Estimating Probabilities of Causation with Machine Learning Models

Shuai Wang[1]   **Ang Li**[1]

[1]Dept. of Computer Science
Florida State University
Tallahassee, FL, USA

Abstract

Probabilities of causation play a crucial role in modern decision-making. This paper addresses the challenge of predicting probabilities of causation for subpopulations with insufficient data using machine learning models. Tian and Pearl first defined and derived tight bounds for three fundamental probabilities of causation: the probability of necessity and sufficiency (PNS), the probability of sufficiency (PS), and the probability of necessity (PN). However, estimating these probabilities requires both experimental and observational distributions specific to each subpopulation, which are often unavailable or impractical to obtain with limited population-level data. We assume that the probabilities of causation for each subpopulation are determined by its characteristics. To estimate these probabilities for subpopulations with insufficient data, we propose using machine learning models that draw insights from subpopulations with sufficient data. Our evaluation of multiple machine learning models indicates that, given sufficient population-level data and an appropriate choice of machine learning model and activation function, PNS can be effectively predicted. Through simulation studies, we show that our multilayer perceptron (MLP) model with the Mish activation function achieves a mean absolute error (MAE) of approximately 0.02 in predicting PNS for 32,768 subpopulations using data from around 2,000 subpopulations.

1 INTRODUCTION

Understanding causal relationships and estimating probabilities of causation are crucial in fields such as healthcare, policy evaluation, and economics [Pearl, 2009, Imbens and Rubin, 2015, Heckman and Pinto, 2015]. Unlike correlation-based methods, causal inference enables decision-makers to determine whether an action or intervention directly leads to a desired outcome. This is particularly essential in personalized medicine, where accurately assessing treatment effects ensures both efficacy and safety [Mueller and Pearl, 2023]. Moreover, causal reasoning enhances machine learning applications by improving accuracy [Li et al., 2020], interpretability, and fairness [Plecko and Bareinboim, 2022] in automated decision-making. Despite its broad significance, estimating probabilities of causation remains challenging due to data limitations. In this paper, we address this challenge by leveraging machine learning techniques to predict probabilities of causation for subpopulations with insufficient data.

The study of probabilities of causation began around 2000 when Pearl [1999] first defined three fundamental probabilities—PNS, PS, and PN—within Structural Causal Models [Galles and Pearl, 1998, Halpern, 2000, Pearl, 2009]. Subsequently, Tian and Pearl [2000] derived tight bounds for these probabilities using Balke's linear programming [Balke, 1995], incorporating both observational and experimental data. Nearly two decades later, Li and Pearl [2019] formally proved these bounds and introduced the unit selection model, a decision-making framework based on their linear combination. More recently, Li and Pearl [2024b] extended the definitions and bounds to a more general form. Additionally, Mueller et al. [2022], as well as Dawid et al. [2017], demonstrated that these bounds could be further refined given specific causal structures.

However, any above estimation of the probabilities of causation requires both observational and experimental data. Additionally, estimating (sub)populations, based on Li et al. [2022c]'s suggestions, requires approximately $1,300$ entries of both data types for each (sub)population, making the process impractical. Li et al. [2022a,b] demonstrated the potential of machine learning models to achieve accurate estimations for (sub)populations. In this research, we select five diverse machine learning models based on the characteristics of the probabilities of causation. We then evaluate

their performance in accomplishing the task.

1.1 CONTRIBUTIONS

Despite the extensive theoretical research on probabilities of causation, practical estimation methods have remained unexplored. Our work provides the first systematic approach to predicting probabilities of causation using machine learning. Specifically, we make the following contributions:

- **First Machine Learning Pipeline for Predicting Probabilities of Causation:** We propose a novel machine learning framework to estimate the bounds of PNS, PS, and PN, filling a critical gap between theoretical causal inference and practical applications.
- **First Accurate Machine Learning Model for PNS Prediction:** We demonstrate that a MLP can accurately predict PNS, proving that machine learning is a feasible and effective tool for estimating probabilities of causation.
- **First Dataset for PNS Bound Prediction:** We construct and release the first synthetic dataset specifically designed to evaluate machine learning models for estimating PNS, providing a foundation for future research in this area.

To the best of our knowledge, no prior work has applied machine learning to the problem of predicting probabilities of causation. Our study establishes a new research direction by bridging causal inference and machine learning for practical estimation tasks.

The remainder of the paper is structured as follows: first, we review key causal inference concepts to provide necessary context. Next, we introduce the model and dataset used in our study. Finally, we present our five machine learning models developed for the task. All code for data generation and machine learning models is included in the appendix.

2 PRELIMINARIES

In this section, we review the fundamental concepts of causal inference necessary for understanding the rest of the paper. We begin by discussing the definitions of PNS, PS, and PN as introduced by Pearl [1999], followed by the definitions of identifiability and the conditions required to identify PNS, PS, and PN [Tian and Pearl, 2000]. Additionally, we examine the tight bounds of PNS, PS, and PN in cases where they are unidentifiable [Tian and Pearl, 2000]. Readers already familiar with these concepts may skip this section.

