Optimal causal decision trees ensemble for improved prediction and causal inference

NEELAM YOUNAS¹, AMJAD ALI², HAFSA HINA¹, MUHAMMAD HAMRAZ ², ZARDAD KHAN ²
AND SAEED ALDAHMANI ³.
¹Department of Econometrics, Pakistan Institute of Development Economics, Pakistan
²Department of Statistics, Abdul Wali Khan University Mardan, Pakistan
³Department of Analytics in the Digital Era, United Arab Emirates University, UAE
Corresponding authors: Zardad Khan (e-mail: zardadkhan@awkum.edu.pk), Amjad Ali (e-mail: amjad.ali@awkum.edu.pk)

ABSTRACT Ensemble methods can be used to identify causal relationships in data for a better understanding and taking the right decision in processes that involve high risk. This paper explores the idea of a causal decision tree forest and proposes a regularized ensemble method by integrating optimal causal trees for improved prediction accuracy while not compromising on accurately estimating heterogeneous treatment effects. The proposed method is based on selecting a subset of the most accurate causal trees from a sufficiently large pool based on their out-of-sample error estimates. The selected trees are integrated to form an ensemble that is used for estimating heterogeneous treatment effect and predicting unseen data. The proposed method is applied on Pakistan’s income function consisting of 27964 observations on wages of workers age 10 and above as an example dataset. The paper gives a detailed simulation study where datasets are generated under 5 different designs. The proposed method is assessed against ordinary least square (OLS), least absolute shrinkage and selection operator (LASSO), Ridge, Causal Tree and the standard decision trees forest via mean square error (MSE), root mean square error (RMSE), mean absolute deviation (MAD) and Pearson correlation (r) as performance metrics. The analyses given in the paper reveal that the proposed method can be used effectively for estimating heterogeneous treatment effects and achieves better prediction performance and as compared to the rest of the methods given in the paper.

INDEX TERMS Causal Inference, Causal Decision Tree, Random Forest, Causal Random Forest, Ensemble Learning, Heterogeneous Treatment Effect

I. INTRODUCTION
The identification of the causal relationships in the data is key to provide a better understanding and the knowledge for taking an accurate decision in processes with risk. Such types of relationships are usually established with the help of experiments which are effective but, at the same time, costly and difficult to conduct [1]. Observational studies can also be used to find the causal relationships in the data [2], which are tested by taking a sample from historical data or by observing the characteristic of interest over a period of time, thereby making the observational study time-consuming.

Machine learning is generally used for accurately predicting unknown data based on learning from known data. However, sometimes, the purpose of using machine learning methods could potentially exceed prediction, such as representing and discovering causal relationships in data and estimating heterogeneous causal effects. This kind of application provides a compact and precise graphical representation of the causal relationships between a set of predictor attributes and an outcome attribute. Various machine learning methods are in use for obtaining the desired results. Typical examples include classification and regression trees, $k$-nearest neighbours models, support vector machines, etc.

The methods for the identification of causal relationship in the data should be capable of identifying the causal effect without any prior knowledge. Moreover, these methods should be capable of dealing high dimensional data sets efficiently. The methods of classification such as decision trees [3] have the ability to identify the causal relationships in the data by using a supervised learning approach where the response variable is known or fixed. Such type of methods are commonly used in medical and social data analyses, for ex-
ample. However, these methods are not specifically designed for identifying the causal relationships in the data, and hence can provide incorrect estimates of causal relationships.

Classification methods, like random forest, are fast and could find causal signals in the data effectively. These methods, with the provision of scalability and automation, are important for exploring causal relationships in large datasets. To this end, researchers have proposed several machine learning methods for exploring causal relationships. Although these methods serve the purpose of finding causal signals, they fail to achieve higher accuracy. Therefore, the aim of this paper is to achieve both causal exploration and high prediction accuracy by proposing a causal decision trees ensemble. This will help economists in predicting and answering various causal questions for policy implementation in a machine learning framework.

Various authors have suggested that combining weak models leads to efficient ensembles. Moreover, combining the outputs of multiple classifiers also reduces generalisation error. Ensemble methods have high efficacy in that the different models involved have different inductive biases, where such diversity reduces variance-error while not increasing the bias error [4]–[6]. As the number of trees in a random forest is often very large, there has been a significant work conducted on the problem of minimising this number to reduce computational cost without decreasing prediction accuracy [7]–[10]. The overall prediction error of a random forest is highly associated with the strength of individual trees and their diversity in the forest. This idea is backed by Breiman’s [11] upper bound for the overall prediction error of random forest given by $\bar{err} \leq \rho \bar{err}$, where $t = 1, 2, \ldots, B$, and $B$ denotes the number of trees in the forest. $\bar{err}$ is the overall prediction error of the forest and $\bar{p}$ represents weighted correlation between residuals from two independent trees, i.e. mean (expected value) of their correlation over entire ensemble, and $err_t$ is the prediction error of some $t^{th}$ tree in the forest.

Generally, a random forest is based on a large number of base trees, and researchers have always tried to minimise this number in order to gradually shrink the cost of computation without negatively affecting the prediction accuracy. Overall, the prediction error of a random forest is strongly connected with the accuracy of individual trees and their diversity in the forest. The proposed method selects a subset of the best causal trees in terms of their individual strength, i.e. accuracy, from a large ensemble grown by the causal random forest. The selected trees are combined in an ensemble for predicting unknown data and estimating heterogeneous treatment effects in the data. The proposed method is applied on an example dataset from the labour force survey (LFS) of Pakistan for causal effects exploration. The paper also gives a detailed simulation study of the proposed method in comparison with causal decision tree, causal random forest, ordinary least square (OLS) linear regression, least absolute shrinkage and selection operator (LASSO) and ridge regression for further assessment. For judging the efficacy of the newly developed method, conditional average treatment effect ($CATE$), average treatment effect ($ATE$), mean square error ($MSE$), root mean square error ($RMSE$), mean absolute error ($MAE$) and Pearson correlation coefficient ($r$) are used as performance measures.

