Optimizing Medical Costs and Resource Utilization Using Causal Inference and Explaining the Model's Predictions

A Project Report

submitted by

TANISHA JAIN

MT21180

in partial fulfilment of the requirements

for the award of the degree of

MASTER OF TECHNOLOGY

ELECTRONICS AND COMMUNICATION ENGINEERING

INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

July, 2023

THESIS CERTIFICATE

This is to certify that the thesis titled **Optimizing Medical Costs and Resource Utilization Using Causal Inference and Explaining the Model's Predictions**, submitted by **Tanisha Jain**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Master of Technology**, is a bonafide record of the research work done by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Ranjitha Prasad
Thesis Supervisor
Assistant Professor
Dept. of Electronics and Communication
IIIT Delhi, 110020

Place: New Delhi

July 2023

ACKNOWLEDGEMENTS

I would like to extend my heartfelt gratitude to Dr. Ranjitha Prasad for her unwavering guidance, exceptional support, and remarkable supervision throughout the course of this research work. Her expertise and insightful feedback have been invaluable in shaping the direction and quality of this thesis. I am profoundly grateful for her dedication and commitment to my academic development. Additionally, I would like to express my deepest appreciation to my family and friends for their unwavering care, unconditional love, and constant blessings throughout this journey.Their encouragement and belief in me have been instrumental in my success.

ABSTRACT

KEYWORDS: Medical Cost ; Counterfactual Inference ; Explainability ; Deep Neural Network

Machine learning (ML) models that accurately predict treatment effects and related healthcare costs can bring significant efficiencies in the healthcare industry. These models could help reduce fatalities resulting from incorrect treatment allocations and contribute to cost-effective healthcare delivery, which is crucial for both developed and developing nations. However, existing literature does not provide any comprehensive framework that effectively estimates both treatment effect and the overall medical expenditure while considering individual treatment effects. To address this gap, we propose CFMedNet, a pioneering counterfactual inference framework that jointly estimates treatment effect and medical costs. This novel framework not only predicts the potential impact and costs of a given treatment and its counterfactual but also provides individual treatment effects for both outcomes. However, a considerable challenge in the adoption of such ML models in healthcare is their perceived 'black box' nature due to limited transparency in decision-making processes. Since medical professionals bear the responsibility for their decisions, it's crucial to have Explainable AI (XAI) models, especially in sensitive domains like healthcare. As an innovative contribution, we introduce a post-hoc explainer, GMM-LIME, specifically designed for multi-output causal inference based counterfactual neural networks. This explainer offers crucial explanations and interpretations of our proposed model, thereby improving its transparency and applicability. This dual contribution of a comprehensive estimation framework and in-depth explanatory tools, holds great potential to significantly progress personalized healthcare, balancing economic efficiency with treatment efficacy. Our work represents an integration of Causal Inference, Deep Learning, and XAI, with results obtained from a semi-synthetic dataset.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Causal Inference

Causal inference (Neal (2020) , Pearl (2010)), as its name suggests, involves studying the underlying cause of a particular effect or outcome. It entails analyzing interventions or treatments to determine if they have indeed caused the observed effect. Additionally, it explores counterfactual scenarios (Höfler (2005), Molnar (2022)), exploring what would happen in an alternative world with a different intervention. This field is highly significant in various domains, and here are some new examples:

- **Healthcare:** (Shi and Norgeot (2022), Prosperi *et al.* (2020))
  For assessing the effectiveness of different COVID-19 vaccines, by conducting rigorous studies and comparing outcomes between individuals who have been vaccinated and those who have not (control group), researchers can establish a causal relationship between the vaccines and their impact on preventing COVID-19. This enables them to determine which vaccines are more effective in providing protection against the virus and informing vaccination strategies for population health.

- **Education:** (Kaur *et al.* (2019), Murnane and Willett (2011))
  If researchers are interested in examining how different teaching methods affect student performance, causal inference can assist them. By comparing outcomes between classrooms that use different teaching approaches, researchers can establish a causal relationship between the teaching method and its impact on academic achievement. This can help identify which teaching method is more effective in improving student outcomes and inform best practices for educators.

- **Policy Making:** (Athey (2015), Kreif and DiazOrdaz (2019))
  Let's consider the scenario of policy making to address unemployment rates. Policymakers need to identify the key factors that contribute to high unemployment. By using causal inference techniques to examine the relationships between different variables and unemployment, policymakers can gain insights into the causal factors driving joblessness. This knowledge empowers them to develop targeted interventions and policies that have the potential to effectively reduce unemployment and promote economic growth.

1.1.1 Statistical Notations and Definitions in Causal Inference

Let's introduce some notation for our discussion. The symbol T will be used to represent the random variable for treatment, Y to represent the random variable for the outcome of interest, and X to represent the covariates. It's important to note that our focus will primarily be on situations where T is binary. However, it's worth mentioning that we can extend our methods to accommodate settings where T can take on more than two values or where T is continuous.

In our context, we define the potential outcome (Rubin (2005)), denoted as $Y(t)$, which represents the outcome that would be observed if a particular treatment, denoted as t, is applied. It's important to note that potential outcomes differ from the observed outcome in the sense that we do not observe all potential outcomes for each individual. Instead, we can potentially observe any of the potential outcomes based on the value of the treatment variable, T.

Let's consider the variables for an individual in the population: T_i represents the treatment received by the i^{th} individual, X_i represents the covariates for that individual, and Y_i represents the observed outcome. In this context, the individual treatment effect **(ITE)** (Jin *et al.* (2023), Shalit *et al.* (2017)) is defined as the difference between the potential outcomes under different treatment conditions.

$$\tau_i = Y_i(1) - Y_i(0) \tag{1.1}$$

1.1.2 Fundamental Problem of Causal Inference

In practical real-world scenarios, it is important to acknowledge that only one outcome is observed among the various potential outcomes (Imbens and Rubin (2015)). For example, when comparing the effects of different vaccines, a patient can only receive one vaccine, and thus only one potential outcome is observed based on the assigned treatment. This poses a fundamental problem because if we cannot observe both potential outcomes, namely $Y_i(1)$ and $Y_i(0)$, we cannot directly observe the causal effect, which is the difference between these potential outcomes $(Y_i(1) - Y_i(0))$.

The unobserved potential outcomes are referred to as *counterfactuals* because they represent what would have happened under alternative treatment conditions. In essence, they depict the outcomes that we cannot directly observe or measure. The existence of counterfactuals makes causal inference challenging as we aim to estimate the causal effects of treatments or interventions based on the observed data.

1.1.3 Solution to the Fundamental Problem

Average Treatment Effect

ATE (Imbens (2004), González Ramírez and Kilic (2019)) is calculated by averaging ITEs, which are the differences in outcomes between different treatment conditions for individuals. It provides an estimate of the overall treatment effect on the population.

$$\tau = \mathbf{E}[Y_i(1) - Y_i(0)] = \mathbf{E}[Y(1) - Y(0)] \tag{1.2}$$

By linearity of expectation:

$$\tau = \mathbf{E}[Y(1) - Y(0)] = \mathbf{E}[Y(1)] - \mathbf{E}[Y(0)] \tag{1.3}$$

The associational difference between the two outcomes is defined as follows:

$$\mathbf{E}[Y \mid T = 1] - \mathbf{E}[Y \mid T = 0] \tag{1.4}$$

The ATE (average treatment effect)(1.2) is a causal measure, while the associational difference is a statistical measure. They are not generally equal because association or correlation does not always imply causation. For example,

- It is important to recognize that a high correlation exists between the number of people drowning in swimming pools and the number of films featuring Nicolas Cage each year. However, it would be erroneous and illogical to conclude that Nicolas Cage is responsible for all those deaths. Correlation alone does not imply causation, and it is crucial to exercise caution and consider other relevant factors before attributing a causal relationship.
- In a study, it was observed that individuals who sleep with their shoes on experience severe headaches the following morning. However, it would be erroneous to conclude that wearing shoes in bed directly causes headaches. Upon further examination, it was discovered that people tend to sleep with their boots on when

they are intoxicated or drunk. The high alcohol consumption from the previous night is more likely to be the cause of the headaches, rather than the act of wearing shoes while sleeping. In this scenario, the common factor is drinking at night, which is a confounding factor that influences both the behavior of sleeping with shoes on and the occurrence of morning headaches. This example is depicted in Fig. (1.1).

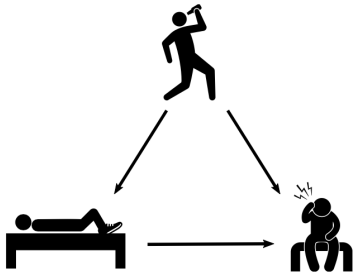

Figure 1.1: In the causal structure, the act of drinking the night before serves as a common cause for both waking up with headaches and sleeping without taking shoes off.

Nevertheless, we can equate (1.3) with (1.4) under certain conditions each of which is explained in subsequent sections.

Ignorability and Exchangeability

The condition of ignorability (Greenland and Robins (2009)) asserts that the potential outcomes are independent of the treatment assignment. In simpler terms, this means that the treatment assignment is completely random, and the potential outcomes are not influenced by the specific treatment received. Mathematically, this condition can be expressed as follows:

$$Y(1), Y(0) \perp T \tag{1.5}$$

This simplifies the average treatment effect (ATE) (1.2) to:

$$\mathbf{E}[Y(1)] - \mathbf{E}[Y(0)] = \mathbf{E}[Y(1) \mid T = 1] - \mathbf{E}[Y(0) \mid T = 0] \tag{1.6}$$

The condition of *ignorability* plays a pivotal role in addressing the issue of confounding variables, which can hinder the equivalence between the associational difference (a statistical quantity) and the average treatment effect (a causal quantity), as discussed earlier. By assuming the condition of ignorability, we can eliminate the influence of confounding variables and establish a more accurate estimation of the causal

effect of the treatment. This condition helps ensure that the treatment assignment is independent of any potential confounders, allowing us to draw valid causal inferences based on the observed data.

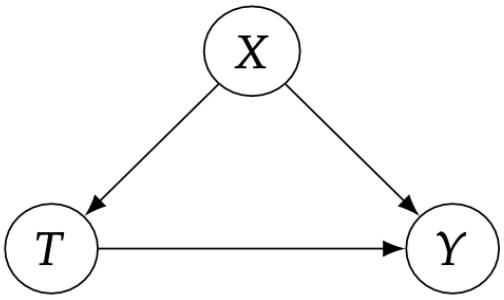

Figure 1.2: The presence of a causal relationship between the variable X and the variable Y is confounded by the variable T.

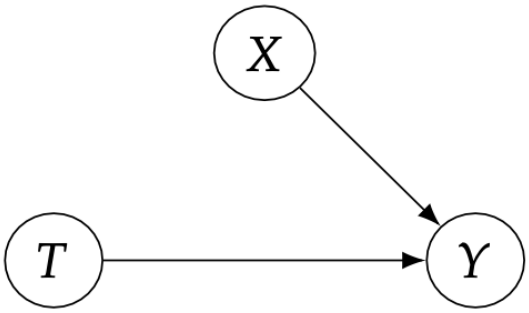

Figure 1.3: In a causal structure where ignorability is upheld, there is no causal link from X to T, indicating the absence of confounding.

The condition of *exchangeability* states that the treatment groups, such as groups A and B, are interchangeable. In other words, if we were to swap the groups, the participants who were originally in treatment group A would experience the same outcomes as the participants in the new treatment group A, and likewise for group B. This condition implies that the two treatment groups are comparable and similar in all aspects, except for the specific treatment received. By assuming exchangeability, we can assume that any differences in outcomes between the treatment groups are solely due to the treatment itself, rather than other factors or characteristics of the groups. This allows for a more valid comparison of treatment effects.

In mathematical sense, the condition of exchangeability can be expressed as follows:

$$\mathbf{E}[Y(1) \mid T = 1] = \mathbf{E}[Y(1) \mid T = 0]$$
$$\mathbf{E}[Y(0) \mid T = 1] = \mathbf{E}[Y(0) \mid T = 0]$$

which implies,

$$\mathbf{E}[Y(1) \mid T = t] = \mathbf{E}[Y(1)] \tag{1.7}$$

$$\mathbf{E}[Y(0) \mid T = t] = \mathbf{E}[Y(0)] \tag{1.8}$$

Equations 1.6, 1.7 and 1.8 evidently imply that the condition of exchangeability and ignorability are mathematically the same.

Conditional Exchangeability and Unconfoundedness

It is not realistic to suppose that treatment groups in observational data are completely interchangeable. In other words, we cannot expect the treatment groups to be identical in all relevant characteristics except for the treatment itself. However, by controlling for key factors through conditioning, we can achieve a form of interchangeability among subgroups. Let X' represent a subset of covariates X conditioned on for unconfoundedness. In this case, the concept of conditional exchangeability or ignorability is defined as follows:

$$(Y(1), Y(0)) \perp T \mid X' \tag{1.9}$$

Conditional exchangeability enables the investigation of the average treatment effect within specific levels of conditioned covariates. It recognizes that while the treatment and potential outcomes may be related without conditioning, they are not associated within those specific levels of X. This concept helps address confounding and provides a framework for analyzing treatment effects while considering the influence of covariates, as:

$$\begin{aligned} \mathbf{E}[Y(1) - Y(0) \mid X] &= \mathbf{E}[Y(1) \mid X] - \mathbf{E}[Y(0) \mid X] \\ &= \mathbf{E}[Y(1) \mid T = 1, X] - \mathbf{E}[Y(0) \mid T = 0, X] \\ &= \mathbf{E}[Y \mid T = 1, X] - \mathbf{E}[Y \mid T = 0, X] \end{aligned} \tag{1.10}$$

We obtain 1.10 through linearity of expectation and conditional exchangeability applied in the same order.

Positivity

Positivity, in the case of binary treatment, requires that every subgroup defined by conditioned covariates has a mixture of treated and control participants. This ensures that all possible treatment options are observed within each subgroup, allowing for a comprehensive analysis of treatment effects across the entire population. The condition of positivity guarantees that the treatment assignment is not overly restricted to specific subgroups, promoting a more balanced and inclusive evaluation of treatment outcomes.Mathematically,

$$0 < P(T = 1 \mid X) < 1 \tag{1.11}$$

No Interference

The notion of no interference indicates that the outcome of every individual is determined only by the treatment provided to that individual and is unaffected by the treatment of someone else. Mathematically,

$$Y_i(t_1, t_2, \ldots, t_{i-1}, t_i, t_{i+1}, \ldots, t_n) = Y_i(t_i) \tag{1.12}$$

Consistency

The assumption that the outcome observed Y is the potential outcome under the observed treatment T is known as consistency. Mathematically,

$$T = t \Rightarrow Y = Y(t) \tag{1.13}$$

which implies

$$Y = Y(T) \tag{1.14}$$

1.1.4 Average Treatment Effect Estimation

Theorem 1.1 (Adjustment Formula) If the assumptions of unconfoundedness, positivity, consistency and no interference hold, the average treatment effect is:-

$$\mathbf{E}[Y(1)Y(0)] = \mathbf{E}_X[\mathbf{E}[Y|T = 1, X] - \mathbf{E}[Y|T = 0, X]] \tag{1.15}$$

Proof.

$$\begin{aligned}\mathbf{E}[Y(1) - Y(0)] &= \mathbf{E}[Y(1)] - \mathbf{E}[Y(0)] \\ &= \mathbf{E}_X[\mathbf{E}[Y(1)|X] - \mathbf{E}[Y(0)|X]] \\ &= \mathbf{E}_X[\mathbf{E}[Y(1)|T = 1, X] - \mathbf{E}[Y(0)|T = 0, X]] \\ &= \mathbf{E}_X[\mathbf{E}[Y|T = 1, X] - \mathbf{E}[Y|T = 0, X]]\end{aligned}$$

□

1.2 Explainable AI

Artificial Intelligence (AI) has become an influential player in multiple sectors of the global economy, and the field of healthcare is no exception. The development of advanced machine learning algorithms has unlocked new possibilities in disease diagnosis, treatment personalization, and healthcare resource management. However, the growing adoption of AI in healthcare comes with a unique set of challenges, the most notable among them being the 'black box' nature of complex models. This obscurity results in a lack of transparency and understanding, which could hinder the trustworthiness of AI solutions, especially in healthcare, where decisions can have life-altering consequences. This concern has spurred the growth of a new sub-field known as Explainable AI (XAI) (Xu *et al.* (2019), Gunning *et al.* (2019), Doran *et al.* (2017), Holzinger (2018)).

