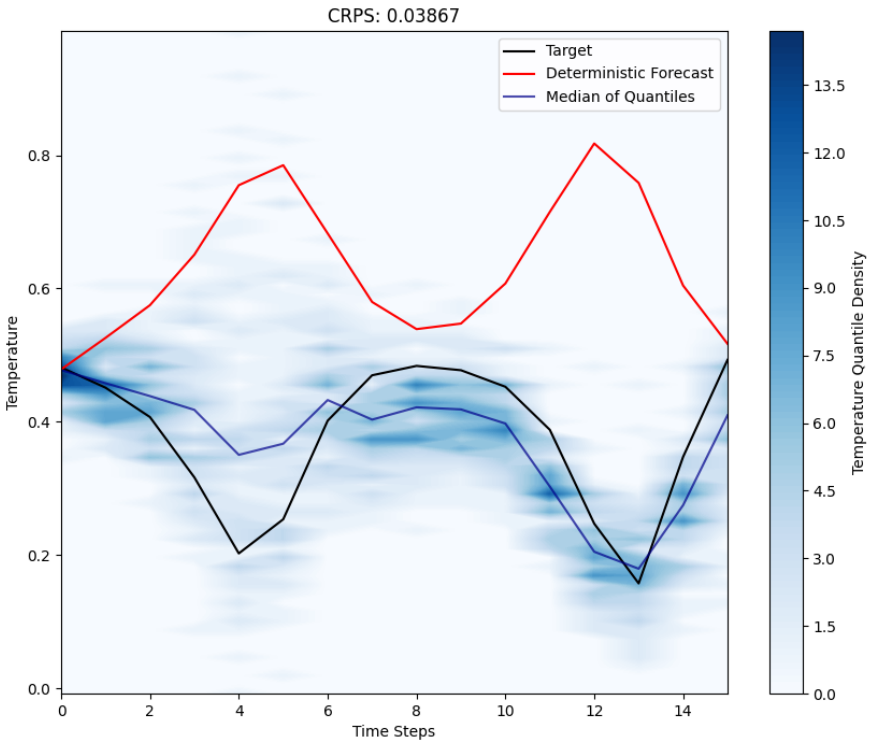


Figure 5.6: Another forecast for the Lorenz system. We can see that the model is less confident in this forecast than in Figure 5.5 but predicts that the chaos will take one of three routes between time steps two and six.

Chapter 6

CONCLUSIONS AND PLANNED WORK

The research outlined in this proposal addresses fundamental challenges in ensuring fairness, robustness, and interpretability in machine learning systems. By introducing tools such as the Objective Fairness Index (OFI), Metric Alignment Trial for Checking Homogeneity (MATCH) test, Cross-Prior Smoothing (CPS), Conditional HydraGAN, and Probabilistic Parallel Time Networks (PTN), we aim to create a strong framework for assessing bias. These methodologies not only enhance the precision of classification metrics and create novel neural frameworks, but also offer a means to align machine learning practices with legal and ethical standards. Through applications ranging from healthcare to chaotic systems forecasting, this research demonstrates a clear trajectory toward building accountable, equitable, and interpretable AI systems. The journey forward involves extending these contributions to broader scenarios, fostering a deeper understanding of AI fairness under diverse practical constraints.

6.1 Summary

In this proposal, we present our research on the Objective Fairness Index (OFI), critical classification metric analyses such as Cross-Prior Score (CPS) and Metric Alignment Trial for Checking Homogeneity (MATCH) test, and the bias mitigation frameworks Conditional HydraGAN and Probabilistic Parallel Time Networks (PTN). OFI is a fairness metric based on international legal precedents; OFI evaluates the fairness of a model by comparing the model's benefit on a given protected group to the benefit of another group. CPS mitigates sample-size-induced bias of classification metrics while MATCH assesses the probability of obtaining some score given the group size. Conditional HydraGAN is a multi-objective generative adversarial network that generates synthetic data to mitigate bias. PTN is a multi-modal hypermodel that approximates a probability distribution for each of multiple time steps. We illustrated how the PTN's probabilistic forecasts are more interpretable and informative than deterministic forecasts in our case studies of weather forecasting in microclimates and difficult systems such as the chaotic Lorenz

system. In real life applications such as economic preparations and weather forecasting, this understanding model uncertainty is critical for the quality of human life. Additionally, these use cases demonstrate PTN's effectiveness and robustness.

6.2 Planned Work

We plan to explore the generalization of OFI to multiclassification, regression, and probabilistic forecasting (enabling PTN evaluations with OFI). Furthermore, properties from social choice theory will be applied to OFI. We will continue investigating critical analyses of classification metrics using the MATCH Test by investigating better approximations for JRMs. Additionally, we will continue researching the PTN by conducting ablation studies and applying it to more datasets, including other complex systems.