Similar to the works mentioned above, we adopt the causal language of Structural Causal Models (SCMs) [Galles and Pearl, 1998, Halpern, 2000]. In this framework, the counterfactual statement "Variable Y would have the value y had

X been x" is denoted as $Y_x = y$, abbreviated as y_x. We consider two types of data: experimental data, expressed as causal effects $P(y_x)$, and observational data, represented by the joint probability function $P(x, y)$. Unless otherwise specified, we assume X and Y are binary variables in a causal model M, with x and y denoting the propositions $X = $ true and $Y = $ true, respectively, and x' and y' representing their complements. For simplicity, we focus on binary variables; extensions to multi-valued cases are discussed by Pearl [2009] (p. 286, footnote 5) and Li and Pearl [2024a].

First, the definitions of three basic probabilities of causation defined using SCM are as follow [Pearl, 1999]:

Definition 1 (Probability of necessity (PN)). Let X and Y be two binary variables in a causal model M, let x and y stand for the propositions $X = true$ and $Y = true$, respectively, and x' and y' for their complements. The probability of necessity is defined as the expression

$$
\begin{aligned}
\text{PN} &\triangleq P(Y_{X=false} = false | X = true, Y = true) \\
&\triangleq P(y'_{x'} | x, y)
\end{aligned}
$$

Definition 2 (Probability of sufficiency (PS)). Let X and Y be two binary variables in a causal model M, let x and y stand for the propositions $X = true$ and $Y = true$, respectively, and x' and y' for their complements. The probability of sufficiency is defined as the expression

$$
\begin{aligned}
\text{PS} &\triangleq P(Y_{X=true} = true | X = false, Y = false) \\
&\triangleq P(y_x | x', y')
\end{aligned}
$$

Definition 3 (Probability of necessity and sufficiency (PNS)). Let X and Y be two binary variables in a causal model M, let x and y stand for the propositions $X = true$ and $Y = true$, respectively, and x' and y' for their complements. The probability of necessity and sufficiency is defined as the expression

$$
\begin{aligned}
\text{PNS} &\triangleq P(Y_{X=true} = true, Y_{X=false} = false) \\
&\triangleq P(y_x, y'_{x'})
\end{aligned}
$$

Then, we review the identification conditions for PNS, PS, and PN [Tian and Pearl, 2000].

Definition 4. (Monotonicity) A Variable Y is said to be monotonic relative to variable X in a causal model M iff

$$
y'_x \wedge y_{x'} = \text{false}.
$$

Theorem 5. *If Y is monotonic relative to X, then PNS, PN, and PS are all identifiable, and*

$$
\begin{aligned}
PNS &= P(y_x) - P(y_{x'}), \\
PN &= \frac{P(y) - P(y_{x'})}{P(x, y)}, \\
PS &= \frac{P(y_x) - P(y)}{P(x', y')}.
\end{aligned}
$$

If PNS, PN, and PS are not identifiable, informative bounds are given by Tian and Pearl [2000].

$$\max \left\{ \begin{array}{c} 0, \\ P(y_x) - P(y_{x'}), \\ P(y) - P(y_{x'}), \\ P(y_x) - P(y) \end{array} \right\} \leq \text{PNS} \tag{1}$$

$$\min \left\{ \begin{array}{c} P(y_x), \\ P(y'_{x'}), \\ P(x,y) + P(x',y'), \\ P(y_x) - P(y_{x'})+ \\ P(x,y') + P(x',y) \end{array} \right\} \geq \text{PNS} \tag{2}$$

$$\max \left\{ \begin{array}{c} 0, \\ \frac{P(y) - P(y_{x'})}{P(x,y)} \end{array} \right\} \leq \text{PN} \tag{3}$$

$$\min \left\{ \begin{array}{c} 1, \\ \frac{P(y'_{x'}) - P(x',y')}{P(x,y)} \end{array} \right\} \geq \text{PN} \tag{4}$$

$$\max \left\{ \begin{array}{c} 0, \\ \frac{P(y') - P(y'_x)}{P(x',y')} \end{array} \right\} \leq \text{PS} \tag{5}$$

$$\min \left\{ \begin{array}{c} 1, \\ \frac{P(y_x) - P(x,y)}{P(x',y')} \end{array} \right\} \geq \text{PS} \tag{6}$$

Therefore, the primary objective of this paper is then to predict Equations 1 to 6 (i.e., the lower and upper bounds of the PNS, PS, and PN) for any (sub)populations using those with sufficient data (i.e., sufficient data to estimate the distributions $P(X, Y)$ and $P(Y_X)$.) Due to space constraints, the focus will be on the bounds of PNS (i.e., Equations 1 and 2). Unless otherwise specified, the discussion will be limited to binary treatment and effect, meaning both X and Y are binary.

3 STRUCTURAL CAUSAL MODEL

In general, the equations in SCMs are in implicitly form (e.g., $Z = f_Z(X, Y, U_Z)$). However, in order to verify the accuracy of the learned bounds of PNS, we need to explicitly define the SCM and the data-generating process to determine the true PNS value and its bounds. Followed the setup in Li et al. [2022a], we will use the following SCM.

$$\begin{cases} Z_i & = U_{Z_i} \text{ for } i \in \{1, ..., 20\}, \\ X & = f_X(M_X, U_X) \\ & = \begin{cases} 1 & \text{if } M_X + U_X > 0.5 \\ 0 & \text{otherwise,} \end{cases} \\ Y & = f_Y(X, M_Y, U_Y) \\ & = \begin{cases} 1 & \text{if } 0 < C_Y \cdot X + M_Y + U_Y < 1 \\ 1 & \text{if } 1 < C_Y \cdot X + M_Y + U_Y < 2 \\ 0 & \text{otherwise.} \end{cases} \end{cases}$$

where X, Y, Z_i are all binary, U_{Z_i}, U_X, U_Y are binary exogenous variables with Bernoulli distributions, C_Y is a constant, and M_X, M_Y are linear combinations of Z_i. The randomly generated value of C_Y, M_X, M_Y and the distributions of U_X, U_Y, U_{Z_i} for the model are provided in the appendix.