XAI refers to methods and techniques in the application of AI, such that the resulting models are understandable and interpretable by human experts. It addresses the transparency issue by revealing the inner workings of AI models and making the results they produce easier to understand. XAI focuses on three primary aspects: interpretability, explainability, and fairness.

Interpretability (Gilpin *et al.* (2018)) refers to the degree to which a human can understand the cause of a decision. In the context of healthcare, this could mean understanding why a particular treatment was recommended by an AI model.

Explainability(Belle and Papantonis (2021), Roscher *et al.* (2020)), on the other hand, refers to the ability to describe in understandable human terms why a model made a certain prediction. For instance, if an AI model predicted a high risk of diabetes for

a patient, explainability would involve outlining the patient characteristics and patterns that led to this conclusion.

Fairness (Arrieta *et al.* (2020), Hagras (2018)) in XAI concerns minimizing bias and ensuring equal treatment across different groups in the data. This can involve steps to remove or account for bias in the data, as well as designing models that treat similar individuals similarly.

In healthcare, the role of XAI is pivotal. It empowers clinicians to comprehend AI outcomes, reinforcing their decision-making processes. Furthermore, it fosters trust in patients, who can better understand and accept AI-driven medical decisions. XAI also contributes towards regulatory compliance and medical ethics by ensuring fairness and transparency.

From enhancing patient trust to improving model accuracy and efficiency, XAI is revolutionizing the use of AI in healthcare. By demystifying AI models and providing clear, understandable explanations of their decisions, XAI paves the way for AI to fulfill its potential in delivering more personalized, effective, and accountable healthcare.

1.2.1 Post-hoc XAI

Post-hoc explainable AI refers to the methods and techniques used to explain the decisions or predictions made by AI models after they have been trained. These methods aim to provide retrospective explanations by examining the model's internal workings, such as feature importance, or by generating additional explanations based on the model's behavior for a given input.

Some post-hoc methods, like LIME (Ribeiro *et al.* (2016)), determine the importance of a feature by perturbing real samples, watching how the ML model's output changes, and then building a local simple model that closely resembles the original model's behavior close to the real samples. These approaches have one drawback: they generate surrounding instances by randomly perturbing feature values, without taking into account the local distribution of features or the density of nearby class labels (Guidotti *et al.* (2018)). These methods support the predictions of opaque models through feature attribution.

As we delve deeper into this thesis, we will explore the intricacies of XAI, its

methodologies, and its impact on the healthcare industry. The following chapters aim to provide a comprehensive understanding of XAI's potential to transform healthcare and the challenges that lie ahead.

1.3 Problem Definition

1.3.1 Treatment Effect and Medical Cost Estimation

The escalating healthcare costs and the necessity for effective treatment present significant challenges within the healthcare system. A solution lies in the precise estimation of treatment effects and corresponding medical costs, which emerges as a critical research direction. Current methods for these estimations often have limitations, such as bias, and inadequate consideration of patient heterogeneity, which is vital for personalized care. The estimation of treatment impact is pivotal for patient outcomes, healthcare systems, policymakers, and payers. It enables identifying effective medical interventions, enhancing clinical decision-making, designing evidence-based treatment protocols, and improving health outcomes. Simultaneously, predicting healthcare costs is strategically important for healthcare providers, payers, and patients. It aids in informed decision-making, managing resources, planning services, policy formulation, risk adjustment, and budgeting. However, the complexities of healthcare data pose challenges in cost estimation, demanding innovative, accurate, and interpretable methodologies. This research is thus a pressing imperative, impacting healthcare efficiency, care quality, and patient experience.

1.3.2 Optimal Treatment Selection

The shift towards personalized patient care emphasizes optimal treatment selection, a process of identifying the most suitable treatment for a patient based on their specific health characteristics. However, due to the intricacy of health variables, the dynamic nature of diseases, and the vast spectrum of treatment options, this process presents a complex challenge. Yet, its importance is undeniable. Optimal treatment selection improves health outcomes, reduces side effects, enhances the quality of life, and boosts treatment adherence for patients. It also enhances clinical decision-making and healthcare delivery for providers by minimizing trial-and-error prescribing, reducing adverse

drug reactions, and helping navigate the treatment landscape. At a system level, it addresses major healthcare challenges such as rising costs, uneven care quality, and inefficient resource allocation, contributing to the sustainability of the healthcare system. Despite its significance, optimal treatment selection is fraught with difficulties, necessitating sophisticated decision-making tools and approaches. The development of innovative methodologies to support healthcare professionals in making complex decisions is a compelling need, marking a critical direction for current research in medical science.

1.3.3 Explainable artificial intelligence(XAI)

This thesis explores the importance of explainable AI in the implementation of deep learning models for estimating treatment effects and medical costs through causal inference in healthcare research. Deep learning techniques offer accurate predictions, but the black-box nature of these models hinders transparency and trust. Explainable AI methods enhance interpretability, allowing researchers to understand the factors influencing treatment effects and costs, identify biases, and communicate findings effectively. Interpretable models ensure accountability and trust by providing justifications for predictions, aiding decision-making by clinicians and patients. Furthermore, explainable AI mitigates biases and unintended consequences by assessing fairness, ethics, and safety. By integrating explainable AI techniques, researchers and practitioners can leverage the power of deep learning while maintaining ethical standards and improving patient outcomes in healthcare decision-making.

CHAPTER 2

RELATED WORK

2.1 Literature Review

2.1.1 Causal Inference

Causal inference plays a vital role in multiple domains, including healthcare, social sciences, and economics, as it involves the fundamental task of understanding cause-and-effect relationships. Traditional causal inference methods often rely on assumptions and limited flexibility. However, deep learning-based approaches have emerged as promising alternatives, offering the potential to capture complex causal relationships without stringent assumptions. In this literature review, we will focus on three notable deep learning-based causal inference models. Deep learning-based causal inference models, such as TARNet, Dragonnet and CFRNet, offer promising approaches to estimate treatment effects and tackle complex causal inference problems. These models leverage the flexibility and power of deep neural networks to capture intricate relationships and handle challenges like unobserved confounders.

TARNet:

The Treatment Agnostic Regression Network (TARNet) proposed by Shalit *et al.* (2017) addresses the challenge of estimating individual treatment effects. TARNet leverages the power of deep neural networks to learn a treatment effect model that is invariant to the specific treatment assignment mechanism. It achieves this by using a novel "doubly robust" estimator that combines an outcome regression and a treatment propensity model. By estimating both models simultaneously, TARNet can accurately estimate the causal effect of a treatment on an individual's outcome. The authors also provide theoretical insights by establishing generalization bounds, ensuring the model's performance on unseen data. These bounds give a measure of confidence in the estimated treatment effects. TARNet's flexibility, scalability, and generalizability make it a valu-

able tool for causal inference tasks, particularly in domains where treatment assignment mechanisms may vary.

Dragonnet:

Dragonnet, proposed by Shi *et al.* (2019), is another deep learning-based causal inference model that adapts neural networks for estimating treatment effects. Dragonnet focuses on the scenario where confounding variables, which affect both treatment assignment and outcome, are unobserved or partially observed. It addresses this challenge by leveraging probabilistic modeling and the concept of doubly robust estimation. Dragonnet uses a variational autoencoder (VAE) to learn a representation of the confounders and combines this with a neural network-based outcome regression and treatment propensity model. By jointly optimizing these models, Dragonnet can estimate treatment effects robustly, even in the presence of unobserved or partially observed confounders. The authors demonstrate the effectiveness of Dragonnet through experiments on real-world datasets.

Counterfactual Regression Network (CFRNet):

Shalit *et al.* (2017) proposed CFRNet framework for estimation of individual treatment effect in balanced representations. We discuss the approach briefly since our work is an extension of this framework and all the assumptions involved hold true for our work as well.

The notations used and assumptions involved are defined as follows:

- Space of covariates $\mathcal{X}$, is a subset of d-dimensional real space $\mathbb{R}^d$ i.e. $\mathcal{X} \subset \mathbb{R}^d$
- The outcome space $\mathcal{Y} \subset \mathbb{R}^d$
- The treatment *a* is binary in nature $\{0, 1\}$
- It is assumed that strong ignorability (1.9) and positivity (1.11) holds.
- The covariates $\mathcal{X}$ are mapped to a representation space $\mathcal{R}$ using the function $\phi : \mathcal{X} \to \mathcal{R}$. It is assumed that ϕ is a one-to-one, twice differentiable function.
- The hypothesis is defined as $h : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$ while the loss function is $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathcal{R}_+$

Various definitions involved in the paper are as follows:

- The treatment effect for an instance *x* is obtained using:

$$\tau(x) = \mathbf{E}\left[Y_1 - Y_0 | (x)\right] \tag{2.1}$$

- The hypothesis is proposed as $f : \mathcal{X} \times \{0, 1\} \rightarrow \mathcal{Y}$ such that $f(x, t) = h(\phi(x), a)$.
- The estimated treatment effect of hypothesis *f* for an instance *x* is :

$$\hat{\tau}(x) = f(x, 1) - f(x, 0) \tag{2.2}$$

- The expectation of square of difference between estimated and actual treatment effect , also called as expected Precision in Estimation of Heterogeneous Effect (PEHE) (Hill (2011)) loss is calculated as:

$$\epsilon_{PEHE} = \int_{\mathcal{X}} (\hat{\tau}_f(x) - \tau(x))^2 p(x) dx \tag{2.3}$$

- In order to compute the distance between treatment and control distributions, a probability distribution metric called as Integral Probability Metric (IPM) is used which is defined as :

$$IPM_G = \sup_{g \in G} \left| \int_S g(s)(p(s) - q(s)) ds \right| \tag{2.4}$$

  where p and q are two probability density functions defined over $\mathcal{S} \subset \mathbb{R}^d$ and G is a family of functions such that $g : \mathcal{X} \rightarrow \mathbb{R}$.

The primary objective of CFRNet is to identify a representation $\phi : \mathcal{X} \rightarrow \mathcal{R}$ and hypothesis $h : \mathcal{R} \times \{0, 1\} \rightarrow \mathcal{Y}$ that minimises the PEHE loss ϵ_{PEHE}. In order to achieve this, Shalit *et al.* (2017) utilized a deep learning architecture to simultaneously model $\phi(x)$ and $h(\phi(x), a)$. The covariates $\mathcal{X}$ are transformed to representation space, $\phi(x)$ which then act as an input to the hypothesis layer segmented into two branches, h_1 and h_0 based on whether treatment assigned is 1 or 0. Also, the difference between the treatment and control distribution is minimized using an IPM term. The architecture is represented in Fig. (2.1).

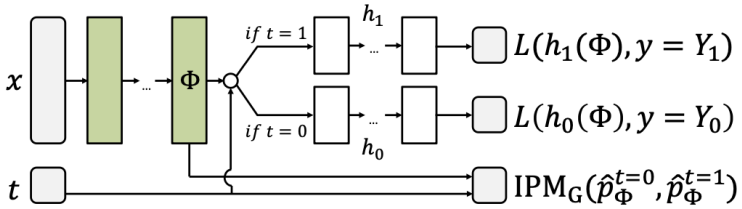

Figure 2.1: CFRNet Architecture

The optimal solution is obtained by minimizing loss function as depicted in equation (2.5) using stochastic gradient descent where the error is backpropagated via both hypothesis and representation networks.

$$\mathcal{L} = \frac{\beta}{n} \sum_{i=1}^{n} s_i L(h(\phi(x_i), a_i), y_i) + \lambda \cdot \mathcal{R}(h) + \gamma \cdot IPM_G(\{\phi(x_i)\}_{i:a_i=0}, \{\phi(x_i)\}_{i:a_i=1}) \quad (2.5)$$

where $s_i = \frac{a_i}{2v} + \frac{1-a_i}{2(1-v)}$,$v = \frac{1}{n}\sum_{i=1}^{n} a_i$ and $\mathcal{R}(\cdot)$ is a model complexity term. Also v refers to the proportion of treated instances $(a_i = 1)$, $\{\beta, \lambda, \gamma\}$ are the hyperparameters and $L(.,.)$ is squared error loss. The IPM term utilized Maximum Mean Discrepancy (Sriperumbudur et al.Sriperumbudur *et al.* (2012)) distance metric.

2.1.2 Medical Cost Prediction

Machine learning techniques have seen extensive use in healthcare, spanning from disease prediction to treatment recommendations and healthcare cost forecasting. Predicting healthcare costs using deep learning has been a burgeoning field in recent years, capitalizing on the rise in available medical data and the surge in computational resources. These predictive models are crucial in forecasting individual and population health expenditures, informing healthcare policies, and optimizing resource allocation.

Ma *et al.* (2019) presented a deep learning framework based on the LSTM architecture to predict individual patient healthcare costs. Their work differed from previous studies by capturing temporal relationships in patients' medical histories. Using administrative claims data, the LSTM model significantly outperformed traditional regression methods and was particularly adept at identifying high-cost patients. The authors Sun *et al.* (2020) introduced CostNet, a novel deep learning framework designed specifically for healthcare cost prediction. CostNet leverages attention mechanisms in its architecture, allowing it to prioritize relevant medical events when predicting costs. They demonstrated the model's superior performance on two large healthcare datasets, outperforming other deep learning models and traditional machine learning techniques. Although Wang *et al.* (2020) study was not specifically focused on cost prediction, it featured an innovative use of Convolutional Neural Networks (CNNs) for chronic disease diagnosis, which is inherently linked to healthcare costs. This work demonstrates the flexibility and potency of CNNs in analyzing electronic health records. Despite being primarily diagnostic, their model has substantial implications for cost prediction

by improving early diagnosis and thus impacting subsequent healthcare costs.Kwon and Kim (2021) addressed healthcare cost prediction by applying Gated Recurrent Unit (GRU) networks. Their model showcased the ability to extract and utilize both static features (like patient demographics) and dynamic features (like medical procedures) from medical records, leading to significantly improved prediction accuracy. The authors highlighted the importance of considering temporal dynamics in patient data when predicting healthcare costs.

In their study, Mateo *et al.* (2021) utilized the Extreme Gradient Boosting (XGBoost) algorithm to predict the most effective treatment for patients with acute bronchiolitis, a common and potentially severe respiratory disease in children. The authors demonstrated the algorithm's effectiveness in understanding complex interdependencies among numerous patient variables, enabling accurate and personalized treatment predictions. In particular, the XGBoost model surpassed traditional machine learning techniques in terms of predictive performance, while also ensuring interpretability via feature importance scores. This study underscores the potential of XGBoost in personalized medicine, particularly in the context of acute diseases where rapid and accurate treatment decisions are critical. In the research Tong *et al.* (2021), the authors explored the use of Bayesian networks and regression methods for predicting healthcare treatment costs. This study demonstrates an effective combination of probabilistic graphical models (Bayesian networks) and regression methods, leveraging the strengths of both approaches. Bayesian networks provided a comprehensive understanding of the probabilistic relationships among various medical variables. On the other hand, regression methods allowed for quantifying the relationships between these variables and the treatment costs. Through this synergy, the model offered robust cost predictions that are critical for healthcare planning and resource allocation. Moreover, the study emphasized the importance of interpretability. Both Bayesian networks, with their visual representation of variable relationships, and regression methods, with their clear mathematical relationships, ensure transparency and interpretability in cost predictions.