The remainder of this paper is organised as follows. Section II provides a summary of the related work done in the literature; Section III presents a detailed description of the method proposed in this paper; Section IV gives the analyses conducted in this paper based on the simulated and real datasets; and Section V concludes the findings.

## II. RELATED WORK

Extensive research has been done in the literature for estimating the parameters of interest, like heterogeneous treatment effects. Some well known methods consist of local maximum likelihood and local generalized method of moments such as [12]–[17]. Some applications of these techniques in the field of economics includes multinomial choice models in a longitudinal data type i.e. [18] and instrumental variables regression i.e. [19]. To estimate the parameters at a particular value of covariates, the core idea is to use kernel weighting function in order to place more weight on nearby observations in the covariate space. The main problem in such types of techniques is that, if the feature space is high dimensional, then the performance of these methods can be suffered from the problem known as “curse of dimensionality” [20].

Authors in [21] replaced kernel weighting with the forest-based weights i.e. weights that are obtained from the trees fraction that contain observation in the same leaf as the response value of covariate vector. Study in [11] proposed a random forest algorithm for non-parametric classification and regression building on insights from the ensemble learning literature [22]–[25]. Random forest as a type of adaptive nearest neighbor prediction is closely built in the studies given in [26] and [27], which is forest-based method for quantile regression and survival analysis. Authors in [28]–[34], have used gradient based test statistics to identify the change points in likelihood models.

According to [35], numerous data mining techniques such as classification, $k$-nearest neighbors and sequential pattern mining are sequentially applied to identify the similarity in decision trees. Also authors in [36] used unified Granger causality analysis (uGCA) framework for sequential medical imaging. The study in [37] proposed hierarchical probabilistic graphical model to simultaneously handle classification of multi-sensor and multi-resolution remote sensing of the same scene. This study consists of hierarchical Markov model along with quadtree structures in order to model the necessary information present in various special scales and a planar Markov model to tackle contextual spatial information at each resolution as well as the ensemble of causal decision trees for pixelwise modeling. Further reading on the methods for causal analysis used in machine learning can be found in the recent literature as given in [38]–[44].

Recursive partitioning models using gradient-based test
statistics were considered in [45]. Several authors in [46]–[48] achieved the statistical stability by using a random forest resampling mechanism. Similarly, another study in [49] adopted a greedy and non-parametric regression technique, utilising gradient-based approximation.

Several studies [27], [46]–[48], [50]–[62] have considered the regression problems by using the random forest algorithms. Athey, Tibshirani and Wager [21] proposed a method which is computationally efficient in generating generalised random forest (GRF). The estimates of this method are consistent and asymptotically normal, thereby providing a valid confidence interval. This method is designed to handle three main tasks, i.e., heterogeneous treatment effects, non-parametric quantile regression and conditional average treatment effects via instrumental variables.

The computational burden of trees ensemble could also be decreased without compromising the prediction accuracy by combining a small number of diverse and accurate trees. This can be achieved by using out-of-bag prediction errors from each training bootstrap sample in order to select the optimal trees on the basis of their individual performance [63]. The proposed method is a modified version of the generalised random forest [21], which involves generating a large number of causal trees and then selecting a proportion of those trees whose error rate is minimum among all the constructed causal trees.

III. METHODS
This section provides a detailed description of the method used in this paper. Before introducing the proposed method, it is deemed important to introduce the causal decision tree as a building block of the suggested optimal causal trees ensemble (OCTE).

A. CAUSAL DECISION TREE
Using the decision tree for estimating the heterogeneous treatment effect is totally different from the classical decision tree used for classification and regression problems. The classical decision tree uses a function mapping characteristics to the response variable about an individual. This can be illustrated from the tree given in Figure 1 as discussed in [64], where the matchmaking mobile app i.e. “Tinder” is used to cure a particular disease. The decision tree is unable to identify the true causal effect since majority of the young people use “Tinder” as compared to old people. Moreover, old people have little chance of recovery from the disease as compared with young people. Thus, the comparison between the two groups, which are not actually comparable, led to a misleading decision tree. A standard decision tree may thus be considered as an appropriate choice in terms of predicting the recovery from the disease but fails to identify the true cause of the disease. Furthermore, its nodes posses no causal interpretation.

On the other hand, the causal decision tree calculates the average of the treated and untreated observations in each node and then computes the difference between these averages, which represents the actual treatment effect in that node.

Estimating the individual treatment effect, i.e. \( \tau_i = Y_{1i} - Y_{0i} \), is not possible in real world problems, because the outcome of the \( i \)th individual is either \( Y_{1i} \) (the sample is treated) or \( Y_{0i} \) (the sample is untreated). One of the two outcomes, i.e. \( Y_{1i} \) or \( Y_{0i} \), has to be predicted, using the counterfactual model (the potential outcome model). For example, suppose an individual has salary \( Y_{0i} \) and education level below secondary level (untreated). We want to know the \( i \)th individual’s salary \( Y_{1i} \) if the person had education above secondary level (treated), i.e., \( X_i = 1 \). The average treatment effect of a group (population) is simply the average of the individual treatment effects included in the population, i.e.,

\[
E[\tau_i] = E[Y_{1i}] - E[Y_{0i}].
\]

A general work flow of a causal decision tree is given in Figure 2.