4 DATA GENERATING PROCESS

Based on the defined model, 20 binary features are considered (i.e., $Z_1, ..., Z_{20}$). We made 15 observable ($Z_1, ..., Z_{15}$) and 5 unobservable, and the exogenous variables are also unobservable, leading to 2^{15} observed subpopulations (i.e., the combination of $Z_1, ..., Z_{15}$ defined a subpopulation).

4.1 INFORMER DATA

To evaluate the learned bounds, the informer data must have access to the actual PNS bounds for each subpopulation. Given the explicit form of the SCM and the distributions of all exogenous variables, the PNS bounds, as well as the experimental and observational distributions, can be computed for each combination of the features $Z_1, \ldots, Z_{15}$ (i.e., a subpopulation) using the SCM. For detailed mathematical formulations, refer to the appendix.

4.2 SAMPLE COLLECTION

A total of $50,000,000$ experimental and $50,000,000$ observational samples were generated as follows for each sample: In both settings, the exogenous variables U_X, U_Y, and U_{Z_i} were randomly generated according to their distributions specified in Section 3. In the experimental setting, X was then assigned according to a Bernoulli(0.5) distribution, while Y and Z_i were computed using the structural functions described in Section 3. In the observational setting, X, Y, and Z_i were all determined by the structural functions. The final datasets include only the observable features $Z_1, \ldots, Z_{15}$, along with X and Y, while $Z_{16}, \ldots, Z_{20}$ were masked.

4.3 DATA FOR MACHINE LEARNING MODELS

We selected subpopulations from the 2^{15} possible groups that contained at least $1,300$ experimental and observational samples ($1,300$ based on Li et al. [2022c]'s suggestions). For these selected subpopulations, we computed the experimental and observational distributions and determined the bounds of PNS using Equations 1 and 2. These results served as the data for our machine learning models (i.e., each data entry consists of 15 features and the PNS bounds as the label.) The obtained data includes $2,054$ entries for the lower bound (LB) and $2,065$ entries for the upper bound (UB) of the PNS.

5 MACHINE LEARNING PREDICTION

To evaluate the feasibility of machine learning in predicting the bounds of the PNS, we employed five distinct machine learning models to assess their effectiveness in this task: Support Vector Machine (SVM) [Cortes, 1995], Random Forest (RF) [Ho, 1995], Gradient Boosting Decision Trees (GBDT) [Friedman, 2001], Transformer [Vaswani, 2017], and Multilayer Perceptron (MLP) [Rumelhart et al., 1986]. These models were chosen to represent a diverse range of machine learning paradigms, including kernel-based methods (SVM), ensemble learning techniques (RF and GBDT), and deep learning approaches (MLP and Transformer). This selection ensures a comprehensive evaluation of their ability to approximate causal quantities across different settings. A detailed pipeline is illustrated in Figure 1.

5.1 SUPPORT VECTOR MACHINE

Support Vector Machines (SVM) [Cortes, 1995] are widely used and well-established supervised learning models. Given their strengths, we selected Support Vector Regression (SVR), a variant of SVM, as the first model for our experiments. To effectively capture complex patterns, we employed the Radial Basis Function (RBF) kernel to map the data into a high-dimensional feature space.

Key hyperparameters include the penalty parameter (C), the insensitive loss threshold (ϵ), and the kernel coefficient (γ). The parameter C controls the trade-off between model complexity and error tolerance, where larger values may lead to overfitting. The threshold ϵ defines the margin of tolerance for errors, while γ determines the influence range of individual data points.

A two-stage hyperparameter tuning strategy was adopted. First, Randomized Search [Bergstra and Bengio, 2012] was employed to efficiently explore the parameter space and identify promising ranges. Then, Grid Search [Bergstra and Bengio, 2012] was used to fine-tune parameters within these ranges. Cross-validation ensured robust generalization throughout the tuning process.

Finally, the mean squared error (MSE) and mean absolute error (MAE) values of the SVR model can be found in Table 1. Confusion matrices are presented in Figures 3a and 3d, while Figures 2a and 2e provide a clearer comparison with the true PNS bounds. For the prediction of the lower bound, SVR demonstrates reasonable effectiveness; however, for the more complex upper bound, it exhibits a significant decline in accuracy.

5.2 RANDOM FOREST

Random Forests (RF) [Ho, 1995] are a widely used ensemble learning method for classification, regression, and other predictive tasks. The core idea behind RF is to construct multiple decision trees during training and aggregate their outputs to enhance overall performance. As an ensemble model, RF exhibits strong robustness, motivating us to assess its effectiveness in predicting PNS bounds.

Key hyperparameters of RF include the number of trees ($n_{\text{estimators}}$), maximum tree depth (max_depth), minimum samples required to split a node (min_samples_split), and the number of features considered for splitting (max_features). Increasing $n_{\text{estimators}}$ generally improves performance but at the expense of higher computational costs. The parameters max_depth, min_samples_split, and max_features regulate tree complexity, balancing bias-variance trade-offs.

For hyperparameter optimization, we employed a two-stage tuning strategy similar to that used for SVM. Table 1 also presents RF's MAE and MSE results, while Figures 3b and 3e show its confusion matrices. A more direct comparison with true PNS bounds is provided in Figures 2b and 2f. RF performs comparably to SVM on the lower bound but exhibits significantly higher accuracy on the upper bound.

5.3 GRADIENT BOOSTING DECISION TREES

Gradient Boosting Decision Trees (GBDT) [Friedman, 2001] is an ensemble learning method that builds models sequentially, with each new tree correcting the errors of its predecessors. Unlike traditional boosting, GBDT optimizes pseudo-residuals, enabling flexible loss function optimization. Simple decision trees serve as weak learners, allowing GBDT to effectively capture complex data patterns.

Key hyperparameters include the number of trees ($n_{\text{estimators}}$), learning rate (learning_rate), maximum tree depth (max_depth), and subsample ratio (subsample). The learning rate determines each tree's contribution, while $n_{\text{estimators}}$ and max_depth regulate model complexity and performance.

Following the approach used for SVM and RF, we applied a two-stage tuning strategy. Again, table 1 presents the MSE and MAE results, while Figures 3c and 3f show the confusion matrices. A more direct comparison with true PNS bounds is provided in Figures 2c and 2g. GBDT demonstrates moderate performance on both the lower and upper bounds.

5.4 TRANSFORMER

The Transformer [Vaswani, 2017], originally developed for Natural Language Processing, has expanded into Computer Vision and become a cornerstone of deep learning, particularly with the rise of large language models. Given its significant impact, this study also evaluates the Transformer for testing.
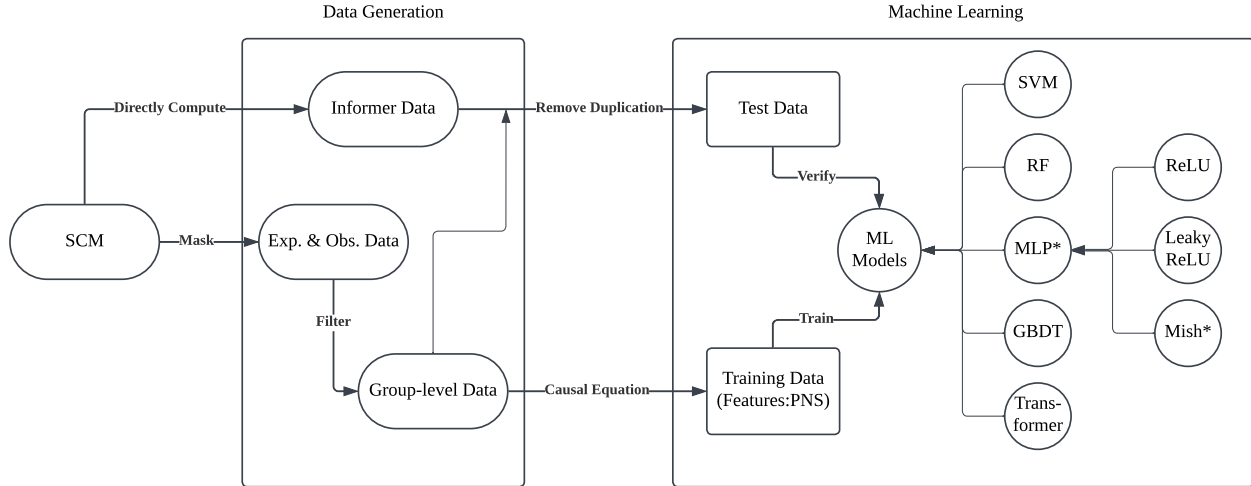
Figure 1: Framework for Causal Data Generation and Machine Learning Prediction.

(a) SVM (Lower bound) (b) RF (Lower bound) (c) GBDT (Lower bound) (d) Transformer (Lower bound)

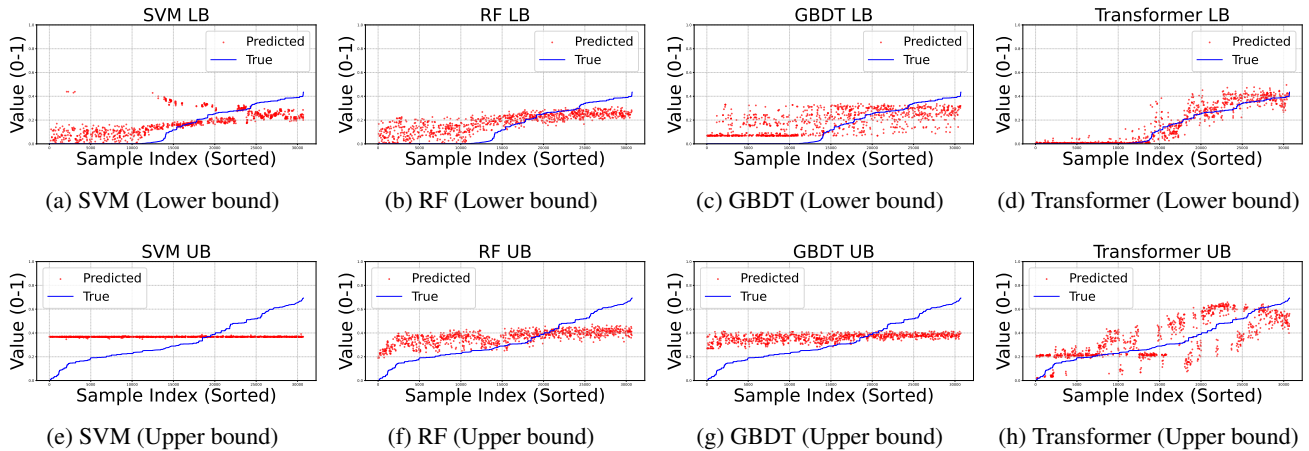(e) SVM (Upper bound) (f) RF (Upper bound) (g) GBDT (Upper bound) (h) Transformer (Upper bound)

Figure 2: Comparison of true and predicted values across different models for both lower and upper bounds.

The model architecture begins with an input layer processing 15-dimensional feature vectors, followed by a linear embedding layer that projects inputs into a 64-dimensional space. Positional encoding is applied to retain feature order information, and two Transformer encoder layers with four attention heads each capture complex feature interactions. The final output is generated through a fully connected layer with a Sigmoid activation function, ensuring predictions remain within the range $[0, 1]$. Key hyperparameters include an embedding dimension of 64, four attention heads, two encoder layers, and a dropout rate of 0.1.

Similarly, table 1 presents the MSE and MAE results, while Figures 3g and 3h show the confusion matrices. A direct comparison with true PNS bounds is provided in Figures 2d and 2h. The Transformer demonstrates strong performance on the lower bound and moderate performance on the upper bound.

5.5 MULTILAYER PERCEPTRON

MLP [Rumelhart et al., 1986] consists of an input layer, one or more hidden layers, and an output layer. With appropriate activation functions, it can effectively model both linear and nonlinear relationships. As a fundamental structure in deep learning, MLP holds significant representativeness, motivating its inclusion in our experiments.

A key consideration for MLP is the choice of activation function, particularly for predicting the lower bound. Since the lower bound of PNS cannot be negative, we initially selected the ReLU [Nair and Hinton, 2010] activation function (7). However, ReLU can lead to the loss of negatively correlated features, prompting us to adopt LeakyReLU [Maas et al., 2013] (8) as a complementary solution. Furthermore, given the considerable number of zero values in the data, the non-differentiability of ReLU and LeakyReLU at $s = 0$ imposes limitations on backpropagation. To address this,

(a) SVM Lower bound.　　(b) RF Lower bound.　　(c) GBDT Lower bound.

(d) SVM Upper bound.　　(e) RF Upper bound.　　(f) GBDT Upper bound.

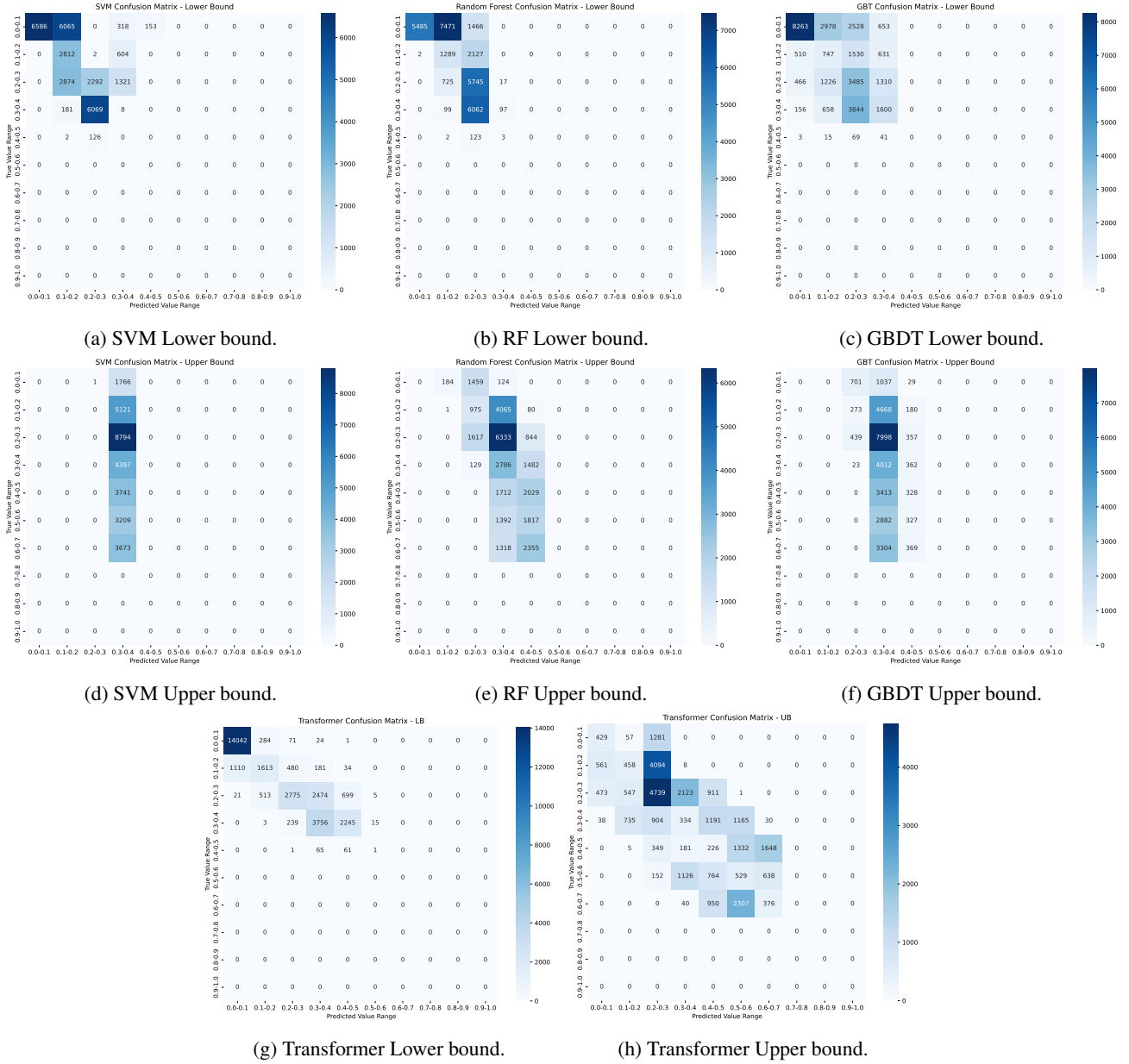(g) Transformer Lower bound.　　(h) Transformer Upper bound.

Figure 3: Confusion matrices of SVM, RF, GBDT, and Transformer models.

we proposed using Mish [Misra, 2019] (9) as an alternative activation function. The corresponding equations are:

$$\text{ReLU}(s) = \max(0, s) \tag{7}$$

$$\text{LeakyReLU}(s) = \begin{cases} s, & \text{if } s \geq 0 \\ \alpha s, & \text{if } s < 0 \end{cases} \tag{8}$$

$$\text{Mish}(s) = s \cdot \tanh(\ln(1 + e^s)) \tag{9}$$

Additionally, we implemented an MLP with the architecture $15 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 1$, utilizing ReLU-like functions

and Sigmoid as activation functions. The model was optimized using the Adam optimizer with a learning rate of 0.01 and trained for 1000 epochs.

Again, the final results are presented in Table 1. With the Mish activation function, the MLP achieved an MSE of **0.0011** on the lower bound and **0.0010** on the upper bound. For MAE, it attained **0.0225** on the lower bound and **0.0247** on the upper bound. The confusion matrix is shown in Figure 4, and a clearer comparison with the true PNS bounds is provided in Figure 5 (Only the best performance comparisons with Mish are shown).

Overall, MLP significantly outperformed other machine

learning models, with Mish yielding the best results among the activation functions. The comparison with the true PNS bounds further confirms that MLP (Mish) provides an accurate and practical model for predicting PNS.

5.6 EXPERIMENTAL COMPARISON

As shown in Table 1, MLP delivers the best overall performance. Among the other four machine learning models, SVM performs well on the lower bound but fails almost entirely on the upper bound. RF shows significantly better results, achieving acceptable performance on both bounds. Despite also being a tree-based model, GBDT underperforms compared to RF, with only a slight improvement over SVM on the upper bound. The Transformer, as an MLP-based model, outperforms the other machine learning models but still falls short of MLP's performance.

For MLP models, the dataset's characteristics around zero (we will discuss these characteristics in the discussion section) lead to notable differences in activation function performance. Basic ReLU shows suboptimal performance on the lower bound, while LeakyReLU, which accounts for negative values, performs slightly better. Mish, which not only handles negative values but also ensures differentiability around zero, achieves the best results.

-

Table 1: Comparison of Model Performance

Model	Dataset	MSE	MAE
SVM	Lower bound	0.0112	0.0868
	Upper bound	0.0304	0.1527
RF	Lower bound	0.0116	0.0919
	Upper bound	0.0205	0.1242
GBDT	Lower bound	0.0159	0.1049
	Upper bound	0.0261	0.1399
Transformer	Lower bound	0.0030	0.0348
	Upper bound	0.0156	0.1060
MLP(ReLU)	Lower bound	0.0045	0.0434
	Upper bound	0.0023	0.0357
MLP(LeakyReLU)	Lower bound	0.0038	0.0379
	Upper bound	0.0024	0.0380
MLP(Mish)	Lower bound	**0.0011**	**0.0225**
	Upper bound	**0.0010**	**0.0247**

6 DISCUSSION

Our study demonstrates that among common machine learning methods, the MLP (Mish) is the most effective and accurate in estimating the bounds of PNS. Below, we discuss key considerations and future directions.

First, in complex settings—especially those involving non-binary probabilities of causation, as indicated in [Li and

Pearl, 2024a, Zhang et al., 2024]—the lower bounds of PNS often approach zero, further emphasizing the potential of MLP (Mish) in such scenarios.

Second, while our model successfully addresses the challenge of predicting from approximately 2,000 subpopulations to 2^{15} subpopulations, a significant issue remains: the 2,000 reliable training samples may require up to 50 million data points at the population level. This is due to data sparsity and the minimum data requirements needed to estimate PNS bounds for specific subpopulations (this paper uses 1,300 observational and experimental data points). However, collecting such a large volume of data is often impractical. Fortunately, we propose three practical approaches to data collection: instead of gathering population-level data, we can directly collect data for the 2,000 predefined subpopulations; the required 1,300 data points can be reduced to approximately 400, as suggested by Li et al. [2022c], while maintaining basic accuracy; and we can rely solely on experimental data, as proposed by Li and Pearl [2023], or use only observational data and estimate experimental data through adjustment formulas [Pearl, 1993, 1995].

A fundamental goal in causal inference is identifying subpopulations exhibiting desirable counterfactual behavior patterns [Li and Pearl, 2019] to effectively guide policy-making and decision implementation. Specifically, this involves finding subpopulations with sufficiently small PNS upper bounds or sufficiently large PNS lower bounds. Therefore, accurately predicting all subpopulations is unnecessary for this task. A promising research direction is determining a minimal training set that can reliably predict these key subpopulations.

Next, the current data-generating process consists of a relatively simple causal structure with only 20 confounders, making MLP the optimal choice. However, if the underlying causal structure were more complex or included additional confounding variables, would more sophisticated models be necessary? Since this paper focuses solely on illustrating the feasibility and potential of machine learning models for predicting causal quantities, a deeper exploration of the relationship between model complexity and causal structure remains an important direction for future research.

Finally, traditional methods for estimating probabilities of causation, such as PNS, PS, and PN, rely on direct computation using observational and experimental data within SCMs. However, they assume sufficient data for each subgroup, which is often unrealistic. Our work addresses this challenge by developing a machine learning-based framework to predict PNS for subpopulations with insufficient data, a gap not covered by traditional causal inference techniques.

Although there are no classical methods explicitly designed for this task, we establish a robust benchmark for evaluation by leveraging the known data-generating process in
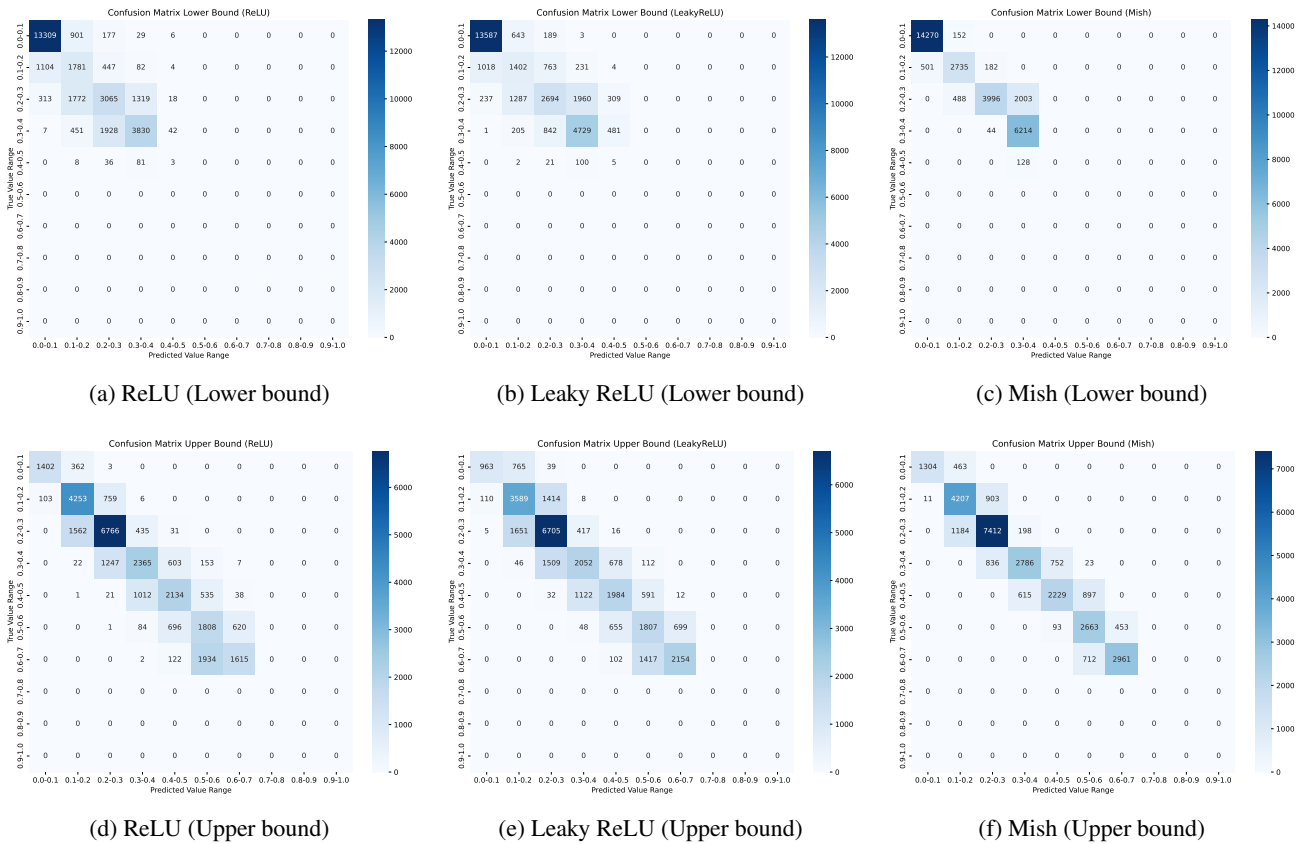
(a) ReLU (Lower bound)   (b) Leaky ReLU (Lower bound)   (c) Mish (Lower bound)

(d) ReLU (Upper bound)   (e) Leaky ReLU (Upper bound)   (f) Mish (Upper bound)

Figure 4: Confusion matrices of MLP with different activation functions: ReLU, Leaky ReLU, and Mish for both lower and upper bounds.
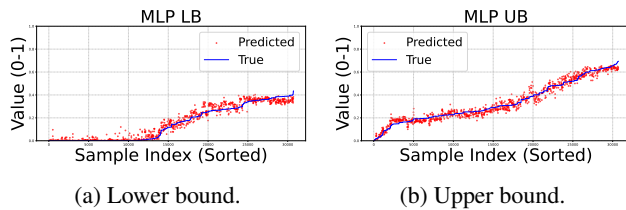

(a) Lower bound.   (b) Upper bound.

Figure 5: Comparison of MLP (Mish) for lower and upper bounds.

our synthetic experiments. For subgroups with insufficient data, we compare our model predictions to the true PNS bounds derived from SCM equations, ensuring that our results are grounded in a well-defined standard. This explicit benchmarking demonstrates the validity of our approach and allows us to assess model accuracy relative to the traditional PNS estimation under ideal conditions.

7 CONCLUSION

In this paper, we demonstrated that the bounds of probabilities of causation can be effectively learned and predicted using machine learning models. Specifically, we proposed five different models to predict the bounds of PNS. Experiments showed that an MLP with the Mish activation function achieved a mean absolute error of approximately 0.02 for an SCM with 15 observed and 5 unobserved confounders. Our results suggest that machine learning is a powerful tool for causal inference, particularly in real-world scenarios where direct estimation using SCM formulas is infeasible due to data limitations. Future research will explore larger datasets with more complex SCMs.

Although our study demonstrates the feasibility of machine learning for estimating probabilities of causation, we acknowledge that our experiments are based on synthetic data generated from a structured SCM. Most existing research on probabilities of causation remains theoretical, often without practical validation, despite claims of real-world applicability. Due to page limitations, we could not extend our study to real-world applications, but this remains a critical direction for future research. We believe that bridging this gap will require developing datasets from real-world causal systems where experimental and observational data can be systematically collected. Our work serves as a first step in this direction, providing a foundation for future studies to explore the practical deployment of machine learning models for causality estimation.

References

Alexander Abraham Balke. *Probabilistic counterfactuals: semantics, computation, and applications*. University of California, Los Angeles, 1995.

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

Corinna Cortes. Support-vector networks. *Machine Learning*, 1995.

Philip Dawid, Monica Musio, and Rossella Murtas. The probability of causation. *Law, Probability and Risk*, 16: 163–179, 2017.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5): 1189–1232, 2001. doi: 10.1214/aos/1013203451.

David Galles and Judea Pearl. An axiomatic characterization of causal counterfactuals. *Foundations of Science*, 3(1): 151–182, 1998.

Joseph Y Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000.

James Heckman and Rodrigo Pinto. Causal analysis after haavelmo. *Econometric Theory*, 31(1):115–151, 2015.

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

Ang Li and Judea Pearl. Unit selection based on counterfactual logic. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Ang Li and Judea Pearl. Probabilities of causation: Role of observational data. In *International Conference on Artificial Intelligence and Statistics*, pages 10012–10027. PMLR, 2023.

Ang Li and Judea Pearl. Probabilities of causation with nonbinary treatment and effect. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20465–20472, 2024a.

Ang Li and Judea Pearl. Unit selection with nonbinary treatment and effect. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38(18), pages 20473–20480, 2024b.

Ang Li, Suming J. Chen, Jingzheng Qin, and Zhen Qin. Training machine learning models with causal logic. In *Companion Proceedings of the Web Conference 2020*, pages 557–561, 2020.

Ang Li, Song Jiang, Yizhou Sun, and Judea Pearl. Learning probabilities of causation from finite population data. *arXiv preprint arXiv:2210.08453*, 2022a.

Ang Li, Song Jiang, Yizhou Sun, and Judea Pearl. Unit selection: Learning benefit function from finite population data. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022b.

Ang Li, Ruirui Mao, and Judea Pearl. Probabilities of causation: Adequate size of experimental and observational samples. In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*, 2022c.

Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, GA, 2013.

Diganta Misra. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*, 2019.

Scott Mueller and Judea Pearl. Perspective on 'harm' in personalized medicine – an alternative perspective. Technical Report R-530, Department of Computer Science, University of California, Los Angeles, CA, 2023. Forthcoming, American Journal of Epidemiology.

Scott Mueller, Ang Li, and Judea Pearl. Causes of effects: Learning individual responses from population data. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, pages 2712–2718, 2022.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

Judea Pearl. Aspects of graphical models connected with causality. *Proceedings of the 49th Session of the international Statistical Institute, Italy*, pages 399–401, 1993.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese*, pages 93–149, 1999.

Judea Pearl. *Causality*. Cambridge university press, 2nd edition, 2009.

Drago Plecko and Elias Bareinboim. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Chi Zhang, Ang Li, Scott Mueller, and Rumen Iliev. Causal ai framework for unit selection in optimizing electric vehicle procurement. In *2nd Workshop on Sustainable AI*, 2024.