# Causal Interpretability and Uncertainty Estimation in Mixture Density Networks*

Gokul Swamy[0000−0002−0554−4141], Arunita Das[0000−0001−6063−5073] ✉, and Shobhit Niranjan[ORCID]

Amazon, Seattle WA 98170, USA
{swagokul, arunita, shobhnir}@amazon.com

**Abstract.** Neural network implementations have predominantly been a black box lacking both in interpretability and estimation of uncertainty. In this study, we propose a novel causal attribution methodology for mixture density networks wherein we outline a framework to compute the causal effect of each feature on the target variable along with the associated uncertainty in the attribution. Our approach allows for the prediction and causal estimation tasks, along with the uncertainty estimation, to be integrated within the same architecture thus obviating the need to train a separate causal model. We report experimental results on two real-world problems comprising of studying the causal impact of bio markers on diabetes progression and the causal impact of certain key features on ecommerce sales. We also evaluate our approach on an open source simulated dataset and compare our results against multiple state-of-the-art baselines.

**Keywords:** Causality · Neural Network · Mixture Density Network.

## 1 Introduction

Most (deep) neural network implementations in practice continue to be black-boxes with scant regard for model interpretability and output stochasticity. This has led to low adoption of these techniques in sensitive arenas such as banking and finance etc. In other instances, deep learning models have fallen foul of fairness standards which could have been thwarted had model interpretability been incorporated into the architecture [8]. Furthermore, traditional neural network models are trained to yield conditional mean estimates (especially in the regression setting) without estimating associated confidence bounds. It is pertinent to estimate the uncertainty in the target predictions as these estimates can help the practitioner implement a robust decisioning system that can for example preclude a treatment being applied to an instance where the uncertainty in the estimate is high. Several works have focused on uncertainty estimation in neural networks through either learning probability distributions [4] of the target

* Supported by Amazon

variable or through sampling based approaches [16]. Interpretability has also received a lot of attention in the recent past with a slew of methods proposed to solve this problem [19, 12]. While these methodologies work well for a large number of scenarios, the feature importance doled out by these algorithms do not constitute a "causal estimate". Further, Sundararajan et. al., [21] presented a set of two axiomatic properties, sensitivity and implementation invariance, that are desirable for any attribution framework and showed that gradient based attribution methods violate the sensitivity criterion while surrogate methods suffer from not being implementation invariant. In causal estimation, we seek to understand the influence of a particular feature on the target variable when adjusted for the confounders [13]. In the absence of a causal estimate, the feature attributions generated by these algorithms could sometimes lead to grossly erroneous conclusions. There are multiple algorithms for causal estimation which involve training a separate causal model such as propensity score matching [15], doubly robust regression [14], double ML [6] and more recently the causal forest [2] and generalized random forest [1] algorithms. In the context of neural networks, a few approaches have been proposed such as DragonNet [18], TarNet [17], NedNet [17], etc that aim to combine the prediction and causal estimation within the same architecture. However, these methods are only suitable for binary treatment regimes and are also somewhat more optimized for the causal estimation task as opposed to the prediction task.

In this study, we present an approach towards estimating the causal impact of each feature in a neural network on a target variable basis a pre-trained prediction model. Our main contributions in this work include:

- We propose a unified model using a mixture density network towards combining the prediction task, output uncertainty prediction and the causal attribution of each feature on the target variable without the need for training a separate causal model.
- The model can generate the causal impact of a feature on the target variable with very little computational overhead.
- Our methodology estimates the associated uncertainty bounds around the causal estimates. The uncertainty bounds are especially pertinent when training in a few-shot setting i.e., low data domains.

The rest of this paper is organized as follows: In section 2, we outline our proposed methodology for causal impact estimation from mixture density networks and in section 3, we highlight the performance of our methodology on two real-world datasets and a benchmark simulated dataset from the public domain. Finally in section 4, we conclude our work and outline scope for future research.