2.1.3 Post-hoc Explanation methods

As machine learning models become increasingly complex, there is a growing need for methods to interpret their decisions - a field of study known as explainable artificial intelligence (XAI). Post-hoc explanation methods are particularly valuable as they

elucidate the reasoning of a model after it has made a prediction. We'll be briefly discussing some popular post-hoc explanation methods below.

LIME

The LIME (Local Interpretable Model-Agnostic Explanations) framework is widely recognized as a popular approach for generating interpretable explanations for machine learning models after they have been trained. This framework offers explanations by assigning feature importance scores as its output, which helps in understanding the model's decision-making process. The fundamental algorithm for LIME is depicted in Fig. (2.2).

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K
  $\mathcal{Z} \leftarrow \{\}$
  for $i \in \{1, 2, 3, ..., N\}$ **do**
    $z'_i \leftarrow sample\_around(x')$
    $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z'_i, f(z_i), \pi_x(z_i) \rangle$
  end for
  $w \leftarrow$ K-Lasso$(\mathcal{Z}, K)$ ▷ with z'_i as features, $f(z)$ as target
  return w

Figure 2.2: Algorithm for LIME (Ribeiro *et al.* (2016))

Mathematically, LIME (Local Interpretable Model-Agnostic Explanations) can be described as follows. Let the sample to be explained be denoted as $\mathbf{x}_k$. The main objective of locally interpretable model-agnostic methods is to provide feature importance specific to $\mathbf{x}_k$. To achieve this, LIME first creates a surrogate dataset by sampling in the vicinity of $\mathbf{x}_k$. This surrogate dataset is denoted as $\mathcal{D}$, where each entry x_i, y_i represents a feature vector x_i in $\mathbb{R}^{n\times 1}$ and its corresponding target variable y_i in $\mathbb{R}$ for the i-th sample in the locality of $\mathbf{x}_k$. In other words, $\mathcal{D}$ can be represented as $\mathcal{D} \in \mathbb{R}^{m\times(n+1)}$, where m is the number of samples in the surrogate dataset.

The target values in $\mathcal{D}$ are obtained using the given black-box prediction model denoted as f_p, such that $f_p(x_i) = y_i$. To explain the decisions of f_p for the instance $\mathbf{x}_k$, an explainable AI method employs an explainer module f_e, which is trained on the surrogate dataset $\mathcal{D}$ using a fixed optimization objective.

In LIME, sparse linear models are used to explain the black-box model $f_p(\cdot)$. It considers a class of models denoted as $\mathcal{G}$, and utilizes a locally weighted square loss

function denoted as $L(\cdot)$. The goal is to find a sparse linear model $g \in G$ that minimizes the loss function $L(\cdot)$, where the model's complexity is controlled through a regularization parameter. This sparse linear model serves as an interpretable approximation of the black-box model and provides feature importance scores, enabling the understanding of the decision-making process of f_p for the instance $\mathbf{x}_k$.

$$L(fp, g, \pi_x) = \sum_{z,z' \in Z} \pi_x(\mathbf{z})(f_p(\mathbf{z}) - g(\mathbf{z'}))^2 \tag{2.6}$$

In LIME, a sparse linear model $g(\cdot) \in \mathcal{G}$ is employed, where $\mathbf{x} \in \mathbb{R}^n$ represents the instance being explained. The perturbed sample **z'** is created by randomly selecting a fraction of non-zero elements from **x**, and $\mathbf{z} \in \mathbb{R}^n$ represents the recovered sample in the original representation.

To create the surrogate dataset $\mathcal{Z}$, an exponential kernel π

x(**z**) = $\exp(-D(\mathbf{x}, \mathbf{z})^2/\sigma^2)$ is used, where $D(\cdot, \cdot)$ is a distance function and σ is the width of the kernel. The value of σ is chosen heuristically as it is not known for a given dataset. The samples around the non-zero values of **x** are randomly drawn to obtain a perturbed sample **z'**, along with the associated labels obtained using the black-box model $f_p(\cdot)$. This perturbed dataset $\mathcal{Z}$ is then used to optimize Equation (2.6), which leads to obtaining an explanation within the LIME framework.

The algorithm outlined in Fig.(2.2) can be subdivided into three key steps for a comprehensive understanding:

1. Sampling and Prediction:
   - Firstly, samples are generated around a specific feature set using a similarity kernel. This kernel function aids in distinguishing samples that differ from the given feature set.
   - Next, the prediction model is utilized to obtain target values for the generated sample sets. This step provides an approximation of the model's behavior on these samples.

2. Feature Selection:
   - After obtaining the target values, a feature selection technique such as LASSO (Least Absolute Shrinkage and Selection Operator) or forward selection is applied to identify the most relevant features.
   - These techniques help in choosing the top K features that have the strongest impact on the model's predictions. By selecting a subset of features, the algorithm aims to provide a concise and interpretable explanation.

3. Output Interpretation:
    - The final step involves presenting the selected top K features in a meaningful and interpretable manner.
    - This output aims to provide insights into the factors that significantly influence the model's predictions, aiding in the understanding and interpretation of the model's decision-making process.

By following these three steps, the algorithm aims to generate explanations that highlight the most important features and their impact on the model's predictions. It facilitates the interpretability and comprehensibility of the machine learning model, helping users gain insights into the key factors driving the model's behavior.

KernelSHAP (SHapley Additive exPlanations)

KernelSHAP (SHapley Additive exPlanations) (Lundberg and Lee (2017)) is an interpretable machine learning framework that provides explanations for individual predictions by assigning importance values to each feature. It is based on the concept of Shapley values from cooperative game theory and utilizes a kernel-based approach to estimate feature contributions. Kernel SHAP offers a flexible and model-agnostic solution for generating reliable and meaningful explanations.

Mathematically, Kernel SHAP can be described as follows. Let f_p represent the black-box prediction model that takes an input instance $\mathbf{x} \in \mathbb{R}^n$ and outputs a prediction. The goal of Kernel SHAP is to estimate the Shapley values ϕ_i for each feature x_i of $\mathbf{x}$, indicating the contribution of that feature to the prediction.

To compute the Shapley values, Kernel SHAP creates a set of coalitions, which are subsets of features. Each coalition S is a combination of features excluding x_i, and it represents all possible feature subsets that exclude the feature x_i. The coalitions are randomly sampled to generate reference instances $\mathbf{z}_S$, which are similar to the input $\mathbf{x}$ but with different combinations of features. The reference instances are used to approximate the expected prediction difference by comparing the model's predictions for the reference instances with the predictions for the input instance.

The Shapley value ϕ_i for feature x_i is computed as the average marginal contribution of the feature across all coalitions. It measures the change in the prediction caused by including feature x_i compared to the average prediction change when considering different feature subsets. This process is repeated for each feature, resulting in a set of

Shapley values that quantify the impact of each feature on the prediction.

The kernel function in Kernel SHAP plays a crucial role in determining the similarity between instances and coalitions. It defines the weighting scheme used to estimate the contribution of each coalition to the Shapley value. A common choice is the Gaussian kernel, given by $\pi_{\mathbf{x}}(\mathbf{z}_S) = e^{-\frac{||\mathbf{X}-\mathbf{Z}_S||_2}{\sigma^2}}$, where σ controls the width of the kernel. The kernel assigns higher weights to coalitions that are closer to the input instance, reflecting their greater relevance to the prediction.

By applying Kernel SHAP, explanations can be obtained at the individual prediction level, allowing users to understand the importance of each feature in the context of a specific prediction. These feature importance values provide valuable insights into the factors influencing the model's decision-making process and enhance the interpretability of complex machine learning models.

We observe that significant strides have been made in developing post-hoc explanation methods for machine learning whether by unifying various interpretation methods, incorporating interpretability into the learning process, or using prior beliefs to shape model learning, these contributions offer novel approaches to illuminate the 'black box' of machine learning. As the field continues to evolve, these foundational works will undoubtedly guide future research in the quest for greater transparency and accountability in AI.

2.2 Novelty

In existing literature, no single framework reliably estimates both treatment effect and the total medical expenditure incurred. Our research addresses this lacuna by proposing an innovative counterfactual inference frameworks. This framework allows for a joint estimation of the treatment effect and medical cost, facilitating balanced representations. Furthermore, we break new ground by introducing a post-hoc explainer, specifically designed for a multi-output causal inference based counterfactual neural networks. This explainer provides valuable explanations and interpretations of our proposed model, enhancing its transparency and utility. This two-fold novel contribution – a comprehensive estimation framework and a detailed explanatory tool – holds the potential to significantly advance the field of personalized healthcare, promoting both economic efficiency and treatment efficacy.

CHAPTER 3

METHODOLOGY

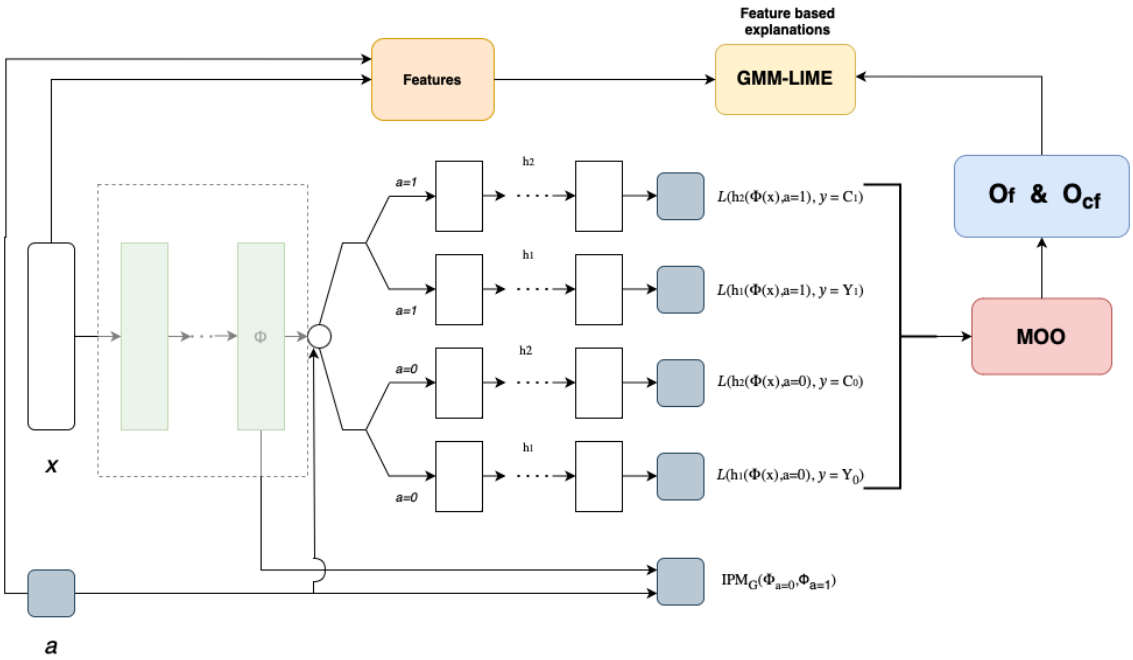

Figure 3.1: Complete work Methodology.

We propose a counterfactual inference-based framework, CFMedNet for jointly estimating treatment outcomes and associated medical costs. Our framework incorporates multi-objective optimization and introduces an explanation method: GMM LIME. We conduct experiments on a semi-synthetic dataset generated through specific processes, providing detailed explanations of each component in subsequent sections. By leveraging counterfactual inference, multi-objective optimization, and explanation methods, our framework offers insights into treatment effectiveness, costs, and interpretable decision-making. The overall methodology is depicted in Fig. (3.1).

3.1 CFMedNet

We propose an approach that reliably estimates the treatment efficacy and medical cost efficiency for patients. This work is an extension of CFRNet framework with an additional hypothesis layer used to model the associated medical costs. The assumptions of

positivity (1.11) and strong ignorability (1.9) in CFRNet holds for our case also. Further, we assume that the censoring is random. The terminologies and notations used in this work are defined as :-

- We define a representation mapping of covariates $\mathcal{X}$ to space $\mathcal{R}$ using one-to-one, twice differentiable function $\phi : \mathcal{X} \to \mathcal{R}$. Also, let $\psi : \mathcal{R} \to \mathcal{X}$ be the inverse function of ϕ, such that $(\psi(\phi(x))) = x$ for all $x \in \mathcal{X}$
- The outcome space $\mathcal{Y}, \mathcal{C} \subset \mathcal{R}$ and the treatment a is binary in nature { 0,1}.
  The outcome space in CFRNet refers to the range of possible outcomes or predictions that the model can generate. By defining two outcome spaces for our proposed architecture i.e CFMedNet, we allow the model to provide separate predictions or estimates for each category within those outcome spaces. This is quite beneficial when we want to analyze and understand the differential effects of treatments or interventions on different outcome categories.
- We define two hypothesis:
  $h_1 : \mathcal{R} \times \{0, 1\} \to \mathcal{Y}$
  $h_2 : \mathcal{R} \times \{0, 1\} \to \mathcal{C}$
  Here h_1, maps the input space $\mathcal{R}$ (which represents the features or covariates) along with a binary treatment indicator $\{0, 1\}$ to the outcome space $\mathcal{Y}$, while h_2, maps the input space along with a binary treatment indicator to the outcome space $\mathcal{C}$.
  By defining two hypotheses and corresponding outcomes in CFMedNet, the model estimates both treatment effects and associated costs. This approach assesses the impact of a treatment in terms of effectiveness and costs. Incorporating both outcomes allows CFMedNet to provide predictions for both aspects, enabling a comprehensive analysis of the treatment's impact on effectiveness and economic implications.
- The respective loss functions are :
  $L_1 : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$
  $L_2 : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_+$
  where both the losses are root mean squared error (RMSE) loss functions.

Definition 3.1. Let $p^{a=1}(x) := p(x|a = 1)$ and $p^{a=0}(x) := p(x|a = 0)$ be the treatment and control distributions respectively.

Definition 3.2. Let p_ϕ be the distribution induced by ϕ over $\mathcal{R}$. Then ${p_\phi}^{a=1}(r) := p_\phi(r|a = 1)$ and ${p_\phi}^{a=0}(r) := p_\phi(r|a = 0)$ are the treatment and control distributions ϕ induced over $\mathcal{R}$.