Decision trees for causal inference are generally used to separate data into buckets in order to estimate the average treatment effects within each node. The process of decision tree learning for causal inference can be separated into two steps for each of these tasks, commonly referred to as the splitting step and the estimation step, respectively. Therefore, a causal decision tree can effectively be used for estimating heterogeneous treatment effects in a computationally efficient manner. However, its prediction in the cases of regression and classification problems is less efficient.

Although a single causal decision tree model is interpretable and fast, the estimates it returns for heterogeneous treatment effects might not be generalisable. Therefore, the ensemble of causal decision trees can solve this problem by providing robust estimates for causal relationships at the cost of interpretability without significantly increasing computational cost. Combining a few accurate and diverse causal decision trees could provide improved estimates and might be taken forward in the direction of improving interpretability of the standard causal decision trees ensemble.

This work aims at improving the causal random forest with the help of best trees selection for size reduction and

![FIGURE 1: A standard classification tree](image)
improved estimation. To achieve this, \( B \) sub-samples are taken from the given training data \( L = (X, Y, W) \), where \( X \) is the feature space, \( Y \) is the response and \( W \) is the binary treatment. A causal decision tree is grown on each sub-sample. The performance of trees is evaluated on the basis of out-of-sample observations and ranked accordingly. Trees having the smallest error estimates on the out-of-sample observations are selected, while the rest of trees are discarded. Then, the selected trees are combined to form the optimal causal tree ensemble.

Partitioning of the given training data \( L = (X, Y, W) \) is carried out randomly into two non-overlapping groups, i.e., \( L_b = (X_b, Y_b, W_b) \) and \( L_v = (X_v, Y_v, W_v) \). Then causal trees are grown, each on a sub-sample \( L_b = (X_b, Y_b, W_b) \) from \( L_b = (X_b, Y_b, W_b) \). While doing so, a random subset of \( p' \) features is selected from the entire set of features at each node of the causal tree; (where, \( p \) is total number of features and \( p' \) is a subset of features taken from total \( p \) features randomly). This will add additional randomness to the causal trees. As the observations in \( L_v = (X_v, Y_v) \) take no part in the training phase of the causal trees, error estimates are calculated for these out-of-sample observations and the trees are ranked in descending order with respect to the error estimates. The error estimates taken in this paper are the standard errors for out-of-sample observations by each tree. The final ensemble is constructed by choosing the top ranked \( M \) causal trees.

**B. STEPS OF THE PROPOSED (OCTE)**

The proposed algorithm considers the following steps to assess treatment effect:

1) Select \( B \) number of sub-samples from the training part of the dataset i.e. \( L = (X, Y, W) \).
2) Use generalized random forest for growing a causal decision tree on each sub-sample.
3) Arrange the causal trees according to their out-of-sample predicted standard errors.
4) The best \( M \) causal trees are chosen having the smallest individual prediction standard error on out-of-sample observations.
5) Combine the \( M \) selected trees to form an optimal causal decision trees forest and use it to predict the treatment effect of new/test data points.

Pseudocode of the proposed method OCTE is given in Algorithm 1 along with an illustrating flow chart in Figure 3.

**Algorithm 1 Pseudocode of the proposed OCTE**

1: \( B \): Number of sub samples taken from training data i.e. \( L = (X, Y, W) \).
2: \( M \): Number top ranked causal trees having minimum out of sample standard error.
3: for \( t = 1 \rightarrow B \) do
4: Grow a causal decision tree using generalized random forest;
5: Compute error rate on out-of-sample observations;
6: Save all the trees;
7: Save the out-of-sample errors;
8: end for
9: Rank the causal trees based on the out-of-sample standard error;
10: Select top ranked \( M \) causal trees;
11: Combine the \( M \) selected causal trees to construct optimal causal tree ensemble (OCTE);
12: Use OCTE for estimating treatment effect and predicting unseen data.
IV. EXPERIMENTS AND RESULTS
In this paper, the proposed OCTE is assessed using five different simulation designs. It is then compared with five state-of-the-art methods, i.e., OLS, LASSO, Ridge, causal tree and causal random forest.

The OCTE is also applied on a real dataset, the nationally representative Labor Force Survey of Pakistan (LFS). The LFS data include records from 2017 to 2018 taken from Pakistan’s Bureau of Statistics. The labor force survey is a nationwide survey containing micro-data from all over the country’s demographic and employment information.

A. SIMULATED DATA
Design 1: Normally Distributed \( U_j \) and Linear Outcome Model
The simulated data models are based on [65]–[68].

1) Each cluster consists of \( n_j \) observations, where \( j = 1, 2, \ldots , J \), (where, \( J \) is the total number of clusters). Each cluster is generated by drawing a random number from a normal distribution with a small standard deviation and rounded mean \( J \) to the closest integer. All generated data contain about 4,000 observations, and a conditional sample size of five is used for \((J, I)\).

2) For all observations i.e., \( i = 1, 2, \ldots , n_j \) for \( j \)th cluster, simulate confounders with individual level \( X_{ij} = (X_{1ij}, X_{2ij}, X_{3ij}) \), measured confounder for cluster level \( Z_j \), and unmeasured confounder for cluster level \( U_j \) i.e.

\[
X_{1ij} \sim Unif(-1, 1), X_{2ij} \sim N(0, 1), \quad X_{3ij} \sim Unif(0, 1),
\]

\[
Z_j \sim Unif(-1, 1) \text{ and } U_j \sim N(0, 1).
\]

3) Status of individual treatment \( W_{ij} \) is generated from the logistic model propensity score as follow.

\[
\logit(e_{ij}) = -0.6 + 0.3X_{1ij} + 0.3X_{2ij} + 0.3X_{3ij} + 0.3Z_j + 0.4X_{2ij}Z_j + 0.4X_{3ij}Z_j + 2U_{ij} + 2U_{1ij}I(X_{3ij} < 0.3),
\]

and \( W_{ij} \sim Bernoulli(e_{ij}) \).