## 2 Methodology

### 2.1 Task Definition

Let us denote the problem domain to consist of a set of $n$ features $X$ as $\{x_1, ..., x_n\}$, where the problem involves mapping the function $y = f(X)$. Given this setting,

our goal is to develop the causal relationship between $x_i$ and $y$ $\forall i \in \{1, .., n\}$ along with the associated uncertainty in the estimates. In order to characterize this uncertainty we model the regression problem using a mixture density network (MDN). For the sake of completeness we briefly elaborate on the MDN in the next subsection and detail our causal estimation framework using MDN as the base architecture in subsequent subsections.

### 2.2 Mixture Density Network

Mixture density networks are trained to output the parameters of a mixture density for every model input. For example, in the case of Gaussian mixtures (GMM), the model would output the mixing coefficients ($\pi$), the variances ($\sigma$) and the means ($\mu$) of each of the constituents in the GMM. The loss function for the MDN is usually the likelihood of the data under the predicted distribution and is given by:

$$\arg_{\theta^*} \min l(\theta^*) = -\frac{1}{|D|} \sum_{(x,y)\in D} \log p(h(x)|x) \tag{1}$$

In the remainder of this text we will reference the Gaussian mixture density model using the functional form $f_{s_i}$ with the subscript $s_i$ indicating the $i^{th}$ mixture component of either the mean ($\mu_i$), the variance ($\sigma_i$) or the mixing coefficient ($\pi_i$). In the simplest case, the mixture density network can be trained to output the mean ($f_\mu$) and standard deviation ($f_\sigma$) of a uni-modal Gaussian distribution of the target variable. We derive the causal framework for this simple case and subsequently extend it for multi-modal Gaussian mixtures.

**Causal Attribution Uni-modal Gaussian MDN** We define the average causal attribution of an input feature $x_i$ on the output $y$ as (see [5]):

$$G^y_{x_i=\alpha} = E[y|do(x_i = \alpha)] - b_{x_i} \tag{2}$$

here the $do(.)$ operator [13] indicates the causal dependence wherein $x_i$ is an intervened variable forced to take on the value $\alpha$, as opposed to the Bayesian dependence given by $E[y|x_i = \alpha]$, where $x_i$ is observed and takes on the value $\alpha$ in accordance with the data distribution. $b_{x_i}$ is some chosen baseline against which the intervention is measured. In the case of a GMM with a single mixture component we can define the average causal estimate to be the expectation of the mean given by:

$$CE^{ATE}_{x_i=\alpha} = E[\mu_y|do(x_i = \alpha)] - b_{x_i} \tag{3}$$

One possible approach to estimate $G^{\mu_y}_{x_i=\alpha}$ would be to perform a summation over $\mu_y$ while sampling all possible states of the input features $\{x_1, ..., x_{i-1}, x_{i+1}, ..., x_n\}$ keeping $x_i = \alpha$. This of course is not computationally feasible even for modestly sized datasets when multiple such computations are desired. To overcome this limitation, we can perform a Taylor expansion of the mixture density network

around $\mu_x = \{\mu_{x1}, ..., \mu_{xn}\}$, with $\mu_{xj}$ being given as: $\mu_{xj} = E[x_j|x_i = \alpha]$, to yield:

$$f_\mu(z) = f_\mu(\mu_x) + \nabla^T f_\mu(\mu_x)(z - \mu_x) + \frac{1}{2}(z - \mu_x)^T \nabla^2 f_\mu(\mu_x)(z - \mu_x) + ... \quad (4)$$

Ignoring the higher order terms, marginalising the input features (except for $x_i$) and computing the expectation we get:

$$CE^{ATE}_{x_i=\alpha} = E[f_\mu(z)|do(x_i = \alpha)] \approx \quad (5)$$
$$f_\mu(\mu_x) + \frac{1}{2} Tr(\nabla^2 f_\mu(\mu_x) E[(z - \mu_x)(z - \mu_x)^T | do(x_i = \alpha)]$$