6.3 List of Publications

To date, I have been the primary author for published five papers and currently have another one under review at AISTATS. Also, Dr. Assefaw and I are preparing other submissions: a journal submission that integrates OFI with social choice theory, and a submission for PTN with more detailed analyses. Other planned work includes generalizing OFI to multiclassification, regression, and probabilistic forecasting, and improving the beta-distribution approximation of Joint-Ratio Metrics. All code and data are available upon request. The list of my publications is as follows:

1. **Facets of Disparate Impact: Evaluating Legally Consistent Bias in Machine Learning**
   Jarren Briscoe, Assefaw Gebremedhin
   Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)
   DOI: 10.1145/3627673.3679925.
2. **Adversarial Creation of a Smart Home Testbed for Novelty Detection**
   Jarren Briscoe, Assefaw Gebremedhin, Lawrence B. Holder, Diane J. Cook
   AAAI Spring Symposium on Designing AI for Open Worlds, 2022.
3. **Reducing Sample Selection Bias in Clinical Data through Generation of Multi-Objective Synthetic Data**

Jarren Briscoe, Chance DeSmet, Katherine Wuestney, Assefaw Gebremedhin, Roschelle Fritz, Diane J. Cook

Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS 2024).

4. **Exploring Geriatric Clinical Data and Mitigating Bias with Multi-Objective Synthetic Data Generation for Equitable Health Predictions**

   Jarren Briscoe, Chance DeSmet, Katherine Wuestney, Assefaw Gebremedhin, Roschelle Fritz, Diane J. Cook

   Journal of Biomedical Engineering and Biosciences (JBEB), 2024.

5. **Specialized Neural Network Pruning for Boolean Abstractions**

   Jarren Briscoe, Brian Rague, Kyle Feuz, Robert Ball

   Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KEOD 2021)

   DOI: 10.5220/0010657800003064.

6.4 Research Timeline

This section presents an overview of my research timeline, highlighting the progression of tasks, roles, and key achievements across different semesters. The tables below are structured chronologically and convey the past activities, ongoing projects, and future objectives within the research framework of my doctorate experience. The current and previous semesters are detailed in Table 6.1, while the future semesters are outlined in Table 6.2.

Table 6.1: Research Activities by Semester

Semester	Position	Research
Before Fall 2021	RA/Industry	Published "Specialized Neural Network Pruning for Boolean Abstractions" to KEOD.
Fall 2021	RA	Conducted literature review on neural networks, GANs, and fairness. Created CWGAN for synthetic data generation and a novel anomaly detector for network traffic.
Spring 2022	RA	Researched defining bias and nuances of bias. Presented CWGAN at the AAAI Spring Symposium.
Summer 2022	Internship	Began internship with METER. Researched temporal geospatial bias of microclimates and improved CWGAN for synthetic data generation.
Fall 2022	GAANN Fellow	Presented at Science In Our Valley seminar on bias and fairness in ML. Continued work on synthetic data generation, defining bias, and probabilistic forecasting. Passed Qualification Exam with a novel definition of bias.
Spring 2023	GAANN Fellow	Researched time-series similarity measures, laying the framework for comparing them with initial experiments. Created the OFI and conducted preliminary analyses of classification metric distributions. Developed the initial Conditional HydraGAN.
Summer 2023	Internship	Researched neural architectures and learning algorithms. Continued work on microclimate forecasting and drafted the "Probabilistic Parallel Time Network" (PTN) paper.
Fall 2023	GAANN Fellow	Generalized the PTN to dynamic systems, including the chaotic Lorenz system. Mentored Serena Peterson to compare time-series measures. Improved the legal motivation for OFI.
Spring 2024	GAANN Fellow	Improved Conditional HydraGAN and proposed desirable properties of classification metrics.
Summer 2024	GAANN Fellow	Submitted OFI to CIKM and Conditional HydraGAN to ICBES. Finalized the PTN architecture.
Fall 2024	GAANN Fellow	Submitted MATCH and CPS to AISTATS 2025. Presented Conditional HydraGAN at ICBES and OFI at CIKM. Published in Journal of Biomedical Engineering and Biosciences. Take Preliminary Exam.

Table 6.2: Research Plans by Semester

Future Semester	Position	Research Plans
Spring 2025	GAANN Fellow	Submit a journal paper relating classification metrics to social choice theory, and improving desirable properties of bias metrics. Submit PTN. Resolve any AISTATS reviewer concerns.
Summer 2025	GAANN Fellow	Improve, submit, and present papers as necessary. Complete dissertation.
End of Summer 2025	GAANN Fellow	Defend dissertation.