4) The potential outcomes \( Y_{ij1}, Y_{ij0} \) and observed response \( Y_{ij} \) are generated from the regression model as follow.

\[
Y_{ij}(w) = -70 + 2X_{1ij} + 2X_{2ij} + 2X_{3ij} + 2Z_j + 2X_{2ij}Z_j + 2U_{ij} + 2X_{1ij}I(X_{3ij} < 0.3) + 2U_{ij} + W(2 + 2X_{3ij} + 2Z_j) + r_{ij},
\]

\[
Y_{ij} = W_{ij}Y_{ij1} + (1 - W_{ij})Y_{ij0},
\]

and \( r_{ij} \sim N(0, 1) \), where \( r_{ij} \) is random error for \( i \)th sample in \( j \)th cluster.

Design 2: Uniformly Distributed \( U_j \) and Linear Outcome Model
This design utilizes a similar model for data generating as Design 1, but the only difference is that \( U_j \) (unmeasured cluster level confounder) has uniform distribution i.e. \( U_j \sim U(-2, 2) \).

Design 3: Uniformly Distributed \( U_j \), Linear Outcome Model, and Misspecified Working Models
This design is also similar to Design 2, but higher order terms in the outcome model are ignored.

Design 4: Uniformly Distributed \( U_j \) and Nonlinear Binary Outcome Model
This design is also similar to Design 2, but the difference is that the outcome model nonlinear and binary. The outcome model is given by

\[
\text{logit}(P(Y_{ij}(w) = 1|X_{ij}, Z_j, U_j)) = -0.6 + 0.3X_{1ij} + 0.3X_{2ij} + 0.3X_{3ij} + 0.3Z_j + 0.3Z_j + \beta_1U_j + \beta_2U_j(0.5 + 0.3X_{3ij}) + 0.3Z_j + \beta_2U_j,
\]

and \( Y_{ij}(w) \sim Bernoulli(P(Y_{ij}(w) = 1|X_{ij}, Z_j, U_j)) \), \( Y_{ij} = W_{ij}Y_{ij1}(1 - W_{ij})Y_{ij0} \).

Design 5: Uniformly Distributed \( U_j \) and Linear Outcome with Exponential Error Model
The construction of Design 5 is almost similar to Design 2 with \( U_j \) as uniformly distributed and linear outcome model.

| TABLE 1: Summary of Designs 1-5 |
|-------------------------------|
| Specification |
| Design | Distribution | Outcome model | Estimated models with higher order terms |
| 1 | N(0, 1) | Linear | Yes |
| 2 | Unif(-2, 2) | Linear | Yes |
| 3 | Unif(-2, 2) | Linear | No |
| 4 | Unif(-2, 2) | Non-Linear | Yes |
| 5 | Unif(-2, 2) | Linear with exponential error term | Yes |

All five designs consist of about 4000 samples and data in each design is divided into 70% training and 30% testing parts. \( MSE, \ RMSE, \ MAD \) and correlation coefficient \( (r) \) are used as performance measures for both conditional average treatment effect (\( CATE \)) i.e. \( \tau_i \) as well as average treatment effect (\( ATE \)) i.e. \( \tau \).

For \( CATE \), 500 realization are made under each design and \( MSE, \ RMSE, \ MAD \) and correlation coefficient \( (r) \) are calculated reporting their average values. Expressions of the metrics used are given bellow.

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\tau_i - \hat{\tau}_i)^2,
\]

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\tau_i - \hat{\tau}_i)^2},
\]

\[
\text{MAD} = \frac{1}{n} \sum_{i=1}^{n} |\tau_i - \hat{\tau}_i|,
\]

\[
r = \frac{\sum_{i=1}^{n} (\tau_i - \hat{\tau}_i)(\tau_i - \hat{\tau}_i)}{\sqrt{\sum_{i=1}^{n} (\tau_i - \hat{\tau}_i)^2} \sqrt{\sum_{i=1}^{n} (\tau_i - \hat{\tau}_i)^2}}.
\]
B. SIMULATED DATA RESULTS

Table 4 shows the results of the proposed method (OCTE) and all the other methods considered in this study, in terms of conditional average treatment effect (CATE). The results suggest that the proposed OCTE outperformed all the other state-of-the-art procedure on almost all the five designs. The results are also shown in the form of bar plots in Figures 4-7. The proposed OCTE provides minimum mean square error (MSE) on first four designs, while causal forest (CF) gives optimal value for Design 5. The OLS, LASSO, Ridge and CT did not perform well on any design. In terms of root mean square error (RMSE), the proposed OCTE performed better than the other methods on four designs and CF outperformed the others on Design 5. Apart from OCTE and CF, the remaining methods did not outperformed the rest of the methods on any of the designs. Similarly, OCTE is giving optimal results in terms of mean absolute deviation (MAD) as compared to the other methods. In terms of Pearson’s product moment correlation coefficients ($r$), the OCTE also outperformed the other procedures on Designs 3 and has similar performance on Designs 1 and 4, while OLS outperformed the rest of the methods on Designs 4 and 5.