Definition 3.3. Given the hypothesis h_1 and h_2, the respective expected loss for an instance and treatment pair (x, a) is:

$$l_{1h,\phi}(x, t) = \int_{\mathcal{Y}} L(Y_t, h_1(\phi(x), a))p(Y_t|x)dY_t \tag{3.1}$$

$$l_{2h,\phi}(x, t) = \int_{\mathcal{C}} L(C_t, h_2(\phi(x), a))p(C_t|x)dC_t \tag{3.2}$$

Definition 3.4. ϵ_{F1} and ϵ_{CF1} denote the expected factual and counterfactual losses of h_1 and ϕ, defined as:

$$\epsilon_{F1}(h_1, \phi) = \int_{\mathcal{X}\times\{0,1\}} l_{1h,\phi}(x, a)p(x, a)dxda \tag{3.3}$$

$$\epsilon_{CF1}(h_1, \phi) = \int_{\mathcal{X}\times\{0,1\}} l_{1h,\phi}(x, a)p(x, 1-a)dxda \tag{3.4}$$

Definition 3.5. ϵ_{F2} and ϵ_{CF2} denote the expected factual and counterfactual losses of h_2 and ϕ, defined as:

$$\epsilon_{F2}(h_2, \phi) = \int_{\mathcal{X}\times\{0,1\}} l_{2h,\phi}(x, a)p(x, a)dxda \tag{3.5}$$

$$\epsilon_{CF2}(h_2, \phi) = \int_{\mathcal{X}\times\{0,1\}} l_{2h,\phi}(x, a)p(x, 1-a)dxda \tag{3.6}$$

Definition 3.6. On similar lines, the expected losses for treatment and control distributions are defined as:

$$\epsilon_{F1}{}^{a=1}(h_1, \phi) = \int_{\mathcal{X}} l_{1h,\phi}(x, 1)p^{a=1}(x)dx \tag{3.7}$$

$$\epsilon_{F1}{}^{a=0}(h_1, \phi) = \int_{\mathcal{X}} l_{1h,\phi}(x, 0)p^{a=0}(x)dx \tag{3.8}$$

$$\epsilon_{F2}{}^{a=1}(h_2, \phi) = \int_{\mathcal{X}} l_{2h,\phi}(x, 1)p^{a=1}(x)dx \tag{3.9}$$

$$\epsilon_{F2}{}^{a=0}(h_2, \phi) = \int_{\mathcal{X}} l_{2h,\phi}(x, 0)p^{a=0}(x)dx \tag{3.10}$$

$$\epsilon_{CF1}{}^{a=1}(h_1, \phi) = \int_{\mathcal{X}} l_{1h,\phi}(x, 1)p^{a=0}(x)dx \tag{3.11}$$

$$\epsilon_{CF1}{}^{a=0}(h_1, \phi) = \int_{\mathcal{X}} l_{1h,\phi}(x, 0)p^{a=1}(x)dx \tag{3.12}$$

$$\epsilon_{CF2}{}^{a=1}(h_2, \phi) = \int_{\mathcal{X}} l_{2h,\phi}(x, 1)p^{a=0}(x)dx \tag{3.13}$$

$$\epsilon_{CF2}{}^{a=0}(h_2, \phi) = \int_{\mathcal{X}} l_{2h,\phi}(x, 0)p^{a=1}(x)dx \tag{3.14}$$

Let $v := p(a = 1)$ be the fraction of population that has been treated.

Since $p(x, a) = v \cdot p^{a=1}(x) + (1v) \cdot p^{a=0}(x)$, we obtain the results as stated in Lemma 3.1 using Definition (3.4) and Definition (3.6).

Lemma 3.1.

$$\begin{aligned}\epsilon_F(h_1, h_2, \phi) = v \cdot (\epsilon_{F1}^{a=1}(h_1, \phi) + \epsilon_{F2}^{a=1}(h_2, \phi)) \\ +(1-v) \cdot (\epsilon_{F1}^{a=0}(h_1, \phi) + \epsilon_{F2}^{a=0}(h_2, \phi))\end{aligned} \tag{3.15}$$

$$\begin{aligned}\epsilon_{CF}(h_1, h_2, \phi) = (1-v) \cdot (\epsilon_{CF1}^{a=1}(h_1, \phi) + \epsilon_{CF2}^{a=1}(h_2, \phi)) \\ +v \cdot (\epsilon_{CF1}^{a=0}(h_1, \phi) + \epsilon_{CF2}^{a=0}(h_2, \phi))\end{aligned} \tag{3.16}$$

In Lemma 3.1, we postulate that the overall expected treated and control loss is a linear combination of both the respective hypothesis losses.

Lemma 3.2. Let $\phi : \mathcal{X} \to \mathcal{R}$ is an invertible representation with ψ as the inverse function. The distributions $p_\phi^{a=1}$ and $p_\phi^{a=0}$ are as defined in Definition 3.1. Let $v := p(a=1)$ be the fraction of population that has been treated. Assume G is a family of functions such that $g : \mathcal{R} \to \mathbb{R}$ and $IPM_G(.,.)$ refers to the integral probability metric induced by G. Let h_1 and h_2 are the two hypothesis as defined earlier. Also, let there exists a constant $B_\phi > 0$ such that for a given treatment, $a = \{0, 1\}$, the function $g(r, a) := \frac{1}{B_\phi} \cdot (l_{1h,\phi}(\psi(r), a) + (l_{2h,\phi}(\psi(r), a)) \in G$. We obtain:

$$\begin{aligned}\epsilon_{CF}(h_1, h_2, \phi) \leq (1-v)\{\epsilon_{F1}^{a=1}(h_1, \phi) + \epsilon_{F2}^{a=1}(h_2, \phi)\} \cdot v\{\epsilon_{F1}^{a=0}(h_1, \phi) + \epsilon_{F2}^{a=0}(h_2, \phi)\} \\ +B_\phi \cdot IPM_G(p_\phi^{a=0}, p_\phi^{a=1})\end{aligned} \tag{3.17}$$

Proof.

$$\begin{aligned}&\epsilon_{CF}(h_1, h_2, \phi) - (1-v) \cdot \{\epsilon_{F1}^{a=1}(h_1, \phi) + \epsilon_{F2}^{a=1}(h_2, \phi)\} - v \cdot \{\epsilon_{F1}^{a=0}(h_1, \phi) + \epsilon_{F2}^{a=0}(h_2, \phi)\} \\ &= (1-v) \cdot \{\epsilon_{CF1}^{a=1}(h_1, \phi) + \epsilon_{CF2}^{a=1}(h_2, \phi)\} + v \cdot \{\epsilon_{CF1}^{a=0}(h_1, \phi) + \epsilon_{CF2}^{a=0}(h_2, \phi)\} \\ &- (1-v) \cdot \{\epsilon_{F1}^{a=1}(h_1, \phi) + \epsilon_{F2}^{a=1}(h_2, \phi)\} - v \cdot \{\epsilon_{F1}^{a=0}(h_1, \phi) + \epsilon_{F2}^{a=0}(h_2, \phi)\}\end{aligned} \tag{3.18}$$

$$\begin{aligned}&= (1-v) \cdot \{\epsilon_{CF1}^{a=1}(h_1, \phi) + \epsilon_{CF2}^{a=1}(h_2, \phi) - \epsilon_{F1}^{a=1}(h_1, \phi) - \epsilon_{F2}^{a=1}(h_2, \phi)\} \\ &+ v \cdot \{\epsilon_{CF1}^{a=0}(h_1, \phi) + \epsilon_{CF2}^{a=0}(h_2, \phi) - \epsilon_{F1}^{a=0}(h_1, \phi) + \epsilon_{F2}^{a=0}(h_2, \phi)\}\end{aligned} \tag{3.19}$$

$$
\begin{aligned}
&= (1-v)\cdot\{\int_{\mathcal{X}} l_{1h,\phi}(x,1)p^{a=0}(x)dx + l_{2h,\phi}(x,1)p^{a=0}(x)dx \\
&- \int_{\mathcal{X}} l_{1h,\phi}(x,1)p^{a=1}(x)dx + l_{2h,\phi}(x,1)p^{a=1}(x)dx\} \\
&+ v\cdot\{\int_{\mathcal{X}} l_{1h,\phi}(x,0)p^{a=1}(x)dx + l_{2h,\phi}(x,0)p^{a=1}(x)dx \\
&- \int_{\mathcal{X}} l_{1h,\phi}(x,0)p^{a=0}(x)dx + l_{2h,\phi}(x,0)p^{a=0}(x)dx\}
\end{aligned} \tag{3.20}
$$

$$
\begin{aligned}
&= (1-v)\cdot\{(\int_{\mathcal{X}} l_{1h,\phi}(x,1) + l_{2h,\phi}(x,0))\cdot(p^{a=1}(x) - p^{a=0}(x))dx\} \\
&\quad +v\cdot\{(\int_{\mathcal{X}} l_{1h,\phi}(x,0) + l_{2h,\phi}(x,1))\cdot(p^{a=0}(x) - p^{a=1}(x))dx\}
\end{aligned} \tag{3.21}
$$

$$
\begin{aligned}
&= B_\phi\cdot(1-v)\{\int_{\mathcal{R}} \frac{1}{B_\phi}\cdot(l_{1h,\phi}(\psi(r),1) + (l_{2h,\phi}(\psi(r),1))\cdot(p_\phi^{a=1}(r) - p_\phi^{a=0})(r)dr\} \\
&\quad +B_\phi\cdot v\{\int_{\mathcal{R}} \frac{1}{B_\phi}\cdot(l_{1h,\phi}(\psi(r),0) + (l_{2h,\phi}(\psi(r),0))\cdot(p_\phi^{a=1}(r) - p_\phi^{a=0})(r)dr\}
\end{aligned} \tag{3.22}
$$

$$
\begin{aligned}
&\leq B_\phi\cdot(1-v)\sup_{g'\in G} \mid \int_R g^{'}(r)\cdot(p_\phi^{a=0}(r) - p_\phi^{a=1}(r))dr \mid \\
&\quad +B_\phi\cdot v\sup_{g'\in G} \mid \int_R g^{'}(r)\cdot(p_\phi^{a=1}(r) - p_\phi^{a=0}(r))dr \mid
\end{aligned} \tag{3.23}
$$

$$
= B_\phi\cdot IPM_G(p_\phi^{a=0}(r), p_\phi^{a=1}(r)) \tag{3.24}
$$

□

Here, equality (3.18) is as per Equation (3.16) of Lemma (3.1), while equality (3.20) is by Definition 3.6 of the expected losses. Further, equality (3.22) is the change of variables and inequalities (3.23) (3.24) are by the definition of function g and term IPM_G respectively.

Definition 3.7. For $a = 0, 1$, we define:

$$
\begin{aligned}
m_a(x) &= \mathbf{E}[Y_a \mid x] \\
n_a(x) &= \mathbf{E}[C_a \mid x]
\end{aligned}
$$

We can rewrite the treatment effect as:

$$\tau(x) = m_1(x) + n_1(x) - m_0(x) - n_0(x)$$

Recall that $f_1 : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$ and $f_2 : \mathcal{X} \times \{0,1\} \to \mathcal{C}$ are the hypotheses such that $f_1(x,a) = h_1(\phi(x),a)$ and $f_2(x,a) = h_2(\phi(x),a)$ for a representation ϕ and hypotheses h_1, h_2 defined over the output of ϕ.

Definition 3.8. The treatment effect is estimated using:

$$\hat{\tau} = f_1(x,1) - f_1(x,0) + \beta(f_2(x,1) - f_2(x,0))$$

where β is a scaling factor to balance the scales of both hypotheses.

Definition 3.9. The expectation of square of difference between estimated and actual treatment effect, also called as expected Precision in Estimation of Heterogeneous Effect (PEHE) loss is calculated as:

$$\epsilon_{PEHE}(f) = \int_{\mathcal{X}} (\hat{\tau}_f(x) - \tau(x))^2 p(x) dx$$

Definition 3.10. The expected variance of Y_a and C_a with respect to distribution $p(x;a)$:

$$\sigma^2{}_{Y_a}(p(x,a)) = \int_{\mathcal{X}\times\mathcal{Y}} (Y_a - m_a(x))^2 \, p(Y_a|x)p(x,a) dY_a dx$$
$$\sigma^2{}_{C_a}(p(x,a)) = \int_{\mathcal{X}\times\mathcal{C}} (C_a - n_a(x))^2 \, p(C_a|x)p(x,a) dC_a dx$$

Lemma 3.3. Given two functions $f_1 : \mathcal{X} \times \{0,1\} \to \mathcal{Y}$ and $f_2 : \mathcal{X} \times \{0,1\} \to \mathcal{C}$, and distribution $p(x,a)$ defined over $\mathcal{X} \times \{0,1\}$:

$$\int_{\mathcal{X}\times\{0,1\}} (f_1(x,a) - m_a(x))^2 p(x,a) dx da + \int_{\mathcal{X}\times\{0,1\}} (f_2(x,a) - n_a(x))^2 p(x,a) dx da$$
$$= \epsilon_F - \sigma^2_{Y_a}(p(x,a)) - \sigma^2_{C_a}(p(x,a)) \tag{3.25}$$

$$\int_{\mathcal{X}\times\{0,1\}} (f_1(x,a) - m_a(x))^2 p(x,1-a) dx da + \int_{\mathcal{X}\times\{0,1\}} (f_2(x,a) - n_a(x))^2 p(x,1-a) dx da$$
$$= \epsilon_{CF} - \sigma^2_{Y_a}(p(x,1-a)) - \sigma^2_{C_a}(p(x,1-a)) \tag{3.26}$$

Proof.

$$\epsilon_F = \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f_1(x,a) - Y_a\right)^2 p(Y_a|x)p(x,a)dY_a dxda + \int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} \left(f_2(x,a) - C_a\right)^2 p(C_a|x)p(x,a)dC_a dxda \tag{3.27}$$

$$\begin{aligned} &= \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f_1(x,a) - m_a(x)\right)^2 p(Y_a|x)p(x,a)dY_a dxda \\ &+ \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(m_a(x) - Y_a\right)^2 p(Y_a|x)p(x,a)dY_a dxda \\ &+ 2\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f_1(x,a) - m_a(x)\right)\cdot\left(m_a(x) - Y_a\right) p(Y_a|x)p(x,a)dY_a dxda \end{aligned} \tag{3.28}$$

$$\begin{aligned} &= \int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} \left(f_2(x,a) - n_a(x)\right)^2 p(C_a|x)p(x,a)dC_a dxda \\ &+ \int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} \left(n_a(x) - C_a\right)^2 p(C_a|x)p(x,a)dC_a dxda \\ &+ 2\int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} \left(f_2(x,a) - n_a(x)\right)\cdot\left(n_a(x) - C_a\right) p(C_a|x)p(x,a)dC_a dxda \end{aligned} \tag{3.29}$$

$$\begin{aligned} &= \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f_1(x,a) - m_a(x)\right)^2 p(Y_a|x)p(x,a)dY_a dxda \\ &+ \int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} \left(f_2(x,a) - n_a(x)\right)^2 p(C_a|x)p(x,a)dC_a dxda \\ &+ \sigma^2_{Y_a}(p(x,a)) + \sigma^2_{C_a}(p(x,a)) \end{aligned} \tag{3.30}$$

$$\epsilon_{CF} = \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f_1(x,a) - Y_a\right)^2 p(Y_a|x)p(x,1-a)dY_a dxda + \int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} \left(f_2(x,a) - C_a\right)^2 p(C_a|x)p(x,1-a)dC_a dxda \tag{3.31}$$

$$\begin{aligned} &= \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f_1(x,a) - m_a(x)\right)^2 p(Y_a|x)p(x,1-a)dY_a dxda \\ &+ \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(m_a(x) - Y_a\right)^2 p(Y_a|x)p(x,1-a)dY_a dxda \\ &+ 2\int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f_1(x,a) - m_a(x)\right)\cdot\left(m_a(x) - Y_a\right) p(Y_a|x)p(x,1-a)dY_a dxda \end{aligned} \tag{3.32}$$

$$
\begin{aligned}
&= \int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} \left(f_2(x,a) - n_a(x)\right)^2 p(C_a|x)p(x,1-a)dC_a dxda \\
&+ \int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} \left(n_a(x) - C_a\right)^2 p(C_a|x)p(x,1-a)dC_a dxda \\
&+ 2\int_{\mathcal{X}\times\{0,1\}\times\mathcal{C}} \left(f_2(x,a) - n_a(x)\right)\cdot\left(n_a(x) - C_a\right) p(C_a|x)p(x,1-a)dC_a dxda
\end{aligned} \tag{3.33}
$$

$$
\begin{aligned}
&= \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f_1(x,a) - m_a(x)\right)^2 p(Y_a|x)p(x,1-a)dY_a dxda \\
&+ \int_{\mathcal{X}\times\{0,1\}\times\mathcal{Y}} \left(f_2(x,a) - n_a(x)\right)^2 p(Y_a|x)p(x,1-a)dY_a dxda \\
&+ \sigma^2_{Y_a}(p(x,1-a)) + \sigma^2_{C_a}(p(x,1-a))
\end{aligned} \tag{3.34}
$$