REFERENCES

Aghbalou, A., Sabourin, A., and Portier, F. (2024). Sharp error bounds for imbalanced classification: how many examples in the minority class? In *International Conference on Artificial Intelligence and Statistics*, pages 838–846. PMLR.

Amarasinghe, K., Rodolfa, K. T., Lamba, H., and Ghani, R. (2023). Explainable machine learning for public policy: Use cases, gaps, and research directions. *Data & Policy*, 5:e5.

Amazon (2021). Amazon ai fairness and explainability whitepaper.

Aminikhanghahi, S. and Cook, D. J. (2019). Enhancing activity recognition using CPD-based activity segmentation. *Pervasive and Mobile Computing*, 53(75-89).

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016a). Machine bias. ProPublica. Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016b). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*. Accessed: 2024-10-08.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein gan.

Bender, G., Kindermans, P.-J., Zoph, B., Vasudevan, V., and Le, Q. (2018). Understanding and simplifying one-shot architecture search. In *International conference on machine learning*, pages 550–559. PMLR.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).

Berry, A. C. (1941). The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the American Mathematical Society*, 49(1):122–136.

Biddle, D. (2017). *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Routledge.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., et al. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(2):e1484.

Bojesomo, A., Al-Marzouqi, H., and Liatsis, P. (2021). Spatiotemporal vision transformer for short time weather forecasting. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 5741–5746.

Brennan, T., Dieterich, W., and Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and behavior*, 36(1):21–40.

Briscoe, J., DeSmet, C., Wuestney, K., Gebremedhin, A., Fritz, R., and Cook, D. J. (2024). Reducing sample selection bias in clinical data through generation of multi-objective synthetic data. In *Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'24)*.

Briscoe, J. and Gebremedhin, A. (2024). Facets of disparate impact: Evaluating legally consistent bias in machine learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 3637–3641, New York, NY, USA. Association for Computing Machinery.

Briscoe, J., Gebremedhin, A., Holder, L. B., and Cook, D. J. (2022). Adversarial creation of a smart home testbed for novelty detection. In *AAAI Spring Symposium on Designing AI for Open Worlds*.

Briscoe, J., Rague, B., Feuz, K., and Ball, R. (2021). Specialized neural network pruning for boolean abstractions. volume 2: KEOD of *IC3K 2021*, pages 178–185. Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management.

Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56:1 – 38.

Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13.

Chicco, D. and Jurman, G. (2023). The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification. *BioData Mining*, 16(1):4.

Chopin, I. and Germaine, C. (2017). A comparative analysis of non-discrimination law in europe: Section 3.1 genuine and determining occupational requirements. Research Report DS-05-17-172-EN-N, European Commission, Directorate-General for Justice and Consumers, Brussels.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Cornacchia, G., Anelli, V. W., Narducci, F., Ragone, A., and Di Sciascio, E. (2023). Counterfactual reasoning for bias evaluation and detection in a fairness under unawareness setting. 372.

Craven, J. P., Rudack, D. E., and Shafer, P. E. (2020). National blend of models: a statistically post-processed multi-model ensemble. *Journal of Operational Meteorology*, 8(1).

Dai, Q., Li, H., Wu, P., Dong, Z., Zhou, X.-H., Zhang, R., Zhang, R., and Sun, J. (2022). A generalized doubly robust learning framework for debiasing post-click conversion rate prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 252–262, New York, NY, USA. Association for Computing Machinery.

Damak, K., Khenissi, S., and Nasraoui, O. (2022). Debiasing the cloze task in sequential recommendation with bidirectional transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 273–282, New York, NY, USA. Association for Computing Machinery.

Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., and Yu, R. (2023). Long-term forecasting with tide: Time-series dense encoder.

Das, S., Donini, M., Gelman, J., Haas, K., Hardt, M., Katzman, J., Kenthapadi, K., Larroy, P., Yilmaz, P., and Zafar, B. (2021a). Fairness measures for machine learning in finance.

Das, S., Donini, M., Gelman, J., Haas, K., Hardt, M., Katzman, J., Kenthapadi, K., Larroy, P., Yilmaz, P., and Zafar, M. B. (2021b). Fairness measures for machine learning in finance. *The Journal of Financial Data Science*.

Dennis J. Aigner, M. d. A. and Wiles, J. (2024). Statistical approaches for assessing disparate impact in fair housing cases. *Statistics and Public Policy*, 11(1):2263038.