If we make the assumption that none of the inputs are causally related to each other (they could still be correlated) then the causal dependence $\mu_{xj} = E[x_j|do(x_i = \alpha)]$ boils down to $\mu_{xj} = E[x_j]$. This causal independence allows for the means and the covariance matrices in eq (5) to be precomputed resulting in fast computations of the average causal effect of all the features taking on any continuous value within the allowable range. In several instances, the average causal estimate does not suffice to completely characterize the impact of a feature value / treatment on a target variable. As an example, while a drug may exhibit a good average treatment effect, it might so happen that the drug does not result in any treatment effect being observed in 20% of the population. While it is possible to infer such findings using methodologies aimed at estimating individual causal effects (ICE), such approaches have several pitfalls including the requirement for a separate model to be learnt to infer the ICE and also the inability to use a single model to estimate the causal impact from a large set of features. In order to overcome these limitations and estimate the uncertainty bounds around the average causal estimate of any of the desired features we make note of the Quantile function of the Gaussian distribution given by:

$$l_p = inf\{x \in R : F(x) \geq p\} = \mu + \sigma(\sqrt{2} * erf^{-1}(2p - 1)) \quad (6)$$

where $F(x)$ is the cumulative distribution function of the Gaussian distribution and $erf^{-1}(.)$ is the inverse error function. By performing an interventional expectation on the $p^{th}$ percentile we can estimate the causal uncertainty bounds as:

$$CE^{lb}_{x_i=\alpha} = E[l_{p_{low}}|do(x_i = \alpha)] - b_{x_i} \quad (7)$$
$$CE^{ub}_{x_i=\alpha} = E[l_{p_{high}}|do(x_i = \alpha)] - b_{x_i}$$

Once again, we can approximate the function for the $p^{th}$ percentile using a Taylor expansion around $\mu_x$ and take the interventional expectation to obtain:

$$CE^{lb}_{x_i=\alpha} \approx f_{lb}(\mu_x) + \frac{1}{2} Tr(\nabla^2 f_{lb}(\mu_x) E[(z - \mu_x)(z - \mu_x)^T | do(x_i = \alpha)] \quad (8)$$
$$CE^{ub}_{x_i=\alpha} \approx f_{ub}(\mu_x) + \frac{1}{2} Tr(\nabla^2 f_{ub}(\mu_x) E[(z - \mu_x)(z - \mu_x)^T | do(x_i = \alpha)]$$

with

$$f_{lb/ub}(x) = f_{\mu}(x) + f_{\sigma}(x) * (\sqrt{2} * erf^{-1}(2p_{low/high} - 1)) \tag{9}$$

It is straightforward to compute (8) as the data covariance remains invariant under $\mu_x$ and the terms $f_{lb/ub}(\mu_x)$ and $\nabla^2 f_{lb/ub}(\mu_x)$ can be easily computed using equation (9) through plugging in the appropriate input and employing gradient back propagation for computing the Hessian. While we have discussed the estimation of the average causal effect so far, it is also possible to compute the conditional average treatment effect (CATE) given by:

$$CE^{CATE}_{x_i=\alpha} = E[\mu_y | X = x, do(x_i = \alpha)] - b_{x_i} \tag{10}$$

by setting $\mu_x = X$ in eq (5) and performing a Taylor expansion of the neural network around the conditional mean of the input features. The uncertainity bounds for the CATE can be estimated in a similar manner using eq's (6) to (9).