□

Equality (3.27)is by the definition of ϵ_F while equality (3.28)is simple mathematical manipulation. Equality (3.29) is due to Definition (3.10) and also because two integral terms approach zero since $m_a(x) = \int_{\mathcal{X}} Y_a p(Y_a|x)dx$ and $n_a(x) = \int_{\mathcal{X}} C_a p(C_a|x)dx$

Theorem 3.1. Assume $\phi : \mathcal{X} \to \mathcal{R}$ is a one-to-one representation function, and ψ is its inverse. Further, assume that ${p_\phi}^{a=0}, {p_\phi}^{a=1}$ are defined as in Definition (3.2). Suppose $v = p(a=1)$. Let G be a family of functions $g : \mathcal{R} \to \mathbb{R}$, and $IPM_G(.,.)$ indicate the integral probability metric induced by G . Consider $h_1 : \mathcal{R} \times \{0,1\} \to \mathcal{Y}, h_2 : \mathcal{R} \times \{0,1\} \to \mathcal{C}$
to be two hypotheses. Let the loss $L(y_1, y_2) = (y_1 - y_2)^2$. Assume there exists a constant $B_\phi > 0$, such that the functions $g_{\Phi,h} := \frac{1}{B_\phi} \cdot (l_{1h,\phi}(\psi(r), a) + (l_{2h,\phi}(\psi(r), a)) \in G$ for a $\in \{0,1\}$. We then have:

$$
\epsilon_{PEHE} \leq 4 \cdot \{\epsilon_F(h_1, h_2, \phi) + \epsilon_{CF}(h_1, h_2, \phi) - 2 \cdot (\sigma^2_Y + \sigma^2_C)\} \tag{3.35}
$$

$$
\begin{aligned}
\leq 4 \cdot \{&\epsilon_F 1^{a=1}(h_1, \phi) + \epsilon_F 2^{a=1}(h_2, \phi) + \epsilon_F 1^{a=0}(h_1, \phi) + \epsilon_F 1^{a=0}(h_2, \phi) \\
&+ B_\phi \cdot IPM_G(p_\phi^{a=0}, p_\phi^{a=1}) - 2 \cdot (\sigma^2_Y + \sigma^2_C)\}
\end{aligned} \tag{3.36}
$$

Proof.

$$\epsilon_{PEHE} = \int_{\mathcal{X}} (\hat{\tau}_f(x) - \tau(x))^2 p(x) dx \tag{3.37}$$

$$\begin{aligned} &= \int_{\mathcal{X}} \{(f_1(x,1) + f_2(x,1) - f_1(x,0) - f_2(x,0)) \\ &- (m_1(x) + n_1(x) - m_0(x) - n_0(x))\}^2 p(x) dx \end{aligned} \tag{3.38}$$

$$\begin{aligned} &= \int_{\mathcal{X}} \{((f_1(x,1) - (m_1(x)) + (f_2(x,1) - n_1(x)) \\ &+ (m_0(x) - f_1(x,0)) + (n_0(x) - f_2(x,0))^2\} p(x) dx \end{aligned} \tag{3.39}$$

$$\begin{aligned} &\leq 2 \cdot \int_{\mathcal{X}} \{(f_1(x,1) - m_1(x) + f_2(x,1) - n_1(x)) \\ &+ (m_0(x) - f_1(x,0) + n_0(x) - f_2(x,0))^2\} p(x) dx \end{aligned} \tag{3.40}$$

$$\begin{aligned} &= 2 \cdot \int_{\mathcal{X}} \{(f_1(x,1) - m_1(x) + f_2(x,1) - n_1(x))^2 p(x, a=1) dx \\ &+ 2 \cdot \int_{\mathcal{X}} (m_0(x) - f_1(x,0) + (n_0(x) - f_2(x,0))^2 p(x, a=0) dx \\ &+ 2 \cdot \int_{\mathcal{X}} \{(f_1(x,1) - m_1(x) + f_2(x,1) - n_1(x))^2 p(x, a=0) dx \\ &+ 2 \cdot \int_{\mathcal{X}} (m_0(x) - f_1(x,0) + (n_0(x) - f_2(x,0))^2 p(x, a=1) dx \end{aligned} \tag{3.41}$$

$$\begin{aligned} &= 2 \cdot \int_{\mathcal{X} \times \{0,1\}} (f_1(x,a) - m_a(x) + f_2(x,a) - n_a(x))^2 p(x,a) dx da \\ &+ 2 \cdot \int_{\mathcal{X} \times \{0,1\}} (f_1(x,a) - m_a(x) + f_2(x,a) - n_a(x))^2 p(x, 1-a) dx da \end{aligned} \tag{3.42}$$

$$\begin{aligned} &\leq 4 \cdot \{ \int_{\mathcal{X} \times \{0,1\}} (f_1(x,a) - m_a(x))^2 p(x,a) dx da \\ &+ \int_{\mathcal{X} \times \{0,1\}} (f_2(x,a) - n_a(x))^2 p(x,a) dx da \\ &+ 4 \cdot \{ \int_{\mathcal{X} \times \{0,1\}} (f_1(x,a) - m_a(x))^2 p(x, 1-a) dx da \\ &+ \int_{\mathcal{X} \times \{0,1\}} (f_2(x,a) - n_a(x))^2 p(x, 1-a) dx da \} \end{aligned} \tag{3.43}$$

$$\epsilon_{PEHE} \leq 4 \cdot \{\epsilon_F(h_1, h_2, \phi) + \epsilon_{CF}(h_1, h_2, \phi) - 2 \cdot (\sigma_Y^2 + \sigma_C^2)\} \tag{3.44}$$

□

Inequality (3.40) and (3.43) are derived using the mathematical identity $(a+b)2 \leq 2 \cdot (a^2+b^2)$.The second inequality in the theorem can be proved by using Lemma (3.1) and Lemma (3.2.) combined.

The architecture of proposed CFMedNet method includes a representation layer and two hypothesis layers (each for medical cost and outcome) with respect to two treatment arms as shown in Figure (3.2).

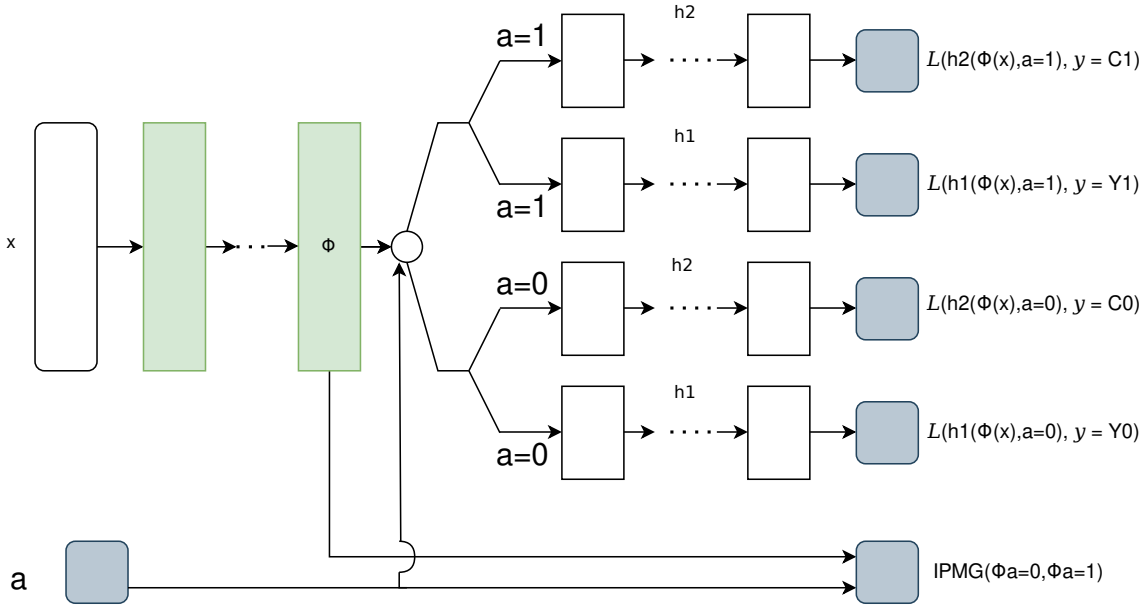

Figure 3.2: CFMedNet Architecture.

The optimal solution is obtained by minimizing loss function as depicted in equation(3.45) using Adam optimizer where the error is backpropagated via both hypotheses layers and representation networks.

$$\mathcal{L} = \frac{\beta}{n} \sum_{i=1}^{n} s_i \cdot (L(h_1(\phi(x_i), a_i), y_i) + L(h_2(\phi(x_i), a_i), c_i)) + \lambda \cdot \mathcal{R}(h) + \gamma \cdot IPM_G(\{\phi(x_i)\}_{i:a_i=0}, \{\phi(x_i)\}_{i:a_i=1}) \tag{3.45}$$

where $s_i = \frac{a_i}{2v} + \frac{1-a_i}{2(1-v)}$,$v = \frac{1}{n}\sum_{i=1}^{n} a_i$ and $\mathcal{R}(\cdot)$ is a model complexity term. Also v refers to the proportion of treated instances $(a_i = 1)$, $\{\beta, \lambda, \gamma\}$ are the hyperparameters and $L(.,.)$ is squared error loss. The IPM term utilized MMD (Sriperumbudur *et al.* (2012)) distance metric.

The described approach involves using deep learning algorithms to estimate the effects of treatment and medical costs for patients. The algorithms consist of multiple fully connected neural networks that are trained separately on different treatment arms in each epoch. The approach evaluates both factual and counterfactual cases to estimate the effect of treatment on patients. To assess the performance of the model, various metrics such as average treatment effect, mean squared error, and PEHE are calculated. In summary, the approach is a complex and advanced method for estimating treatment effects and medical costs, which could have practical applications in the field of healthcare and medicine.

3.2 XAI Models for Counterfactual Inference

The utilization of machine learning (ML) models, such as the one we propose, to predict treatment effects and corresponding medical costs can lead to substantial time and resource savings in the healthcare sector. Notably, it also has the potential to decrease mortality rates by preventing inappropriate treatment assignment. However, a significant obstacle is that medical practitioners often possess limited understanding of ML models, viewing them as "black boxes" due to the lack of transparency in the decision-making process. The healthcare domain is a particularly sensitive area; the lack of explainability or interpretability of AI/ML models makes their deployment in a clinical setting a delicate task, especially when it comes to decision-making. Despite the fact that AI/ML models are designed to assist doctors in their decision-making process, it is the doctors who ultimately bear responsibility for the medical decisions made. Therefore, the explainability of AI models becomes a crucial factor in the development and deployment of AI-based healthcare solutions. Hence, efficient Explainable AI (XAI) models are crucial for frameworks similar to our proposed architecture, CFMedNet, to build trust among doctors regarding the decisions made by our model.

3.2.1 Approach Selection

Local Interpretable Model-Agnostic Explanations or LIME, is a technique developed to elucidate the decisions of any machine learning model by offering localized explanations for specific predictions. Surrogate sampling is a key component of LIME as it facilitates the creation of an interpretable model around the prediction of interest. LIME

emphasizes the local explanation of predictions, meaning it selects a particular data instance for which we aim to decipher the prediction. This selection sets the stage for surrogate sampling, where LIME generates a new dataset by introducing slight alterations, or perturbations, to the selected data instance. Each of these modified versions then receives a prediction from the original model. Following this, each perturbed instance is allocated a weight based on its similarity to the original instance, with higher weights given to those closer to the original. Subsequently, LIME trains an easily interpretable model using this newly weighted dataset. The surrogate model, thus developed, replicates the actions of the black box model within the vicinity of the selected instance, thereby shedding light on the reasons behind a particular prediction. Therefore, surrogate sampling is fundamental in enabling local interpretability in machine learning. Fundamentally, surrogate sampling empowers LIME to form a 'local neighborhood' around the instance we are trying to interpret. The surrogate model, trained within this neighborhood, delivers a simplified and more interpretable approximation of the original complex model's behavior within this local scope.

3.2.2 GMM-LIME

Gaussian Mixture Model LIME (GMM-LIME) relies on surrogate sampling to generate a set, $\mathcal{D}$, composed of random samples (or perturbations) centered around a particular instance, $\mathbf{x}$. The need for surrogate sampling stems from the requirement for creating local, classifier-based explanations that accurately reflect the class imbalances inherent to the data.

Our objective is to form a surrogate dataset that is balanced and includes samples from all classes, though in differing quantities. For this, we use a Gaussian sampling technique, similar to the one mentioned in Ribeiro *et al.* (2016), where we expand the standard deviation to increase the sampling neighborhood and thus secure surrogate instances from all classes.

To further address class imbalance within $\mathcal{D}$, we utilize Gaussian Mixture Models (GMM). A GMM is a probabilistic model based on the assumption that all instances are generated from a finite number of Gaussian distributions with parameters that are yet to be determined Pedregosa *et al.* (2018). With the help of the bootstrapped samples, we train a GMM of c Gaussians, which we then use to oversample the minority classes,

ultimately achieving a balanced surrogate dataset. The following algorithm describes the sampling process from a GMM. (Nanavati and Prasad (2023)).

Algorithm 1 GMM – Sampling from a Gaussian Mixture Model

Require: Imbalanced Surrogate dataset $\mathcal{D}$, and corresponding labels $\mathbf{y}_{\mathcal{D}}$, Number of classes c

1: Fit a GMM on $\mathcal{D}$ to get cluster mean and variances.
2: Identify minority classes, based on the number of instances in each cluster.
3: Sample the required number of minority class instances.

Ensure: Oversampled Data from the Gaussian Mixture Model

Although these sampling techniques may not necessarily enhance the quality of the samples, they are critical for reducing imbalance in $\mathcal{D}$. In the following sections, we show that GMM-LIME outperforms other perturbation-based methods such as LIME in terms of stability. Following this, we implement a forward feature selection process as proposed in LIME to select the top k features and present their corresponding explanation scores. The initialization step of the algorithm involves specifying the black-box model, denoted as f, and the instance x in the feature space. Additionally, the number of surrogate samples, denoted as n, and the number of Gaussian Mixture components, denoted as C, are set. The following step is to train a Gaussian Mixture Model (GMM) using Algorithm[**??**]. This involves fitting the GMM to the entire dataset. After the model training, the posterior probability of each GMM component for a specific instance x is computed. This helps determine the probability that x belongs to each cluster. Sampling is performed to generate N samples, denoted as $x_1, x_2, ..., x_N$, based on the posterior probabilities of the GMM components for x. The sampling process is biased towards selecting instances that are close to x and gives more weight to more probable clusters.

Once the samples are generated, the distance between each sample instance x_i and x is computed. A suitable distance metric, such as Euclidean distance, is used for this purpose. Weighting is then performed by calculating weights, denoted as w, using the formula $w = \exp\left(-d^2/(2\sigma^2)\right)$. The width parameter σ of the Gaussian kernel is set as the median of the distances. The weights decrease as the distance increases, giving more influence to samples that are closer to x. Next, a simple interpretable model, such as linear regression, is trained using the samples, their corresponding target values $f(x_i)$, and the sample weights w. This simple model approaches the behavior of the complex model around the instance x.

Finally, the explanation is extracted from the simple model. The explanation provides insights into how the complex model makes predictions specifically at the instance x.