DeSmet, C. and Cook, D. (2024). Hydragan: A cooperative agent model for multi-objective data generation. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Diehl, A., Pelorosso, L., Delrieux, C., Saulo, C., Ruiz, J., Gröller, M. E., and Bruckner, S. (2015). Visual analysis of spatio-temporal data: Applications in weather forecasting. In *Computer Graphics Forum*, volume 34, pages 381–390. Wiley Online Library.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(4):1–36.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580.

Eisenstein, J. (2019). *Introduction to Natural Language Processing*. MIT Press.

Esseen, C.-G. (1942). On the liapunoff limit of error in the theory of probability. *Arkiv för Matematik, Astronomi och Fysik*, A28(9):1–19.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 259–268, New York, NY, USA. Association for Computing Machinery.

Feng, J., Gossmann, A., Pirracchio, R., Petrick, N., Pennello, G. A., and Sahiner, B. (2024). Is this model reliable for everyone? testing for strong calibration. In *International Conference on Artificial Intelligence and Statistics*, pages 181–189. PMLR.

Freund, J. E., Miller, I., and Miller, M. (2014). *John E. Freund's Mathematical Statistics with Applications*. Pearson Education Limited, 8 edition. pp. 193.

Gao, J., Han, S., Zhu, H., Yang, S., Jiang, Y., Xu, J., and Zheng, B. (2023). Rec4ad: A free lunch to mitigate sample selection bias for ads ctr prediction in taobao. In *Proceedings of the*

32nd ACM International Conference on Information and Knowledge Management, CIKM '23, page 4574–4580, New York, NY, USA. Association for Computing Machinery.

Gastwirth, J. L. and Miao, W. (2009). Formal statistical analysis of the data in disparate impact cases provides sounder inferences than the U. S. government's 'four-fifths' rule: an examination of the statistical evidence in Ricci v. DeStefano. *Law, Probability and Risk*, 8(2):171–191.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.

Goutte, C. and Gaussier, E. (2005). A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer.

Griggs v. Duke Power Co. (1971). 401 u.s. 424. Supreme Court of the United States.

Guess, D. and Sailor, W. (1993). Chaos theory and the study of human behavior: Implications for special education and developmental disabilities. *The Journal of Special Education*, 27(1):16–34.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, 30.

Gursoy, F. and Kakadiaris, I. A. (2022). Equal confusion fairness: Measuring group-based disparities in automated decision systems. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 137–146.

gustav Esseen, C. (1956). A moment inequality with an application to the central limit theorem. *Scandinavian Actuarial Journal*, 1956:160–170.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413.

He, J., Wang, W., Huang, M., Wang, S., and Guan, X. (2021). Bayesian inference under small sample sizes using general noninformative priors. *Mathematics*, 9(21).

Hong, J., Zhu, Z., Yu, S., Wang, Z., Dodge, H. H., and Zhou, J. (2021). Federated adversarial debiasing for fair and transferable representations. KDD '21, page 617–627, New York, NY, USA. Association for Computing Machinery.

Johnson, N. L., Kemp, A. W., and Kotz, S. (2005). *Univariate Discrete Distributions*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 3rd edition.

Knuth, D. E. (1968). *The Art of Computer Programming, Volume I: Fundamental Algorithms*. Addison-Wesley.

Kwon, S., Kim, S., Lee, S., Kim, J.-Y., An, S., and Kim, K. (2023). Addressing selection bias in computerized adaptive testing: A user-wise aggregate influence function approach. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 4674–4680, New York, NY, USA. Association for Computing Machinery.

Lanczos, C. (1964). A precision approximation of the gamma function. *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis*, 1(1):86–96.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How we analyzed the compas recidivism algorithm. *ProPublica*. Accessed: 2024-10-08.

Le, T. and Deng, A. (2023). The price is right: Removing a/b test bias in a marketplace of expirable goods. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 4681–4687, New York, NY, USA. Association for Computing Machinery.

Lemoine, N. P. (2019). Moving beyond noninformative priors: why and how to choose weakly informative priors in bayesian analyses. *Oikos*, 128(7):912–928.

Liu, H., Simonyan, K., and Yang, Y. (2019). Darts: Differentiable architecture search.

Liu, Y., Sun, Y., Xue, B., Zhang, M., and Yen, G. G. (2020). A survey on evolutionary neural architecture search. *CoRR*, abs/2008.10937.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141.