**Causal Attribution for the multi-modal Gaussian MDN** In several instances a unimodal distribution is insufficient to characterize the target distribution and it is required to have more than one mixture component to better characterize the output stochasticity. The output of a MDN employing a GMM, with multiple mixture components, comprises of the set $[\pi_i, \mu_i, \sigma_i] \ \forall \ i \in \{1, 2...M\}$ where $M$ is the number of mixtures in the GMM. For a multi-modal Gaussian mixture model we define the average causal estimate to be the interventional expectation of the target mean given by:

$$CE^{ATE}_{x_i=\alpha} = E[\mu_y | do(x_i = \alpha)] - b_{x_i} \tag{11}$$

where, $\mu_y = \sum_{i=1}^{M} \pi_i \mu_i$, is the mean of the Gaussian Mixture. As before, $\mu_y$ can be approximated using a Taylor expansion and the interventional expectation can be computed similar to equation (5). In order to estimate the uncertainty around the average causal estimate it is required to obtain the Quantile function ($l_p$) for the Gaussian Mixture. Unlike the uni-modal case the Quantile function for the GMM does not have a closed form analytic expression. In order to overcome this impediment, we model $l_p$ as a feed forward neural network ($f_p$) with the input to this network being the outputs of the MDN and the percentile $p$. A permutation of the mixture components has no bearing on the underlying distribution and therefore the proposed feed forward network must be invariant to the permutations of the input space. In order to incorporate this property into our network, we make use of the deep set architecture defined in [23] wherein the feed forward neural network is constrained to have the following form:

$$h(.) = \rho(\sum_{i=1}^{M} \phi(\beta_i)) \tag{12}$$

where $\beta_i$ is the vector $[\pi_i, \mu_i, \sigma_i, p]$ and $\beta = \cup_{i=1}^{M} \{\beta_i\}$. The function $\phi : R^{4\times 1} \rightarrow R^{n_h \times 1}$ can be thought of as a distinct neural network with a $4 \times 1$ input and a

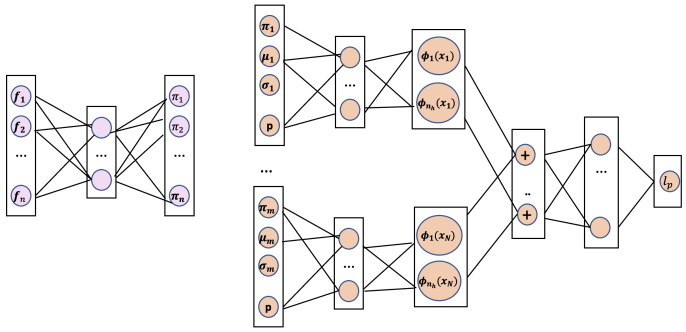

**Fig. 1.** Left (pink): Representation of a mixture density network (MDN) with M-mixtures. Right (orange): A permutation invariant Quantile prediction network with the distribution parameters as the input.

$n_h \times 1$ output. $\phi$ is applied to each individual element in the set $\beta$ and the results are aggregated before passing it through additional layers of a neural network defined by $\rho$. The overall architecture of the MDN and its extension towards predicting the value of the $p^{th}$ GMM quantile is depicted in figure 1. Given the approximation for the quantile function, we can easily compute the uncertainty bounds around the average causal estimate using equation (8) while replacing $f_{lb/ub}$ with the feed forward network $f_p$ and using gradient back propagation to compute the hessian $\nabla^2 f_p(\mu_x)$. In the next section we highlight the results of our approach on two real-world datasets from the e-commerce and healthcare domains and subsequently compare our results against multiple state-of-the-art baselines using the benchmark IHDP causal inferencing dataset.

## 3 Results

### 3.1 Baselines

We compare the proposed approach (Causal MDN) against multiple baselines including causal forest double machine learning (DML) [1], DragonNet [18], TarNet [17] and NedNet [18] causal inferencing architectures. DragonNet, TarNet and NedNet are neural network based causal estimation architectures that combine the outcome prediction and the causal estimation within the same architecture while the causal forest DML algorithm requires training a separate model for estimating the causal attribution of each feature. However, since the DragonNet, TarNet and NedNet baselines are only suitable for a binary treatment regime, we use the causal forest DML algorithm as a baseline for the two real-world problems and do a full comparison of all the baselines on the benchmark synthetic IHDP dataset [11].