3.2.3 Weighted Multi-Objective Optimization (MOO)

We propose employing weighted multi-objective optimization to simultaneously optimise multiple objectives in a problem by allocating weights to each. The approach of using weighted multi-objective optimization to convert an n-tuple output from a neural network to a single-valued output can be particularly beneficial when we aim to explain the model predictions using explainer models like LIME (Local Interpretable Model-Agnostic Explanations). These models predominantly clarify single-value output models, as these explanations are simpler to comprehend. Thus, by transforming the n-tuple output to a single value, we enhance the simplicity of the model's predictions.

Moreover, this method allows us to assign varied weights to different objectives according to their significance. Therefore, we can adapt the single-value output to mirror the trade-offs, such as treatment efficacy and medical cost efficiency, which hold the highest relevance to the problem in focus. This structured approach to handle trade-offs ensures that the ultimate model output and corresponding explanations bear significance. This strategy also facilitates decision making, as it allows decision-makers to effortlessly compare different instances for single-valued outputs. This becomes an essential factor in sectors where interpretability and decision-making work synergistically, such as in healthcare.

In the context of our multi-output network, which encompasses treatment efficacy and medical cost efficiency, the output comprises a tuple of values. To convert this tuple output into a single output, we utilize weighted multi-objective optimization. We use survey-based weights derived from a survey Brenan (2022) that collects insights from domain experts, defining the relative importance of each objective. These weights aid in formulating an optimization problem that represents the trade-off between treatment efficacy and medical cost efficiency. Considering the objectives and their associated weights, a linear-weighted method is applied to convert the tuple output into a single output, customized to meet the demands of the problem at hand. The process is elucidated in the following algorithm.

Algorithm 2 Weighted Multi-objective optimization

Require: Let $\mathcal{X}$ be the set of all features.

1: a neural network $f(x,t)$ that maps input (x,t) to output (Y_0, Y_1, C_0, C_1), where $x \in \mathcal{X}$ and $t \in \{0,1\}$.
2: The output of the neural network is represented by the high-dimensional feature space $F \in \mathbb{R}^4$.
3: Define a multi-objective loss function $g(F)$ that maps feature space F to a scalar value representing the trade-off between treatment efficacy and cost efficiency.
4: Using multi-objective optimization techniques, solve the following optimization problem to obtain a single value for output loss, O: minimize $g(F)$ subject to $f(x,t) = (Y_0, Y_1, C_0, C_1)$
5: Employ an explainability module to generate an interpretable explanation for the neural network's predictions by approximating the local behavior of $f(x,t)$ in the vicinity of the input (x,t,O).

CHAPTER 4

EXPERIMENTATION AND EVALUATION

4.1 Dataset

We have used a semi-synthetic dataset for simulation of experiments. It is a type of dataset that is generated by combining real-world data with synthetic data. This dataset is curated by using the methodology of the Atlantic Causal Inference Conference (ACIC) Data Analysis Challenge 2017 (Hahn *et al.* (2019)) where the outcome variable and treatment is generated through a data generating process (DGP) with covariates taken from Infant Health and Development Program, or IHDP (Gross *et al.* (1993)). The outcome variable in the DGP is obtained depending on the following four types of errors:-

- additive, independent, identically distributed,
- additive, group correlated,
- additive, heteroskedastic,
- non-additive, independent, identically distributed.

In all instances, the error terms were generated following Gaussian distributions. Further the DGP has "high" or "low" settings depending on:

- magnitude of the causal effect, ξ (which can take value either 0.33 or 2)
- strength of the confounding, κ (which can take value either (0.5,0) or (-1,3))
- noise level in the response variable, (which can take value either 0.25 or 1.25)

In the subsequent section generation process of treatment outcome and cost variable is explained. It should be noted that in the DGP, only 8 features are utilized among a total of 58 from the IHDP data:

- X_1 : Age of mother (continuous),
- X_3 : The number of cigarettes smoked by a mother in a day (continuous),
- X_{10} : The endocrine condition of the mother (binary),
- X_{14} : The nervous system condition of the mother (binary),
- X_{15} : Complications during childbirth experienced by mother (binary),
- X_{21} : Birth place of mother (categorical),
- X_{24} : Race of mother (binary),
- X_{43} : Bilirubin of child (continuous).

Table (4.1) provides the correlations among these variables for reference.

$correlation$	X_1	X_3	X_{10}	X_{14}	X_{15}	X_{21}	X_{24}	X_{43}
X_1	1.00	0.04	-0.07	-0.03	-0.04	-0.07	0.03	-0.01
X_3	0.04	1.00	-0.02	0.03	-0.02	-0.10	-0.16	0.13
X_{10}	-0.07	-0.02	1.00	0.04	0.09	-0.02	-0.10	-0.07
X_{14}	-0.03	0.03	0.04	1.00	0.09	-0.03	-0.08	0.07
X_{15}	-0.04	-0.02	0.09	0.09	1.00	-0.03	0.04	-0.04
X_{21}	-0.07	-0.10	-0.02	-0.03	-0.03	1.00	0.20	-0.00
X_{24}	0.03 -	-0.16	-0.10	-0.08	0.04	0.20	1.00	-0.11
X_{43}	-0.01	0.13	-0.07	0.07	-0.04	-0.00	-0.11	1.00

Table 4.1: Correlation matrix of control variables

Treatment Outcome Synthesis

The treatment outcome synthesis is done using the DGP as below :

$$f(x) = x_1 + x_{43} + 0.3(x_{10} - 1) \tag{4.1}$$

$$\pi(x) = Pr(Z_i = 1) = \frac{1}{1 + exp(\kappa_1 f(x) + \kappa_2)} \tag{4.2}$$

$$\mu(x) = -sin(\phi(\pi(x))) + x_{43} \tag{4.3}$$

$$\tau(x) = \xi(x_3 x_{24} + (x_{14} - 1) - (x_{15} - 1)) \tag{4.4}$$

$$\sigma(x) = 0.4 + \frac{x_{21} - 1}{15} \tag{4.5}$$

where x_i is the i^{th} covariate in the IHDP data, and $\phi(.)$ refers to a normal random variable's cumulative distribution function. It is assumed that the errors(ϵ), have an independent, identical standard normal distribution. Let

$$\sigma_y = \eta \sqrt{\text{Var}\left(\mu(x) + \pi(x)\tau(x)\right)}$$

where variance is taken over observed samples. The outcome variable is then computed using:-

$$Y_i^f = \mu(x_i) + \tau(x_i)Z_i + \sigma_y \epsilon_i \tag{4.6}$$

where Z_i is the treatment value corresponding to i^{th} instance. We have used the settings such that $\kappa_1 = 3$, $\kappa_2 = -1$, $\eta = 0.25$ and $\xi = 2$. For counterfactual outcome generation, the same process is followed except that in equation(4.6), counterfactual treatment value is used such that the outcome generation equation becomes:

$$Y_i^{cf} = \mu(x_i) + \tau(x_i)(1 - Z_i) + \sigma_y \epsilon_i \tag{4.7}$$

The range of values for outcome generated by equation (4.6) is

{min: -6.093, max: 82.588}

while for counterfactual outcome generated by equation (4.7), it is

{min: -7.145, max: 88.213}

Medical Cost Synthesis

For medical cost synthesis, the DGP is modified as follows:-

$$f(x) = x_1 + x_{43} + 0.3(x_{10} - 1) \tag{4.8}$$

$$\pi(x) = Pr(Z_i = 1) = \frac{1}{1 + exp(\kappa_1 f(x) + \kappa_2)} \tag{4.9}$$

$$\mu(x) = \mid \log(\pi(x)) \mid + x_{43} \tag{4.10}$$

$$\tau(x) = \alpha - \xi(x_3 x_{24} + \mid (x_{14} - 1) - (x_{15} - 1) \mid) \tag{4.11}$$

$$\sigma(x) = 0.4 + \frac{x_{21} - 1}{15} \tag{4.12}$$

where x_i is the i^{th} covariate in the IHDP data and α is a parameter to regulate the overlapping of control and treatment cost distributions. Here, the value of α is kept equal to 400. It is assumed that the errors(ϵ) in medical cost have an independent, identical exponential distribution.

Then the factual cost and counterfactual cost is generated using equation (4.7) and equation (4.8) respectively with the identical parametric settings as in outcome synthesis.

The range of values for generated medical cost is

{min: 74.606, max: 539.082}

and for counterfactual medical cost, range is

{min: 49.228, max: 618.223}.

The final dataset consists of 4302 instances, 58 IHDP attributes, treatment column, factual and counterfactual treatment outcome column as well as factual and counterfactual medical cost column.

The histogram plots of factual and counterfactual medical costs with respect to treated and control population are depicted in Fig. 4.1 and Fig. 4.2.

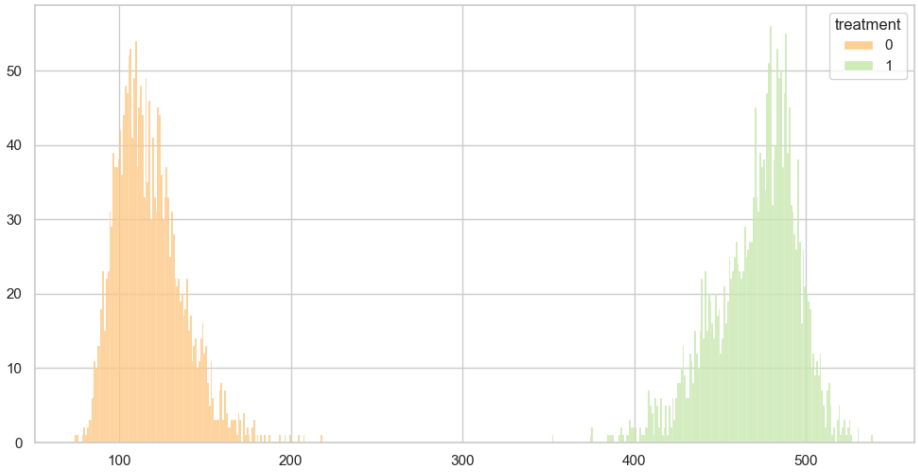

Figure 4.1: Histogram plot of Factual Medical Cost in Semi-Synthetic Dataset (Treated=1, Control=0)

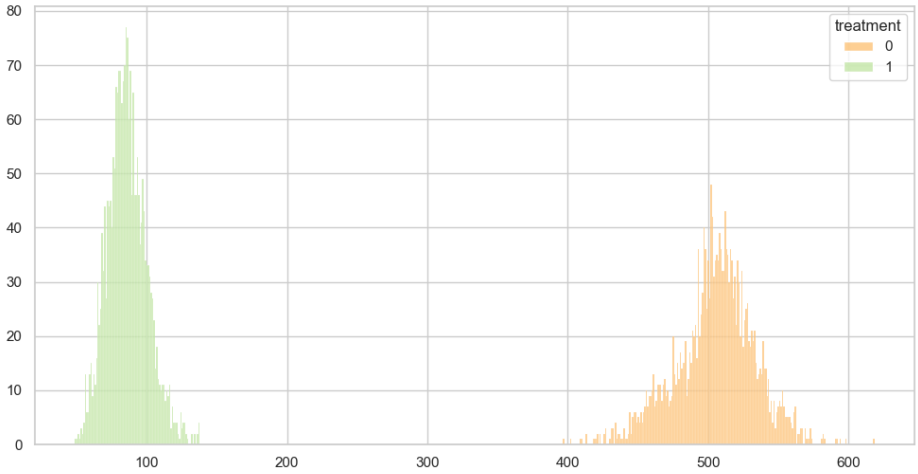

Figure 4.2: Histogram plot of Counterfactual Medical Cost in Semi-Synthetic Dataset (Treated=1, Control=0)

4.2 CFMedNet

A series of experiments utilizing the CFMedNet method were conducted on our semi-synthetic dataset. The factual treatment outcome values spanned from -6.093 to 82.588, while the counterfactual treatment outcome values ranged from -7.145 to 88.213. Similarly, the factual medical cost values ranged from 74.606 to 539.082, and the counterfactual medical cost values varied between 49.228 and 618.223. Further, Scatter Plots (Fig. 4.5)comparing Input and Representation Network Transformed Data have been visualised (MasaAsami (2022)).

- The representation network used in the CFMedNet method consists of a multi-layer perceptron with 3 hidden layers, each having 48 nodes, and an output layer with 48 nodes.
- The dropout rate for the representation network is set to 0.145, and the Rectified Linear Unit (ReLU) activation function is used for the hidden layers.
- The neural network architecture for outcome and medical cost hypotheses also consists of a multi-layer perceptron with 3 hidden layers, each having 32 nodes.
- The dropout rate for these networks is again 0.145, and ReLU activation function is used for the hidden layers.
- The same network architecture is applied to both the control and treated arms in both treatment outcome and medical cost categories.
- The Adam optimizer is used with a weight decay of 0.5.
- The learning rate decay step size is set to 100, and the multiplicative factor for learning rate decay is 0.97.
- Fig.(4.5) shows a scatter plot of the input data and the transformed data from the representation network using t-SNE across two dimensions.
- A Maximum Mean Discrepancy value of zero indicates that the control and treated groups are comparable in terms of representation layer output, indicating no treatment bias.
- The dataset is split into an 80:20 train-test ratio. Further, grid search is performed for 8000 epochs with hyperparameter γ values of 10^8,10^5 and 100, and learning rate values of 0.1 and 0.01.
- The results of the grid search are presented in Table (4.2) and Table (4.3).
- The hyperparameter γ represents the importance of the IPM loss term, with higher values indicating a higher contribution to the overall loss.