Table 5 shows the results for average treatment effect (ATE) for the 5 scenarios. It is evident from the results that the proposed OCTE had better achievements than the other methods in four scenarios in terms of MSE, RMSE and MAD, and it outperformed them in three scenarios in terms of correlation. The CF method, on the other hand, gave the same correlation value in Design 3 as that of OCTE and outperformed the other methods in Design 4. However, LASSO had the best achievement in Design 5 in terms of correlation. For a visual illustration, the results are also shown in the form of bar plots in Figures 12-15. The results obtained in Table 4 and Table 5 can also be seen in the form of bar-plots in Figures 4-7 and Figures 12-15, respectively.

For further assessment, boxplots of MSE, RMSE, MAD and correlation values of CATE obtained from all the 500 runs of the simulated data for each design are displayed in Figures 8, 9, 10 and 11. The boxplots reveal that the OCTE is more consistent in comparison with the other well-known methods.

Moreover, the execution or running times (in seconds) of the new method and the causal forest method (CF) are also given in Table 2, where it can be noticed that, as the number of trees grow, the execution increases linearly, i.e. $f(B) \leq C \times O(B)$, where $B$ is the number of trees. The execution time of the OCTE is greater than that of the CF method, due to the additional tree selection step. To reduce the execution time, Step 3 of the proposed algorithm can be parallelised using existing tools, such as the “parallel” R package [69].

C. LABOUR FORCE SURVEY OF PAKISTAN (LFSP) AS AN EXAMPLE

In this research, the nationally representative labor force survey of Pakistan (LFSP) data from 2017 to 2018 is taken from Pakistan Bureau of Statistics. LFSP is a nationwide survey consisting of labours employment information from all over the country at micro level.

The working sample used is based on those in wage employment and comprises a total of 272610 workers. The analysis is restricted to those older than 10. Missing values and unusable observations are discarded, leaving a total of 27964 observations. The variables used to analyse each worker’s wage include hours worked, education, occupation, residence (in urban or rural area and in one of the four provinces), schooling attainment, gender, employment status, marital status, experience, industry, kind of enterprise and training. A brief description of the variables is given in Table 3.

### Table 2: Execution time (in seconds) of OCTE and CF for different number of trees.

| Number of trees | OCTE  | CF   |
|-----------------|-------|------|
| 50              | 7.306 | 0.296|
| 100             | 7.65  | 0.347|
| 200             | 8.412 | 0.474|
| 300             | 8.497 | 0.985|
| 400             | 10.412| 0.986|
| 500             | 11.685| 1.198|
| 1000            | 17.622| 2.309|
| 2000            | 27.281| 4.688|

### Table 3: Labour Force Survey of Pakistan (LFSP) data description

| Variable | Description |
|----------|-------------|
| $Y$      | Continuous outcome: Logged monthly income of workers. | $W$ | Education: Used as binary treatment; 1 = Education above matric, 0 = Education below matric. |
| $X_1$    | Experience: Experience is calculated using (Age-year of schooling-6). | $X_2$ | Kind of Enterprise: A categorical variable with four groups, Government, Private, Public enterprise, other. |
| $X_3$    | Industry: Industry has 22 categories of different industries, such as Manufacturing, Electricity, etc. | $X_4$ | Hours worked: Hours worked is the sum of total hours an individual work. |
| $X_5$    | Gender: Tow classes, male and female. | $X_6$ | Marital status: Marital status is classified as widow, single, married, divorced. |
| $X_7$    | Employment status: Categorical with four classes, paid employees, Employer, Self Employed, Contributing Family Helpers | $X_8$ | Occupation: It is a categorical variable with nine classes and shows the occupation of employees, such as manager, professional, etc. |
| $X_9$    | Training: Binary variable and shows whether training has been given to the employee for a job or not. | $X_{10}$ | Region: Residence of employee, which has two categories, rural and urban. |
| $X_{11}$ | Province: Province of residence, which has four categories i.e. Khyber Pakhtunkhwa, Punjab, Sindh and Balochistan. |
| Design | Methods         | MSE  | RMSE  | MAD   | r      |
|-------|----------------|------|-------|-------|--------|
|       | OLS Regression | 0.31432 | 0.55620 | 0.45744 | 0.97429 |
| 1     | Ridge Regression | 0.31560 | 0.55735 | 0.45846 | 0.97426 |
|       | Causal Tree     | 8.06559 | 2.81935 | 2.2457 | 0.51432 |
|       | Causal Forest   | 0.13283 | 0.36201 | 0.28706 | 0.97146 |
|       | OCTE            | 0.11038 | 0.32978 | 0.26069 | 0.97429 |
|       | LASSO Regression| 0.32334 | 0.56417 | 0.46353 | 0.97274 |
| 2     | Ridge Regression| 0.32475 | 0.56542 | 0.46473 | 0.97276 |
|       | Causal Tree     | 8.80697 | 2.94933 | 2.35961 | 0.45639 |
|       | Causal Forest   | 0.12809 | 0.35515 | 0.27959 | 0.97215 |
|       | OCTE            | 0.10797 | 0.32611 | 0.25628 | 0.97405 |
|       | LASSO Regression| 0.32448 | 0.56518 | 0.46445 | 0.97271 |
| 3     | Ridge Regression| 0.32475 | 0.56542 | 0.46473 | 0.97276 |
|       | Causal Tree     | 8.53528 | 2.87226 | 2.33565 | 0.47508 |
|       | Causal Forest   | 0.13646 | 0.36671 | 0.28950 | 0.97146 |
|       | OCTE            | 0.11038 | 0.32978 | 0.26069 | 0.97405 |
|       | LASSO Regression| 0.27212 | 0.51680 | 0.42195 | 0.97447 |
| 4     | Ridge Regression| 0.27234 | 0.51780 | 0.42278 | 0.97442 |
|       | Causal Tree     | 8.06559 | 2.81935 | 2.2457 | 0.51432 |
|       | Causal Forest   | 0.13283 | 0.36201 | 0.28706 | 0.97146 |
|       | OCTE            | 0.10797 | 0.32611 | 0.25628 | 0.97405 |
|       | LASSO Regression| 0.31560 | 0.55735 | 0.45846 | 0.97426 |
| 5     | Ridge Regression| 0.31556 | 0.55732 | 0.45850 | 0.97421 |
|       | Causal Tree     | 8.06559 | 2.81935 | 2.2457 | 0.51432 |
|       | Causal Forest   | 0.13283 | 0.36201 | 0.28706 | 0.97146 |
|       | OCTE            | 0.10797 | 0.32611 | 0.25628 | 0.97405 |
|       | LASSO Regression| 0.32448 | 0.56518 | 0.46445 | 0.97271 |