### 3.2 Causal Impact of Bio-markers in Diabetes progression

In this section, we test the robustness of our approach on an external diabetes dataset [9] consisting of 442 data samples. The feature set comprises of ten

features and one target variable indicative of disease progression after one year. The input features and the target variable are all normalized to lie in the range of [0,1] and the baseline is set to 0.5 for all the features. We train a MDN model on this dataset with two hidden layers of size 64 and 32 followed by ReLU activation. The final layer is an MDN layer outputting the parameters of a unimodal Gaussian distribution ($M = 1$). We employ the proposed methodology to ascertain the average causal estimate and the associated uncertainty for each of the features as in the previous section. We compare our method against the state-of-the-art double machine learning (DML) based causal forest estimator [1] by measuring the average treatment effect (ATE) for each feature taking on values in the range [0,1] against a baseline of 0.5. The data generating process for a DML estimator is given by:

$$y = \theta(x) * T + g(x) + \epsilon,\ T = m(x) + \eta \tag{13}$$

where T is the treatment variable set to the feature of interest and g(.) and m(.) are chosen to be random forest regressors. The ATE can then be estimated by computing the expectation:

$$ATE(t) = E[y(T = t) - y(T = baseline)] \forall t \in [0, 1] \tag{14}$$

It must be noted that in DML, the effect model between the outcome and treatment is linear. The Microsoft EconML [3] package was used for training the DML causal forest estimator and required training of ten separate models to characterize the causal impact of each of the ten features with the remaining features comprising of the confounder set.

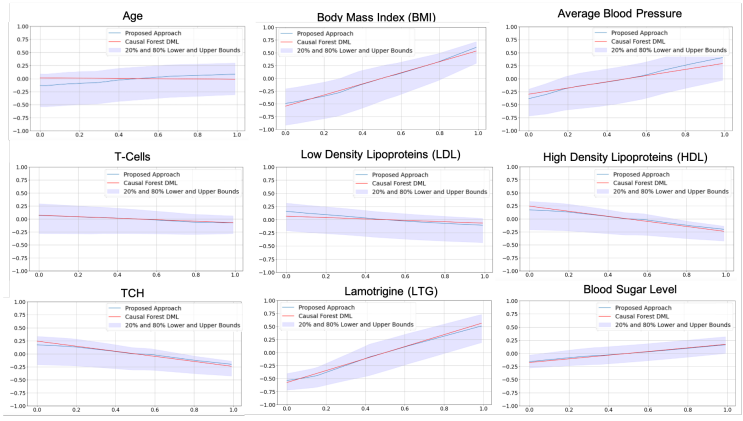


**Fig. 2.** Average Causal Estimate along with the uncertainty for different interventions of the features in the Diabetes progression prediction problem

The plots in figure 2 depict the average causal estimate and the 20% and 80% uncertainty bounds for each of the features along with the causal estimates from the causal forest DML algorithm (red line). The causal estimates from the pretrained MDN network closely mirror the causal estimates from the dedicated

causal models. Table 1 captures the root-mean-squared errors between the causal estimates from the proposed method and the Causal Forest DML methodology.

It is also apparent from the plot that BMI has the highest causal impact with an increasing BMI being suggestive of more rapid disease progression. This correlates well with clinical literature as detailed in [10]. The variance in the causal estimate, with BMI as the intervened feature, is higher for lower BMI's and decreases with increasing BMI. This finding is corroborated by clinical studies [7] which found a strong correlation between variance of glucose increase with decreasing BMI ($p - value \approx 0.02$). Clinical literature studies also indicate a higher risk of diabetes progression with increasing arterial blood pressures [22], increasing LTG and decreasing levels of LDL and HDL proteins [7]. A mild increase in risk has also been observed with increasing age [10]. These trends are evident in our causal estimates as can be visualized from figure 2.