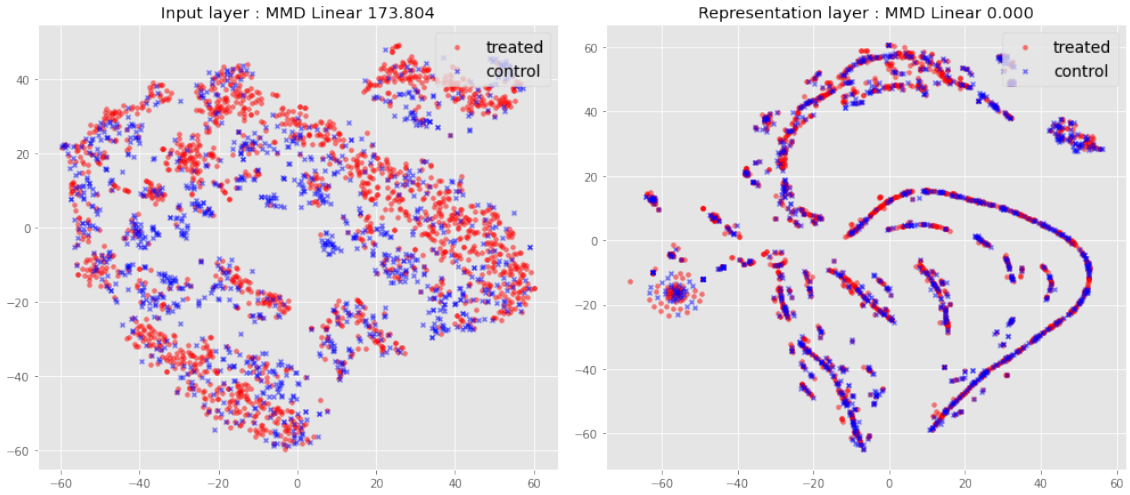

Figure 4.3: Scatter Plot comparing Input and Representation Network Transformed Data

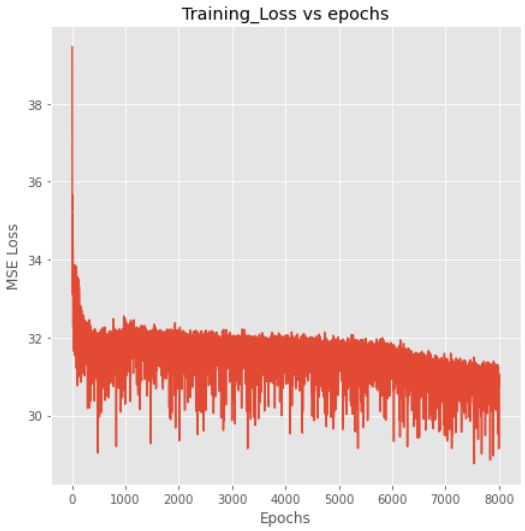

Figure 4.4: Training loss vs epochs

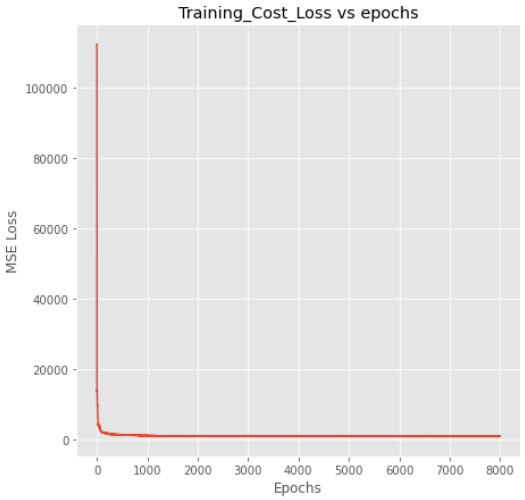

Figure 4.5: Training cost loss vs epochs

where,

	$\gamma = 10^8$	$\gamma = 10^5$	$\gamma = 10^2$	$\gamma = 10^8$	$\gamma = 10^5$	$\gamma = 10^2$
	Learning rate=0.1			Learning rate=0.01		
ATE	430.406	314.222	296.615	466.354	466.109	467.644
ϵ_{ATE}	45.268	70.915	88.522	81.216	80.972	82.506
ATT	430.727	310.735	297.297	466.651	465.662	466.213
ϵ_{ATT}	47.585	72.405	85.844	83.510	82.520	83.072
ATC	430.074	317.832	295.909	466.047	466.573	469.125
ϵ_{ATC}	42.869	69.373	91.295	78.842	79.368	81.920
RMSE	107.304	139.955	126.832	86.239	86.254	85.850
$\sqrt{PEHE}$	107.012	126.013	92.494	95.408	90.062	91.648

Table 4.2: CFMedNet results for Medical cost on semi-synthetic ACIC dataset

	$\gamma = 10^8$	$\gamma = 10^5$	$\gamma = 10^2$	$\gamma = 10^8$	$\gamma = 10^5$	$\gamma = 10^2$
	Learning rate=0.1			Learning rate=0.01		
ATE	3.063	1.918	3.556	3.397	1.626	1.258
ϵ_{ATE}	1.096	1.121	0.515	0.643	1.414	1.782
ATT	3.893	1.889	3.588	1.236	1.487	1.194
ϵ_{ATT}	1.169	0.774	0.924	1.418	1.176	1.469
ATC	2.707	1.948	3.522	1.881	1.770	1.324
ϵ_{ATC}	1.126	1.481	0.092	1.987	1.659	2.106
RMSE	5.090	5.523	5.532	5.6527	5.409	5.329
$\sqrt{PEHE}$	6.252	6.706	6.788	6.531	6.748	7.144

Table 4.3: CFMedNet results for Treatment outcome on semi-synthetic ACIC dataset

- ATE: Average Treatment Effect
- ϵ_{ATE} : Error in Average Treatment Effect
- ATT : Average Treatment Effect on Treated Population
- ϵ_{ATT} : Error in Average Treatment Effect on Treated Population
- ATC : Average Treatment Effect On Control Population
- ϵ_{ATC} : Error in Average Treatment Effect on Control Population
- RMSE : Root Mean Square Error
- $\sqrt{PEHE}$: Error in Precision in Estimation of Heterogeneous Effect

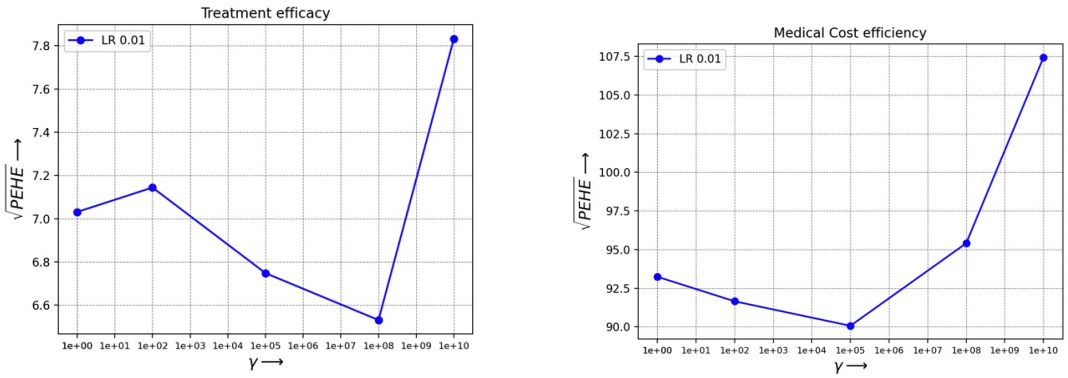

Figure 4.6: $\sqrt{PEHE}$ vs γ

4.3 XAI Decision Models

4.3.1 GMM-LIME

GMM-LIME (Gaussian Mixture Model-LIME) and LIME (Local Interpretable Model-agnostic Explanations) are compared for a specific instance in our dataset. Upon analyzing the results on the test instance using LIME and GMM-LIME, we observed notable differences in their local predictions and actual values. LIME's local prediction for Treatment efficacy for Y_f was 10.022 while the actual value is 6.084 similarly LIME's local prediction for Treatment efficacy for Y_cf was and 12.425 while the actual is 10.514 (Fig.4.7), whereas GMM-LIME's local prediction for Treatment efficacy for Y_f was 9.328 and for Y_cf was 11.482 (Fig.4.8).It is evident that GMM-LIME results are closer to the actual values.

Figure 4.7: LIME results (Random Sampling) for Treatment efficacy for Y_f and Y_{cf} respectively

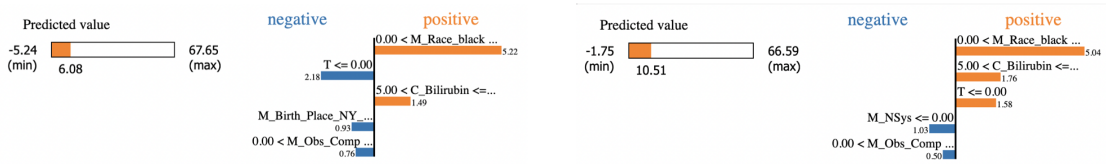

Figure 4.8: GMM-LIME (GMM sampling) for Treatment efficacy for Y_f and Y_{cf} respectively

Similarly, in the case of Medical cost efficiency, LIME yielded local predictions for C_f and C_{cf} as 118.703 and 502.136 (Fig.4.9) while the actual values were 105.345 and 506.095 respectively, while GMM-LIME provided 112.244 and 505.271 as local predictions for C_f and C_{cf} respectively (Fig.4.10).

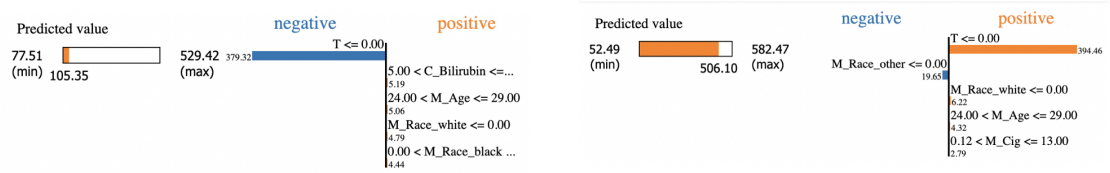

Figure 4.9: LIME results (Random Sampling) for Medical cost efficiency for C_f and C_{cf} respectively

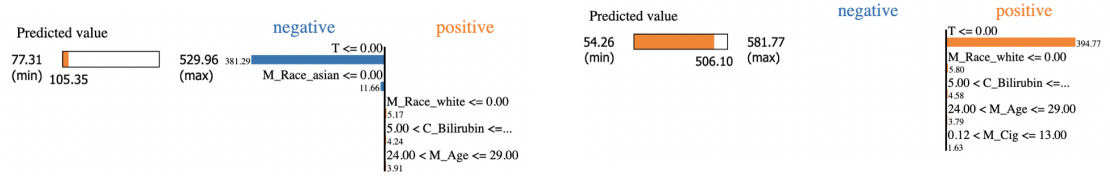

Figure 4.10: GMM-LIME results (GMM sampling) for Medical cost efficiency for C_f and C_{cf} respectively

Moving on to the weighted optimized output, combining Y_f and C_f versus Y_{cf} and C_{cf} (factuals vs counterfactuals), LIME's local predictions were 0.152 compared to actual values of 0.113 for factual outcome whereas the local predictions was 0.475 compared to the counterfactual outcome actual value as 0.420 (Fig.4.11). Further GMM-LIME resulted the local predict as 0.143 for the factual output whereas it gave 0.442 for the counterfactual output (Fig.4.12). These findings highlight the improved performance of GMM-LIME over LIME in terms of producing predictions that are closer to the actual values in our dataset.

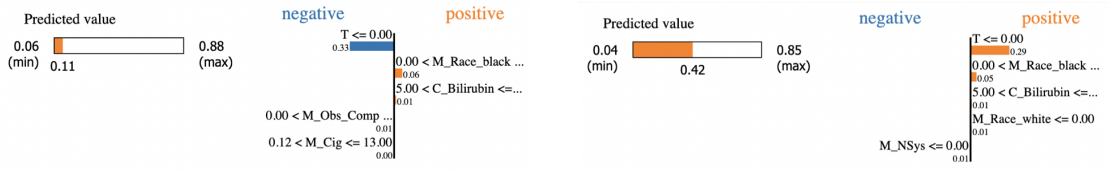

Figure 4.11: LIME results (Random Sampling) for weighted optimized output i.e. combining Treatment efficacy and Medical cost efficiency (factuals vs counterfactuals)

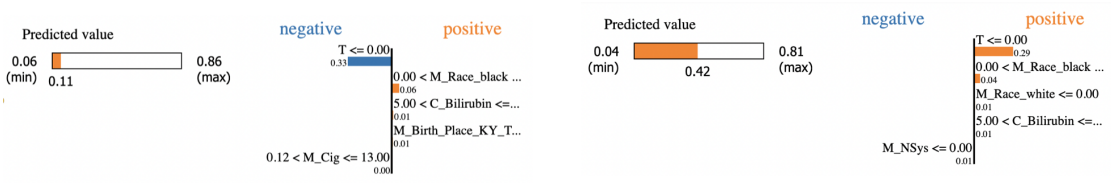

Figure 4.12: GMM-LIME results (GMM sampling) for weighted optimized output i.e. combining Treatment efficacy and Medical cost efficiency (factuals vs counterfactuals)

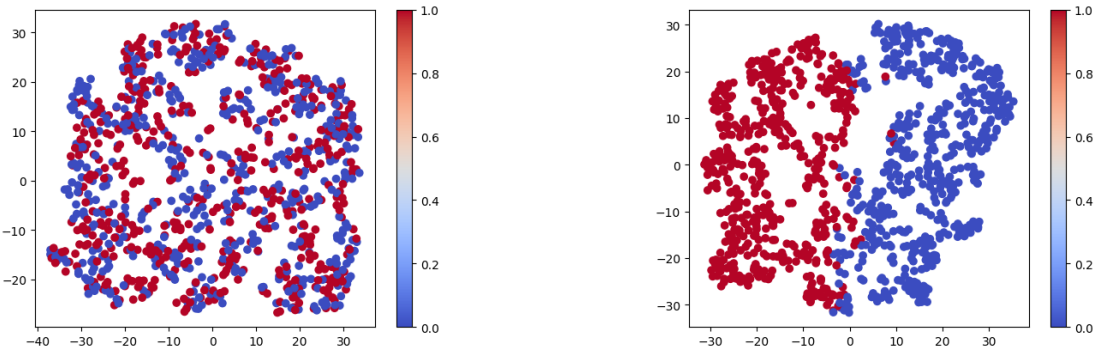

Figure 4.13: Visualizing 1000 test instances through Random sampling vs GMM Sampling for $T = 0$ and $T = 1$

4.3.2 SHAP (SHapley Additive exPlanations)

The SHAP (SHapley Additive exPlanations) results demonstrate consistency in our approach of weighted optimized output, particularly regarding the features that contribute to the final value. The SHAP framework provides insights into feature importance and their individual contributions to the model's predictions. By analyzing the SHAP values, we can observe the consistent patterns in feature importance across different instances. The features that consistently show high SHAP values indicate their significant impact on the final output. This consistency in feature importance strengthens the validity of our weighted optimized output approach. Additionally, examining the SHAP values helps us understand the direction and magnitude of each feature's influence. Positive SHAP values indicate a positive contribution to the output, while negative values imply a negative impact. Overall, the consistency observed in the SHAP results reinforces the effectiveness and reliability of our weighted optimized output approach, as it aligns with the feature contributions consistently highlighted by the SHAP framework.

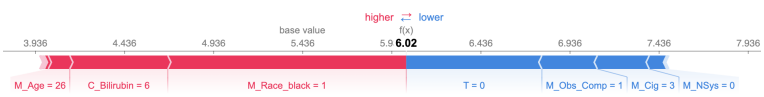

Figure 4.14: SHAP results on Treatment efficacy (Y_f)

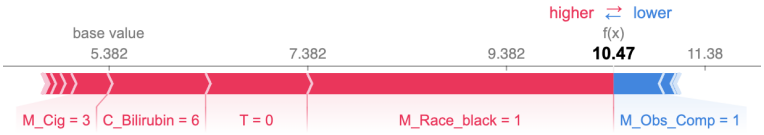

Figure 4.15: SHAP results on Treatment efficacy (Y_{cf})

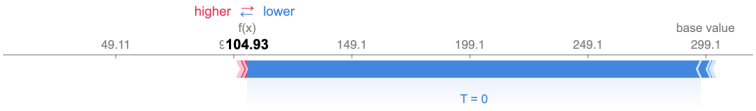

Figure 4.16: SHAP results on Medical cost efficiency (C_f)

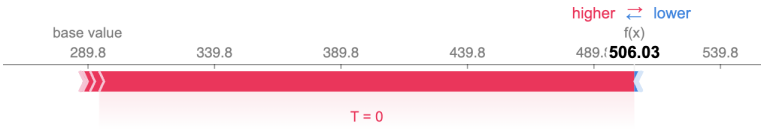

Figure 4.17: SHAP results on Medical cost efficiency (C_{cf})

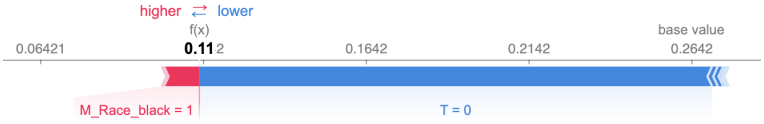

Figure 4.18: SHAP results on weighted optimised output (i.e. Y_f and C_f)

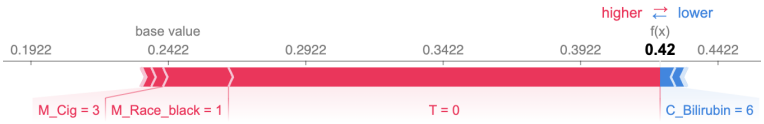

Figure 4.19: SHAP results on weighted optimised output (i.e. Y_{cf} and C_{cf})

4.4 Measuring XAI Effectiveness

4.4.1 Qualitative Results

User Experience

User experience plays a crucial role in assessing the utility of an explanation. Since there is a qualitative way to measure, it involves collecting feedback from the domain experts about their understanding of the explanation, whether it met their expectations, and whether it provided valuable insights that were previously unknown or misunderstood.