| Design | Methods         | MSE  | RMSE  | MAD   | r      |
|-------|----------------|------|-------|-------|--------|
|       | OLS Regression | 0.14080 | 0.37524 | 0.36957 | 0.77690 |
| 1     | Ridge Regression | 0.14187 | 0.37666 | 0.37101 | 0.77718 |
|       | Causal Tree     | 2.90104 | 1.70325 | 1.66662 | 0.45533 |
|       | Causal Forest   | 0.03190 | 0.17860 | 0.17043 | 0.85308 |
|       | OCTE            | 0.01665 | 0.12902 | 0.11817 | 0.85886 |
|       | LASSO Regression| 0.14545 | 0.38135 | 0.37425 | 0.83498 |
| 2     | Ridge Regression| 0.14608 | 0.38221 | 0.37513 | 0.83415 |
|       | Causal Tree     | 3.35401 | 1.83139 | 1.79410 | 0.48703 |
|       | Causal Forest   | 0.02936 | 0.17133 | 0.16403 | 0.90414 |
|       | OCTE            | 0.01512 | 0.12297 | 0.11456 | 0.90952 |
|       | LASSO Regression| 0.14431 | 0.37989 | 0.37279 | 0.83500 |
| 3     | Ridge Regression| 0.14545 | 0.38135 | 0.37425 | 0.83498 |
|       | Causal Tree     | 2.90104 | 1.70325 | 1.66662 | 0.45533 |
|       | Causal Forest   | 0.03190 | 0.17860 | 0.17043 | 0.85308 |
|       | OCTE            | 0.01665 | 0.12902 | 0.11817 | 0.85886 |
|       | LASSO Regression| 0.09665 | 0.31088 | 0.30538 | 0.85601 |
| 4     | Ridge Regression| 0.09756 | 0.31234 | 0.30685 | 0.85570 |
|       | Causal Tree     | 0.09794 | 0.31295 | 0.30747 | 0.85629 |
|       | Causal Forest   | 0.35188 | 1.83081 | 1.78372 | 0.27637 |
|       | OCTE            | 0.01725 | 0.13135 | 0.12203 | 0.89044 |
|       | LASSO Regression| 0.31838 | 0.56425 | 0.56392 | -0.01819 |
| 5     | Ridge Regression| 0.31699 | 0.56275 | 0.56240 | -0.05131 |
|       | Causal Tree     | 0.31762 | 0.56358 | 0.56323 | -0.05942 |
|       | Causal Forest   | 0.23373 | 0.48346 | 0.48189 | -0.16226 |
|       | OCTE            | 0.32029 | 0.56594 | 0.56562 | -0.03722 |
|       | LASSO Regression| 0.18050 | 0.42486 | 0.39127 | 0.45418 |
|       | LASSO Regression| 0.18270 | 0.42744 | 0.39414 | 0.45434 |
|       | Causal Tree     | 3.96988 | 1.99246 | 1.93983 | 0.24032 |
|       | Causal Forest   | 0.07522 | 0.27426 | 0.22826 | 0.40452 |
|       | OCTE            | 0.06731 | 0.25944 | 0.20970 | 0.37665 |
FIGURE 4: Bar plots of MSE computed for CATE.

FIGURE 5: Bar plots of RMSE computed for CATE.

FIGURE 6: Bar plots of MAD computed for CATE.

FIGURE 7: Bar plots of $r$ computed for CATE.

FIGURE 8: Boxplots of MSE computed for CATE.

FIGURE 9: Boxplots of RMSE computed for CATE.
FIGURE 10: Boxplots of $MAD$ computed for CATE.

FIGURE 13: Bar plots of $RMSE$ computed for ATE.

FIGURE 11: Boxplots of $r$ computed for CATE.

FIGURE 14: Bar plots of $MAD$ computed for ATE.

FIGURE 12: Bar plots of $MSE$ computed for ATE.

FIGURE 15: Bar plots of $r$ computed for ATE.
D. LFSP DATA RESULTS

Figure 16, shows the conditional average treatment effect (CATE) of the proposed OCTE computed for LFSP data. It can be observed from the figure that the treatment (education) has a positive effect on the average income of the individuals. This implies that higher education leads to higher income of the individuals. Similar conclusion could be drawn from the box-plot constructed in the same figure.

Figures 17-19 discuss the heterogeneity of education in the variables “Province”, “Region” and “Gender”, respectively. From Figure 17, it is clear that in province Khyber Pakhtunkhwa, the average effect of education on the income of the individuals is less than the rest of the provinces. Punjab and Sindh possess almost equal average effects of education on the income of the individuals. Figure 18 indicates the heterogeneity of education between the rural and urban areas where it is clear that in rural areas, the education has less effect on the income of the individuals. In case of variable “Gender”, education has minimum effect on the income of males as compared to females (Figure 19). In a nut shell, education is observed to have heterogeneous effect among the variables i.e. gender and region, whereas in variable province it has an approximately homogeneous effect.

FIGURE 16: CATE computed by OCTE for LFSP Data.

FIGURE 17: Heterogeneity based on Provinces for LFSP Data.

FIGURE 18: Heterogeneity based on Region for LFSP Data.

FIGURE 19: Heterogeneity based on Gender for LFSP Data.
V. CONCLUSION
This research proposed a causal tree selection method based on out-of-sample standard error. The procedure grows a large number of causal trees, each on a random sub-sample taken from the training data. The proposed method estimates standard errors of all the trees based on out-of-sample observations. The causal trees grown are ranked with respect to the standard errors, and the top M trees are selected. The top ranked trees are combined for the final ensemble. A novel OCTE is assessed based on simulated data, generated under five different designs and compared with common procedures, OLS, LASSO, ridge, causal tree and causal decision trees forest. For assessing the proposed OCTE, performance metrics MSE, RMSE, MAD and Pearson’s correlation coefficient (r) are used. In general, the proposed algorithm outperformed the rest of the methods in almost all the cases.

The OCTE method demonstrated improved prediction performance compared to the rest of the methods considered in the paper. Moreover, the method is effective in estimating heterogeneous causal treatment effects. Since the method uses only few accurate causal trees, the idea could be further be extended to mitigate the interpretability issue of the standard causal trees forest.

REFERENCES
[1] Jennie E Brand, Jiuhui Xu, Bernard Koch, and Pablo Geraldo. Uncovering societal effect heterogeneity using tree-based machine learning. Sociological Methodology, page 008117501993503, 2021.
[2] Robert E Schapire. The strength of weak learnability. Machine learning, 5(2):197–227, 1990.
[3] Pedro Domingos. Using partitioning to speed up specific-to-general rule induction. In Proceedings of the AAI-96 Workshop on Integrating Multiple Learned Models, pages 29–34. Citeseer, 1996.
[4] Tom Mitchell. Machine learning. McGraw hill Burr Ridge, 1997.
[5] Kagan Tumer and Joydeep Ghosh. Error correlation and error reduction in ensemble classifiers. Connection science, 8(3-4):385–404, 1996.
[6] Kamal M Ali and Michael J Pazzani. Error reduction through learning multiple descriptions. Machine learning, 24(3):173–202, 1996.
[7] Simon Bernard, Laurent Heutte, and Sébastien Adam. On the selection of decision trees in random forests. In 2009 International Joint Conference on Neural Networks, pages 302–307. IEEE, 2009.
[8] Nicolai Meinshausen. Node harvest. The Annals of Applied Statistics, pages 2049–2072, 2010.
[9] Thais Mayumi Oshiro, Pedro Santoro Perez, and José Augusto Baranaukasas. How many trees in a random forest? In International workshop on machine learning and data mining in pattern recognition, pages 154–168. Springer, 2012.
[10] Patrice Latinne, Olivier Debeir, and Christine Decaestecker. Limiting the number of trees in random forests. In International workshop on multiple classifier systems, pages 178–187. Springer, 2001.
[11] Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001.
[12] Jianqing Fan, Mark Farmen, and Irene Gijbels. Local maximum likelihood estimation and inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(3):591–608, 1998.
[13] Whitney K Newey. Kernel estimation of partial means and a general variance estimator. Econometric Theory, 10(2):1–21, 1994.
[14] Joan G Staniswalis. The kernel estimate of a regression function in likelihood-based models. Journal of the American Statistical Association, 84(405):276–283, 1989.
[15] Charles J Stone. Consistent nonparametric regression. The annals of statistics, pages 595–620, 1977.
[16] Robert Tibshirani and Trevor Hastie. Local likelihood estimation. Journal of the American Statistical Association, 82(398):559–567, 1987.
[17] Arthur Lewbel. A local generalized method of moments estimator. Economics Letters, 94(1):124–128, 2007.
[18] Bo E Honoré and Ekaterini Kyriazidou. Panel data discrete choice models with lagged dependent variables. Econometrica, 68(4):839–874, 2000.
[19] Liangqun Su, Irina Murtaazashvili, and Aman Ullah. Local linear glm estimation of functional coefficient iv models with an application to estimating the rate of return to schooling. Journal of Business & Economic Statistics, 31(2):184–207, 2013.
[20] James M Robins and Ya’acov Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. Statistics in medicine, 16(3):285–319, 1997.
[21] Susan Athey, Julie Tibshirani, and Stefan Wagner. Generalized random forests. The Annals of Statistics, 47(2):1148–1178, 2019.
[22] Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. Neural computation, 9(7):1545–1588, 1997.
[23] Lez Brediman. Bagging predictors. Machine learning, 24(2):123–140, 1996.
[24] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning, 40(2):139–157, 2000.
[25] Tin Kam Ho. The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8):832–844, 1998.
[26] Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. Bagging survival trees. Statistics in medicine, 23(1):57–91, 2004.
[27] Nicolai Meinshausen and Greg Ridgeway. Quantile regression forests. Journal of Machine Learning Research, 7(6), 2006.
[28] Donald WK Andrews. Tests for parameter instability and structural change with unknown change point. Econometrica: Journal of the Econometric Society, pages 821–856, 1993.
[29] Bruce E Hansen. Testing for parameter instability in linear models. Journal of policy Modeling, 14(4):517–533, 1992.
[30] Nils Lid Jholt and Alexander Koning. Tests for constancy of model parameters over time. Journal of Nonparametric Statistics, 14(1-2):113–132, 2002.
[31] Jukka Nyblom. Testing for the constancy of parameters over time. Journal of the American Statistical Association, 84(405):223–230, 1989.
[32] Achim Zeileis. A unified approach to structural change tests based on ml scores, f statistics, and ols residuals. Econometric Reviews, 24(4):445–466, 2005.
[33] Achim Zeileis and Kurt Hornik. Generalized m-fluctuation tests for parameter instability. Statistica Neerlandica, 61(4):488–508, 2007.
[34] Werner Ploberger and Walter Kramer. The cusum test with ols residuals. Econometrica: Journal of the Econometric Society, pages 271–285, 1992.
[35] Gözde Bakirtürk and Derya Birant. Dtreeesim: A new approach to compute decision tree similarity using re-mining. Turkish Journal of Electrical Engineering & Computer Sciences, 25(1):108–125, 2017.
[36] Zhengshui Hu, Fei Li, Xuewei Wang, and Qiang Lin. Description length guided unified granger causality analysis. IEEE Access, 9:13704–13716, 2021.
[37] Martina Pastorino, Alessandro Montaldo, Luca Fronda, Ilksen Hedhili, Gabriele Moser, Sebastian B Serpico, and Josiane Zerubia. Multisensor and multireolution remote sensing image classification through a causal hierarchical markov framework and decision tree ensembles. Remote Sensing, 13(5):849, 2021.
[38] Fei Yan, Junqiao Ma, Mo Li, Ru Niu, and Tao Tang. An automated accident causal scenario identification method for fully automatic operation system based on stkp. IEEE Access, 9:11051–11064, 2021.
[39] Alex Javier Peñafiel Palacios, Jesús Estupiñán Ricardo, Iyo Alexis Cruz Piza, and Marceja Esther España Herrera. Phenomenological hermeneutical method and neurotrophic cognitive maps in the causal analysis of transgressions against the homeless. Neutrosophic Sets and Systems, pages 821–856, 1993.
randomized experiments for robust prediction. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pages 211–219, 2021.

[44] Yiğit Kıcık, Tim AD Henderson, and Andy Podgurski. Improving fault localization by integrating value and predicate based causal inference techniques. In 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE), pages 649–660. IEEE, 2021.

[45] Achim Zeileis, Torsten Hothorn, and Kurt Hornik. Model-based recursive partitioning. Journal of Computational and Graphical Statistics, 17(2):492–514, 2008.

[46] Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. The Journal of Machine Learning Research, 17(1):841–881, 2016.

[47] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. The Annals of Statistics, 43(4):1716–1741, 2015.

[48] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242, 2018.

[49] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189–1232, 2001.

[50] Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. arXiv preprint arXiv:1407.3939, 2014.

[51] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. Journal of the American Statistical Association, 101(474):578–590, 2006.

[52] Gérard Biau. Analysis of a random forests model. The Journal of Machine Learning Research, 13:1063–1095, 2012.

[53] Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research, 9(9), 2008.

[54] Gérard Biau and Erwan Scornet. A random forest guided tour. Test, 25(2):197–227, 2016.

[55] Peter Bühlmann and Bin Yu. Analyzing bagging. The annals of Statistics, 30(4):927–961, 2002.

[56] Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In International conference on machine learning, pages 665–673. PMLR, 2014.

[57] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. Machine learning, 63(1):3–42, 2006.

[58] Hemant Ishwaran and Udaya B Kogalur. Consistency of random survival forests. Statistics & probability letters, 80(13-14):1056–1064, 2010.

[59] Joseph Sexton and Petter Laake. Standard errors for bagged and random forest estimators. Computational Statistics & Data Analysis, 53(3):801–811, 2009.

[60] Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. arXiv preprint arXiv:1503.06388, 2015.

[61] Ruoning Zhu, Donglin Zeng, and Michael R Kosorok. Reinforcement learning trees. Journal of the American Statistical Association, 110(512):1770–1784, 2015.

[62] Gang Fang, Peng Xu, and Wenbin Liu. Automated ischemic stroke subtyping based on machine learning approach. IEEE Access, 8:118426–118432, 2020.

[63] Zardad Khan, Asma Gul, Aris Perperoglou, Miftahuddin Miftahuddin, Osama Mahmoud, Werner Adler, and Berthold Lausen. Ensemble of optimal trees, random forest and random projection ensemble classification. Advances in Data Analysis and Classification, 14(1):97–116, 2020.

[64] Juyong Li, Saisai Ma, Thuc Le, Lin Liu, and Jixue Liu. Causal decision trees. IEEE Transactions on Knowledge and Data Engineering, 29(8):257–271, 2016.

[65] Bruno Arpino and Mealli Fabrizia. The specification of the propensity score in multilevel studies. In Seventh scientific meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society. Salvatore Ingrassia and Roberto Rocci, 2009.

[66] Fan Li, Alan M Zaslavsky, and Mary Beth Landrum. Propensity score weighting with multilevel data. Statistics in medicine, 32(19):3373–3387, 2013.

[67] Megan S Schuler, Wanghuan Chu, and Donna Coffman. Propensity score weighting for a continuous exposure with multilevel data. Health Services and Outcomes Research Methodology, 16(4):271–292, 2016.

[68] Johan Zetterqvist, Stijn Vansteelandt, Yudi Pawitan, and Arvid Sjölander. Doubly robust methods for handling confounding by cluster. Biostatistics, 17(2):264–276, 2016.

[69] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2021.