**Table 1.** RMSE of Causal MDN against Causal Forest DML

| **Feature Name** | Age | BMI | BP | T-cells | LDL | HDL | TCH | LTG | Blood Sugar Level |
|---|---|---|---|---|---|---|---|---|---|
| **RMSE** | 0.063 | 0.035 | 0.058 | 0.011 | 0.084 | 0.057 | 0.024 | 0.036 | 0.021 |

### 3.3 Causal Impact of a Key Feature on Downstream Customer Spends

In most instances ecommerce deliveries are unattended wherein the package is left outside the customers door without the customer being physically present to receive the shipment. However, in certain instances where there is a high propensity of a shipment being lost or a shipment containing high-value contents it is required of the customer to physically receive the shipment (attended delivery). This requirement adds friction to the delivery process and in this study we are

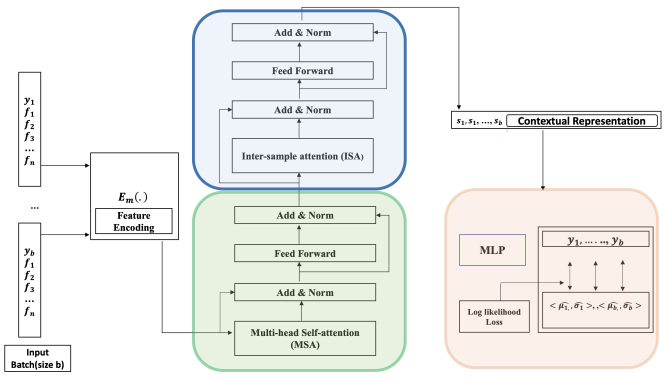


**Fig. 3.** The transformer architecture adapted from [20] for predicting downstream spend

primarily concerned with estimating the causal impact of the percentage of attended deliveries on a customers one-month downstream spend. The base model

for predicting a customers downstream spend comprises of the transformer architecture for tabular data proposed in [20] with the final layer being replaced with a mixture density network comprising of a uni-modal mixture component. The input to the model comprises of 180 customer level features involving the customers past spends, frequently shopped categories, percentage of attended deliveries, payment means adopted, etc. The layer wise details of the model architecture are shown in figure 3. The plots in figure 4 show the causal trends for two selected features (past spends and attended delivery percentage) along with the associated uncertainties. The red lines are the estimates from the doubly debiased causal forest methodology wherein a causal model was trained separately for each feature as in the previous example. While the causal estimates for the attended delivery treatment are very close, there is a significant deviation for the past spends feature primarily on account of the causal forest DML algorithm mapping the estimates to a linear trend. We compare the computation time of two approaches on a d2.8xlarge AWS cloud instance. Our method while directly operating on the pre-trained MDN architecture was 71x faster than having to train a causal forest DML algorithm from scratch thus offering a significant benefit for practical applications with stringent computational and latency constraints.

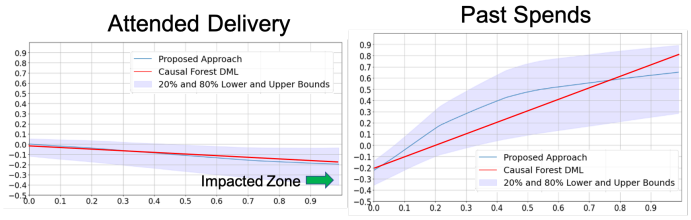


**Fig. 4.** Average Causal Estimate along with the uncertainty for two key features. The impacted zone in the lower bound of the causal uncertainty arises from customers who are more negatively impacted by the proposed treatment

A look at the causal attribution graph for the past spends indicates that past spends are highly causative of future spends which is to be expected. Causal attribution from the attended delivery feature is indicative of the fact that a majority of the customers are largely agnostic of how they receive their packages. However, the small negative slope might imply some customers tend to reduce their spends when faced with repeated attended deliveries as they may find the experience inconvenient.

**Interpretation of Causal Uncertainty** The uncertainty in the causal estimates for past spends show an increasing trend with increase in past spends. This can be explained by the fact that some customers with high past spends tend to either lower or increase their spends by a significant amount whereas customers with low past spends do not often drastically increase their spends. A similar trend is observed for attended delivery percentage feature where a larger

variance in the causal impact is observed with increasing percentages of attended delivery. This is again explained by the fact that for a cohort of customers, attended delivery might be grossly inconvenient which might cause them to curtail their spends to a large extent. This fact cannot be established by looking at the average causal estimates alone and the uncertainty in the causal estimates allows us the opportunity to infer findings that impact only a much smaller cohort and take remedial actions to prevent a detrimental outcome. We used the CATE estimation proposed in eqn (10) to identify this customer cohort and adjusted our policies to minimize the attended delivery percentage for this customer cohort while simultaneously optimizing for the costs incurred on account of lost packages. Basis a randomized trial we were able to measure a 10 basis points improvement (p value $\leq 0.005$) in customer downstream spends leading to a significant increase in sales revenue (not disclosed) basis the proposed intervention.

### 3.4 ATE Estimation for the Benchmark IHDP Dataset

In order to evaluate our algorithm on a public benchmark, we make use of the popular IHDP dataset [11]. IHDP is a semi-syntehtic dataset constructed from the infant health and development program. The dataset studies the impact of home visits by specialists on cognitive health. Given its a semi-synthetic dataset we have both the factual as well as counterfactual measurements available for each sample in the training dataset thus allowing multiple methods to be compared based on the efficacy of the counterfactual estimation. The dataset comprises of 26 covariate features which are potential confounders for estimating the average treatment effect. For the purpose of this comparison, we took 50 realizations of the IHDP dataset with each realization consisting of 747 observations and compute the ATE for each of these realizations basis the proposed method, DragonNet, TarNet, NedNet and causal forest DML architectures. The results of our experiment are summarized in figure 5 and in table 2 where it is evident that the Causal MDN is able to outperform the other baselines by a significant margin. The blue shaded area in figure 5 represents the uncertainty bounds in the ATE estimate and interestingly the uncertainty bounds are larger when the error between the actual and predicted ATE (from the proposed approach) is higher (e.g., DATASET ID 2, 5, 14, 21, 28) thus indicating that the proposed model is able to accurately characterize the uncertainty in the causal estimate which can be used to guide the final decisioning.

**Table 2.** RMSE of Causal MDN Vs Baselines for IHDP Dataset

| Model | Causal MDN | Dragonnet | Tarnet | Nednet | Causal Forest DML |
|---|---|---|---|---|---|
| Overall RMSE | **0.22** | 0.45 | 0.58 | 0.71 | 0.86 |

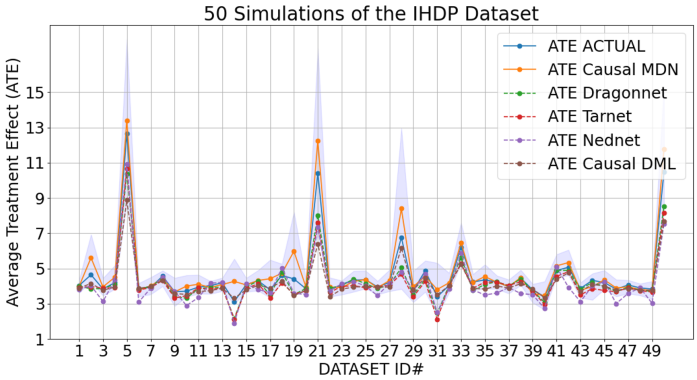


**Fig. 5.** ATE plot for 50 Simulations of the IHDP Dataset

## 4 Conclusion

In this study we have presented a novel causal estimation algorithm enabling the twin pillars of causal interpretability and uncertainty prediction for a host of applications. The proposed model works by approximating the neural network using a second order Taylor expansion and employing a mixture density network architecture to model the underlying uncertainty in the causal estimates. The proposed architecture does not require a separate causal model to be learnt for every feature and is capable of generating the causal estimates for the entire set of features (or any desired set of features) with minimal computational overload thus making it highly suited for many practical applications with computational and latency constraints. Also, while other contemporary models are able to generate uncertainty estimates using bootstrap aggregation or Monte Carlo sampling, our approach elegantly blends the uncertainty estimation within the proposed network architecture. We have compared our method against multiple state-of-the-art baselines and show that our method is superior both in terms of predictive performance as well as computational efficiency. Finally, our methodology is not limited to the mixture density network alone and can be adapted to other architectures where the prediction function is continuous and differentiable.

## References

1. Athey, S., Tibshirani, J., Wager, S.: Generalized random forests. The Annals of Statistics **47**(2), 1148 – 1178 (2019). https://doi.org/10.1214/18-AOS1709, https://doi.org/10.1214/18-AOS1709
2. Athey, S., Wager, S.: Estimating treatment effects with causal forests: An application (2019)
3. Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., Syrgkanis, V.: EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML (2019), version 0.x
4. Bishop, C.M.: Mixture density networks. Tech. rep. (1994)

5. Chattopadhyay, A., Manupriya, P., Sarkar, A., Balasubramanian, V.N.: Neural network attributions: A causal perspective. In: International Conference on Machine Learning. pp. 981–990. PMLR (2019)
6. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J.: Double/debiased machine learning for treatment and causal parameters (2017)
7. Das, R.N.: Diabetes and obesity determinants based on blood serum (2018)
8. Du, M., Yang, F., Zou, N., Hu, X.: Fairness in deep learning: A computational perspective (2020)
9. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al.: Least angle regression. Annals of statistics **32**(2), 407–499 (2004)
10. Fonseca, V.A.: Defining and characterizing the progression of type 2 diabetes (2009)
11. Hill, J.L.: Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics **20**(1), 217–240 (2011). https://doi.org/10.1198/jcgs.2010.08162, https://doi.org/10.1198/jcgs.2010.08162
12. Lundberg, S., Lee, S.: A unified approach to interpreting model predictions. CoRR **abs/1705.07874** (2017), http://arxiv.org/abs/1705.07874
13. Pearl, J.: Causality. Cambridge university press (2009)
14. Robins, J., Sued, M., Lei-Gomez, Q., Rotnitzky, A.: Double-robust and efficient methods for estimating the causal effects of a binary treatment (2020)
15. Rubin, D.B., Thomas, N.: Combining propensity score matching with additional adjustments for prognostic covariates. Journal of the American Statistical Association **95**(450), 573–585 (2000)
16. Schupbach, J., Sheppard, J.W., Forrester, T.: Quantifying uncertainty in neural network ensembles using u-statistics. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2020). https://doi.org/10.1109/IJCNN48605.2020.9206810
17. Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms (2016). https://doi.org/10.48550/ARXIV.1606.03976, https://arxiv.org/abs/1606.03976
18. Shi, C., Blei, D.M., Veitch, V.: Adapting neural networks for the estimation of treatment effects (2019). https://doi.org/10.48550/ARXIV.1906.02120, https://arxiv.org/abs/1906.02120
19. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. CoRR **abs/1704.02685** (2017), http://arxiv.org/abs/1704.02685
20. Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C.B., Goldstein, T.: SAINT: improved neural networks for tabular data via row attention and contrastive pre-training. CoRR **abs/2106.01342** (2021), https://arxiv.org/abs/2106.01342
21. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks (2017)
22. Vasilis Tsimihodimos, Clicerio Gonzalez-Villalpando, J.B.M., Ferrannini, E.: Hypertension and diabetes mellitus (2003)
23. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., Smola, A.: Deep sets (2018)