Loshchilov, I. and Hutter, F. (2017). Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Mangiarotti, S., Peyre, M., Zhang, Y., Huc, M., Roger, F., and Kerr, Y. (2020). Chaos theory applied to the outbreak of covid-19: an ancillary approach to decision making in pandemic context. *Epidemiology & Infection*, 148.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

May, R. M., Goebbert, K. H., Thielen, J. E., Leeman, J. R., Camron, M. D., Bruick, Z., Bruning, E. C., Manser, R. P., Arms, S. C., and Marsh, P. T. (2022). Metpy: A meteorological python library for data analysis and visualization. *Bulletin of the American Meteorological Society*, 103(10):E2273 – E2284.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *CoRR*, abs/1411.1784.

Moore, C., Ferguson, E., and Guerin, P. (2023). Pretrial Risk Assessment on the Ground: Algorithms, Judgments, Meaning, and Policy. *MIT Case Studies in Social and Ethical Responsibilities of Computing*, Summer 2023. https://mit-serc.pubpub.org/pub/czviu6qc.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

OpenAI (2023). Gpt-4: Generative pre-trained transformer 4. Technical report, OpenAI.

Pavlyshenko, B. (2018). Using stacking approaches for machine learning models. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, pages 255–258.

Peizer, D. B. and Pratt, J. W. (1968). A normal approximation for binomial, f, beta, and other common, related tail probabilities, i. *Journal of the American Statistical Association*, 63(324):1416–1456.

ProPublica (2016). Compas analysis github repository. Accessed: 2024-10-08.

Ricci v. DeStefano (2009). 557 u.s. 557. Supreme Court of the United States.

Robin, G. (1984). Grandes valeurs de la fonction somme des diviseurs et hypothése de riemann. *J. Math. Pures Appl.*, 63:187–213.

Rudner, T. G., Zhang, Y. S., Wilson, A. G., and Kempe, J. (2024). Mind the gap: Improving robustness to subpopulation shifts with group-aware priors. In *International Conference on Artificial Intelligence and Statistics*, pages 127–135. PMLR.

Russo, F. and Tonia, F. (2023). Causal discovery and knowledge injection for contestable neural networks. In *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023)*, volume 372 of *Frontiers in Artificial Intelligence and Applications*. IOS Press.

Shahi, S., Fenton, F. H., and Cherry, E. M. (2022). Prediction of chaotic time series using recurrent neural networks and reservoir computing techniques: A comparative study. *Machine Learning with Applications*, 8:100300.

Shevtsova, I. G. (2010). An improvement of convergence rate estimates in the lyapunov theorem. In *Doklady Mathematics*, volume 82.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Takeuchi, I., Le, Q., Sears, T., and Smola, A. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7:1231–1264.

U.S. Equal Employment Opportunity Commission (1978). Uniform guidelines on employee selection procedures. *Code of Federal Regulations, Title 29, Section 1607.4*. Available at: https://www.law.cornell.edu/cfr/text/29/1607.4.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE.

Verma, V. K., Brahma, D., and Rai, P. (2020). Meta-learning for generalized zero-shot learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6062–6069.

Wang, Z., Huang, N., Sun, F., Ren, P., Chen, Z., Luo, H., de Rijke, M., and Ren, Z. (2022). Debiasing learning for membership inference attacks against recommender systems. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 1959–1968, New York, NY, USA. Association for Computing Machinery.

Ward, L. M. and West, R. L. (1998). Modeling human chaotic behavior: Nonlinear forecasting analysis of logistic iteration. *Nonlinear Dynamics, Psychology, and Life Sciences*, 2:261–282.

White, C., Neiswanger, W., and Savani, Y. (2021). Bananas: Bayesian optimization with neural architectures for neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10293–10301.

Ye, K., Walpole, R. E., Myers, R. H., and Myers, S. L. (2024). *Probability and Statistics for Engineers and Scientists*. Pearson, updated 9th edition. Section 6.5.

Yuan, T., Zhu, J., Ren, K., Wang, W., Wang, X., and Li, X. (2022). Neural network driven by space-time partial differential equation for predicting sea surface temperature. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 656–665.

Zang, X., Iqbal, S., Zhu, Y., Liu, X., and Zhao, J. (2016). Applications of chaotic dynamics in robotics. *International Journal of Advanced Robotic Systems*, 13(2):60.

Zhang, J. M., Harman, M., Ma, L., and Liu, Y. (2020). Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*.

Zhang, W., Wang, Z., Kim, J., Cheng, C., Oommen, T., Ravikumar, P., and Weiss, J. (2023). Individual fairness under uncertainty. 372.

Zhou, H., Yang, M., Wang, J., and Pan, W. (2019). Bayesnas: A bayesian approach for neural architecture search. In *International conference on machine learning*, pages 7603–7613. PMLR.

Zoph, B. and Le, Q. V. (2016). Neural architecture search with reinforcement learning. *CoRR*, abs/1611.01578.