User Utility

It refers to the usefulness and value of the explanation to the user. It assesses whether the provided explanation helps the user gain insights into why a specific treatment was chosen over an alternative, and whether it aids in understanding the factors driving the procedure by which decisions are made utilizing the counterfactual model.

Task Execution

By analyzing artificial decision tasks and the user's comprehension, we can assess the impact of the explanation on their decision and task performance. Understanding the reasoning behind a decision enhances the user's comprehension of the intelligent system's task performance. If necessary, the explanations can guide modifications to the logic of autonomous systems based on the analyzed task performance, leading to improved system performance.

Trust Assessment

Through a comprehensive assessment of clarity, utility, and their impact on decision-making, users can develop a proper understanding of the system. This understanding enables them to appropriately utilize the system, ensuring trust is maintained at a reasonable level. By knowing when and how to use the system effectively, users can make informed decisions and rely on its capabilities to achieve desired outcomes.

4.4.2 Quantitaive Results

Stability in repeated explanations

To assess the variability of explanations over several executions, we ran GMM-LIME along with the baseline method LIME. We performed these runs using an incrementally increasing number of surrogate samples: 500, 1000, 1500, 2000, and 2500. We collected 20 successive explanations for 10 randomly chosen index samples for our dataset. The consistency in explanations across different runs, specifically the i^th and j^th runs, was evaluated using Jaccard's Distance (J) (Saini and Prasad (2022),Zafar and Khan (2019)). Computationally, Jaccard's Distance between two feature sets, X_i and

X_j, is defined as:

$$J(X_i, X_j) = \frac{|X_i \cap X_j|}{|X_i \cup X_j|}$$

In this context, X_i and X_j represent the sets of top-5 features for the i^{th} and j^{th} iterations, respectively. The interpretation of Jaccard's Distance is straightforward: a value of 1 suggests identical feature sets (perfect consistency), while a value of 0 indicates no common features (complete inconsistency). Therefore, a highly consistent explainer module would result in a higher value of this metric. We calculated the average Jaccard's Distance over all combinations of iterations and the selected index samples. The results, visualized in Fig.(4.20), averaged over the varying surrogate sample sizes (500, 1000, 1500, 2000, 2500), demonstrated a discernible pattern. The outcome of this comparison was telling. Regardless of the surrogate sample size, GMM-LIME consistently outperformed LIME. This result reinforces the efficacy of the GMM-LIME, affirming its superiority in providing stable and faithful explanations.

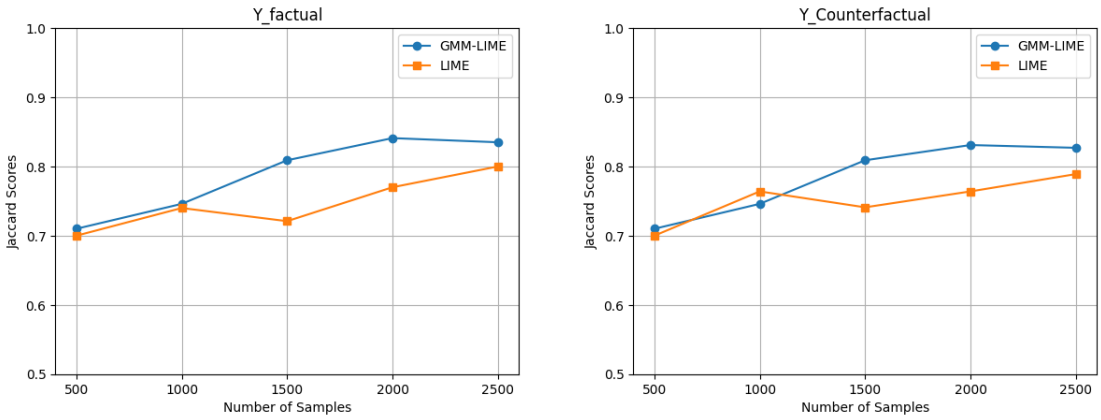

Figure 4.20: Comparing average Jaccard scores for 10 random test instances using varying numbers of GMM-LIME surrogate samples with a state-of-the-art method.

Prediction Closeness to Actual values

We calculate the error between the local predictions and the actual values of the instances of interest. This allows us to assess the accuracy and performance of the interpretable model generated by LIME and GMM-LIME. Similarly, to calculate the local prediction error using SHAP, we first obtain the SHAP values for the instances of interest. These values represent the contributions of each feature towards the prediction for a particular instance. Using the SHAP values, we reconstruct the predicted values by summing the contributions of each feature. Once we have the reconstructed predicted values, we compare them with the actual values of the instances to calculate the prediction error. The error between the local predictions and the actual values can be measured using various metrics depending on the nature of the task. Here, for regression task, we use the root mean square error (RMSE). This metric provides a quantitative measure of the discrepancy between the predicted values and the ground truth values Table(4.4).

The Root Mean Square Error (RMSE) is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where n is the number of samples, y_i represents the actual values, and $\hat{y}_i$ represents the predicted values.

XAI Model	$RMSE_{O_f}$	$RMSE_{O_{cf}}$
LIME	0.058	0.062
GMM-LIME	0.011	0.019
SHAP	0.048	0.054

Table 4.4: RMSE scores for factual (O_f) and counterfactual (O_{cf}) outcomes

where $RMSE_{O_f}$ is the RMSE score in case of factual outcome whereas $RMSE_{O_{cf}}$ is the RMSE score in case of counterfactual outcome.

4.5 Conclusion

We have developed an innovative counterfactual inference framework, CFMedNet, to simultaneously measure the effectiveness of a treatment and its cost efficiency. We assessed this model using semi-synthetic data, yielding promising results. CFMedNet produced semi-synthetic medical costs ranging from 49 to 619, with a root mean square error as low as 85.850. Furthermore, the $\sqrt{PEHE}$ reached a lower boundary of 90.062 for semi-synthetic tests, indicating that our methods are proficient at determining the treatment impact in a multi-outcome prediction challenge, where one of the outcomes is medical cost. As far as we know, no prior algorithm successfully assessed treatment efficacy and cost efficiency for each alternative therapy, evaluating the impact of each intervention on all desired outcomes. Our research fills this void by providing a solution.

We also presented an explanation framework based on Gaussian Mixture Model (GMM) sampling for our CFMedNet model, called GMM-LIME. GMM-LIME is an enhancement of the pre-existing LIME framework that introduces a new sampling method using GMM. It alters the reference sample to generate surrogate samples through the oversampling of minority class instances using GMMs, aiming for a balanced dataset. Feature based explanations are obtained through the same. Furthermore, considering multi-output networks like ours, we proposed a weighted optimization algorithm that can effectively explain multi-output situations based on established explainers such as LIME and SHAP.

4.6 Future Work

This research has the potential for further extension in various aspects.

- In this study, we made the assumption of a binary treatment scenario. However, there is potential for extending this work to encompass cases with multiple treatments.
- In our research, we focused on tabular data as the input. However, there is an opportunity to extend this work to incorporate multimodal models, which involve multiple types of data such as text, images, and numerical features.

REFERENCES

1. **Arrieta, A. B.**, **N. Díaz-Rodríguez**, **J. Del Ser**, **A. Bennetot**, **S. Tabik**, **A. Barbado**, **S. García**, **S. Gil-López**, **D. Molina**, **R. Benjamins**, *et al.* (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, **58**, 82–115.

2. **Athey, S.**, Machine learning and causal inference for policy evaluation. *In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015.

3. **Belle, V.** and **I. Papantonis** (2021). Principles and practice of explainable machine learning. *Frontiers in big Data*, 39.

4. **Brenan, M.** (2022). "record high in u.s. put off medical care due to cost in 2022". URL `https://news.gallup.com/poll/468053/record-high-put-off-medical-care-due-cost-2022.aspx`.

5. **Doran, D.**, **S. Schulz**, and **T. R. Besold** (2017). What does explainable ai really mean? a new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.

6. **Gilpin, L. H.**, **D. Bau**, **B. Z. Yuan**, **A. Bajwa**, **M. Specter**, and **L. Kagal**, Explaining explanations: An overview of interpretability of machine learning. *In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018.

7. **González Ramírez, J.** and **M. Kilic**, *Applied Causal Analysis (with R)*. Chapman and Hall/CRC, Boca Raton, FL, 2019. ISBN 978-1-138-57592-8.

8. **Greenland, S.** and **J. M. Robins** (2009). Identifiability, exchangeability and confounding revisited. *Epidemiologic Perspectives & Innovations*, **6**(1), 1–9.

9. **Gross**, **R. T.**, and **others** (1993). Infant health and development program (IHDP): Enhancing the outcomes of low birth weight, premature infants in the united states, 1985-1988.

10. **Guidotti, R.**, **A. Monreale**, **S. Ruggieri**, **D. Pedreschi**, **F. Turini**, and **F. Giannotti** (2018). Local rule-based explanations of black box decision systems.

11. **Gunning, D.**, **M. Stefik**, **J. Choi**, **T. Miller**, **S. Stumpf**, and **G.-Z. Yang** (2019). Xai—explainable artificial intelligence. *Science robotics*, **4**(37), eaay7120.

12. **Hagras, H.** (2018). Toward human-understandable, explainable ai. *Computer*, **51**(9), 28–36.

13. **Hahn, P.**, **V. Dorie**, and **J. Murray** (2019). Atlantic causal inference conference (acic) data analysis challenge 2017.

14. **Hill, J. L.** (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, **20**(1), 217–240. URL `https://doi.org/10.1198/jcgs.2010.08162`.

15. **Holzinger, A.**, From machine learning to explainable ai. *In 2018 world symposium on digital intelligence for systems and machines (DISA)*. IEEE, 2018.

16. **Höfler, M.** (2005). Causal inference based on counterfactuals. *BMC Medical Research Methodology*, **5**, 28.

17. **Imbens, G. W.** (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, **86**(1), 4–29.

18. **Imbens, G. W.** and **D. B. Rubin**, *Statistical Tools for Causal Inference*. Cambridge University Press, New York, NY, 2015. ISBN 978-1-107-02428-8.

19. **Jin, Y.**, **Z. Ren**, and **E. J. Candès** (2023). Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, **120**(6), e2214889120.

20. **Kaur, P.**, **A. Polyzou**, and **G. Karypis**, Causal inference in higher education. *In Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale*. ACM, 2019. URL `https://doi.org/10.1145%2F3330430.3333663`.

21. **Kreif, N.** and **K. DiazOrdaz** (2019). Machine learning in policy evaluation: new tools for causal inference.

22. **Kwon, H.** and **Y. Kim** (2021). Prediction of patient's medical expenses using gated recurrent unit networks. *IEEE Journal of Biomedical and Health Informatics*, **25**(3), 693–704.

23. **Lundberg, S.** and **S.-I. Lee** (2017). A unified approach to interpreting model predictions.

24. **Ma, F.**, **R. Chitta**, **J. Zhou**, **Q. You**, **T. Sun**, and **J. Gao** (2019). Deep learning for individual patient cost prediction in healthcare. *arXiv preprint arXiv:1911.05580*.

25. **MasaAsami** (2022). Introduction to CFR. `https://github.com/MasaAsami/introduction_to_CFR`. 2022.

26. **Mateo, J.**, **J. Rius-Peris**, **A. Maraña-Pérez**, **A. Valiente-Armero**, and **A. Torres** (2021). Extreme gradient boosting machine learning method for predicting medical treatment in patients with acute bronchiolitis. *Biocybernetics and Biomedical Engineering*, **41**(2), 792–801. ISSN 0208-5216. URL `https://www.sciencedirect.com/science/article/pii/S020852162100053X`.

27. **Molnar, C.**, *Interpretable Machine Learning*. 2022, 2 edition. URL `https://christophm.github.io/interpretable-ml-book`.

28. **Murnane, R. J.** and **J. B. Willett**, *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press, 2011.

29. **Nanavati, P.** and **R. Prasad** (2023). Climax: An exploration of classifier-based contrastive explanations.

30. **Neal, B.** (2020). Introduction to causal inference from a machine learning perspective. URL `https://www.bradyneal.com/Introduction_to_Causal_Inference-Dec17_2020-Neal.pdf`.

31. **Pearl, J.** (2010). An introduction to causal inference. *International Journal of Biostatistics*, **6**(2), Article 7.

32. **Pedregosa, F.**, **G. Varoquaux**, **A. Gramfort**, **V. Michel**, **B. Thirion**, **O. Grisel**, **M. Blondel**, **A. Müller**, **J. Nothman**, **G. Louppe**, **P. Prettenhofer**, **R. Weiss**, **V. Dubourg**, **J. Vanderplas**, **A. Passos**, **D. Cournapeau**, **M. Brucher**, **M. Perrot**, and **Édouard Duchesnay** (2018). Scikit-learn: Machine learning in python.

33. **Prosperi, M.**, **Y. Guo**, **M. Sperrin**, **J. S. Koopman**, **J. S. Min**, **X. He**, **S. Rich**, **M. Wang**, **I. E. Buchan**, and **J. Bian** (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence*, **2**(7), 369–375.

34. **Ribeiro, M. T.**, **S. Singh**, and **C. Guestrin** (2016). "why should i trust you?": Explaining the predictions of any classifier.

35. **Roscher, R.**, **B. Bohn**, **M. F. Duarte**, and **J. Garcke** (2020). Explainable machine learning for scientific insights and discoveries. *Ieee Access*, **8**, 42200–42216.

36. **Rubin, D. B.** (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, **100**(469), 322–331.

37. **Saini, A.** and **R. Prasad** (2022). Select wisely and explain: Active learning and probabilistic local post-hoc explainability.

38. **Shalit, U.**, **F. D. Johansson**, and **D. Sontag** (2017). Estimating individual treatment effect: generalization bounds and algorithms.

39. **Shi, C.**, **D. M. Blei**, and **V. Veitch** (2019). Adapting neural networks for the estimation of treatment effects.

40. **Shi, J.** and **B. Norgeot** (2022). Learning causal effects from observational data in healthcare: A review and summary. *Frontiers in Medicine*, **9**, 864882.

41. **Sriperumbudur, B. K.**, **K. Fukumizu**, **A. Gretton**, **B. Schölkopf**, and **G. R. G. Lanckriet** (2012). On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, **6**(none), 1550 – 1599. URL `https://doi.org/10.1214/12-EJS722`.

42. **Sun, Z.**, **B. Wang**, and **J. Li** (2020). Costnet: Deep learning framework for predicting healthcare costs. *arXiv preprint arXiv:2002.05718*.

43. **Tong, L.-L.**, **J.-B. Gu**, **J.-J. Li**, **G.-X. Liu**, **S.-W. Jin**, and **A.-Y. Yan** (2021). Application of bayesian network and regression method in treatment cost prediction. *BMC Medical Informatics and Decision Making*, **21**.

44. **Wang, H.**, **X. Wu**, **B. Liu**, **X. Chen**, **M. Zhu**, and **Y. Pan** (2020). Application of convolutional neural networks in the diagnosis and prediction of chronic diseases. *Journal of Medical Systems*, **44**(2).

45. **Xu, F.**, **H. Uszkoreit**, **Y. Du**, **W. Fan**, **D. Zhao**, and **J. Zhu**, Explainable ai: A brief survey on history, research areas, approaches and challenges. *In Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer, 2019.

46. **Zafar, M. R.** and **N. M. Khan** (2019). Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems.