

CFC-ATE: Causal Feature Construction via Average Treatment Effect

1st Asmae Lamsaf
NOVA-LINCS
IT: Instituto de Telecomunicações
University of Beira Interior, Portugal
a.lamsaf@ubi.pt

2nd Hugo Proença
IT: Instituto de Telecomunicações
University of Beira Interior, Portugal
hugomcp@di.ubi.pt .

3rd João C. Neves
NOVA-LINCS
University of Beira Interior, Portugal
jcneves@di.ubi.pt

Abstract—Dimensionality reduction is a crucial step in data preprocessing, particularly for high-dimensional datasets, where the excessive number of features increases the risk of overfitting in machine learning models. Traditional dimensionality reduction methods rely on statistical associations or the relative position of the feature embeddings in the hyper-space to map original features to a compact subspace that preserves the most relevant information of the data. However, these methods fail to capture the causal relationships among variables during the transformation process, leading to a loss of structural coherence of the data in low-dimensional spaces. By employing causal discovery and causal inference, it is possible to simplify these problems, effectively merging critical features while reducing both complexity and dimensionality. Our paper introduces a novel approach, Causal Feature Construction via Average Treatment Effect (CFC-ATE), which leverages causal discovery and inference to create more interpretable and reliable features for predictive modeling. Our methodology consists of the following phases: i) leveraging the causal structure of data through the inference of the causal graph. ii) transforming features through the use of the average treatment effect conditioned on the causal structure of the data. The experiments on diverse real-world datasets and synthetic datasets demonstrate the effectiveness of CFC-ATE in improving model performance by comparing it with three methods of feature selection and three benchmark dimensionality reduction techniques.

Index Terms—Causal discovery, Causal inference, Feature construction, Feature selection, Average treatment effect.

I. INTRODUCTION

In machine learning and data science, dealing with multi-dimensional data sets can be challenging, since they may contain irrelevant or irrelevant information, which can obscure useful observations and make results more ambiguous. To overcome this, it is necessary to reduce the number of components. By simplifying the dataset, models are faster, more accurate, and easier to obtain. Methods such as feature selection and dimensionality reduction help by eliminating redundant or redundant features. This not only improves the efficiency of the models but also makes it easier to identify meaningful patterns in the data.

Recently, exploring causality has become a key approach to managing complexity in different systems. This approach simplifies complex causal relationships by focusing on key

factors and ignoring less important ones. It reduces dimensionality through feature selection or by combining features based on their correlation or causality [1]. This process makes datasets easier to handle, resulting in predictive models that are simpler to interpret and more reliable, or in clustering that reduces the number of iterations in the process [35], [36]. Thus, causality has been crucial in exploring complex phenomena scientifically [2]–[8].

Dimensionality reduction methods often require transforming the original data to create new variables [9]. Some approaches achieve this through feature selection [37], which identifies the most relevant variables, or feature construction [10], [11], where new variables are created based on existing ones to capture important patterns in the data. Popular methods such as principal component analysis (PCA) and kernel PCA (KPCA) [10] simplify data by searching for key features based on how variables are related. Autoencoders work by filtering the data particles and then try to reconstruct it over the original data. They are useful for tasks such as data compression and anomaly detection [11], [12], but can sometimes overfit, especially with complex data. T-SNE (t-distributed stochastic neighbor embedding) is an alternative, which is usually high-dimensional data used for visualisation, but can be slow and difficult to maintain properly [13]. Causality refers to the idea that changes in one variable cause changes in another, often represented using Directed Acyclic Graphs (DAGs) [14], [15]. It plays a key role in machine learning, feature selection, and feature construction [8], [16], [17]. Existing methods like causality-PCA handle continuous variables but do not address datasets with both discrete and continuous variables [8]. Our paper introduces Causal Feature Construction via Average Treatment Effect (CFC-ATE), a new dimensionality reduction method that works with both variable types, leveraging causal relationships to create meaningful features. In experiments on real and synthetic datasets, CFC-ATE outperformed traditional feature selection methods and other causality-based techniques.

This research introduces the CFC-ATE method as a new approach for feature construction based on causal relationships.

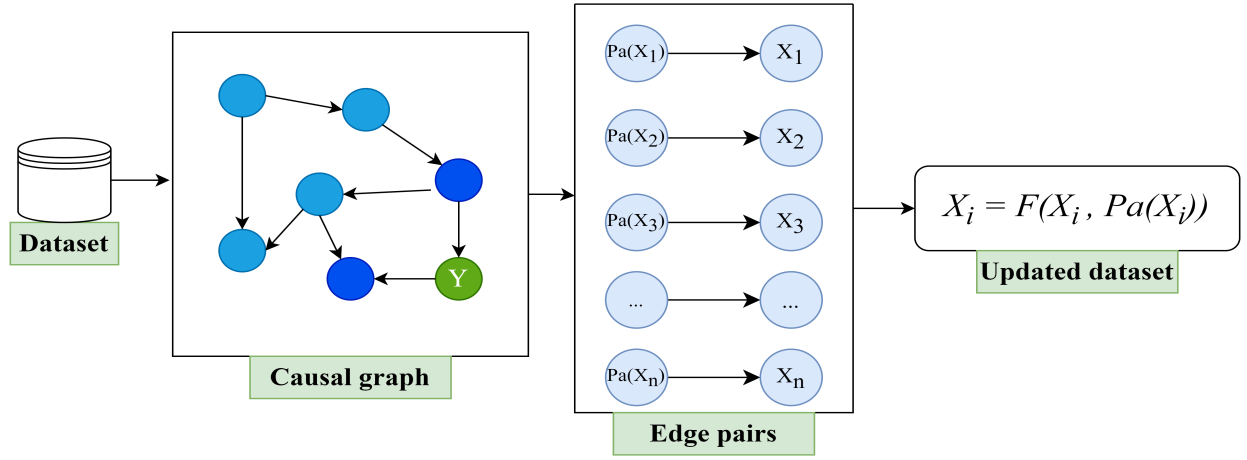


Fig. 1: The figure illustrates the process of CFC-ATE method, starting with a dataset to construct a causal graph, identifying parent-child pairs, and using these to update each variable using the function F . This creates a new dataset with causal information to improve predictive modeling.

Our key contributions are:

- We present the CFC-ATE method, which uniquely combines causal discovery and average treatment effect analysis for feature construction, improving both the interpretability and performance of machine learning models;
- We advance the feature construction field by integrating causal relationships into the process, thereby not only reducing dimensionality but also keeping essential causal information, which increases the robustness and applicability of the models across different scenarios.

II. RELATED WORK

Feature Construction. Feature extraction methods simplify high-dimensional datasets by transforming raw variables into more meaningful features [9]. Principal Component Analysis (PCA) and its variants, like Kernel PCA (KPCA), are commonly used techniques that reduce data complexity by identifying key features based on correlations, preserving important information for modeling [10]. As the field evolved, methods like autoencoders were introduced, which compress data into a simpler form and attempt to reconstruct it [11]. These are useful for tasks such as data compression and anomaly detection, but can struggle with complex data sets and overfit, especially in unsupervised learning [12]. T-SNE (t-distributed stochastic neighbour) is used visualization of high-dimensional data is the tool, which reduces its resolution while preserving local and global structure [13]. However, it is computationally expensive and sensitive to parameters, which can complicate the process.

Causality means that changes in a cause variable lead to changes in an effect variable, often implying correlation, though the reverse is not always true [14]. Causal relationships are usually sparser than correlations and can be represented by Directed Acyclic Graphs (DAGs), where edges show direct causal influence [15]. Causality is gaining importance in machine learning for improving model explanation, robust-

ness, and predictive performance, especially for intervention-based decisions [16]. However, using causal relationships for feature extraction remains underexplored. [8] addresses this by proposing causality-based PCA methods using average treatment effects for continuous variables, but it is limited to continuous data.

Feature construction using causality is crucial because it not only reduces the dimensionality of datasets but also preserves meaningful causal relationships between variables. A method that adapts to both discrete and continuous data is especially important, as it can be applied universally across different types of datasets. Furthermore, this approach is valuable for both regression and classification tasks, providing a flexible solution that enhances model performance by focusing on the underlying causal structure of the data. In our work, we propose a method adaptable to both discrete and continuous features to construct new features from the dataset, reducing dimensionality based on causal relationships between the features.

Causal discovery and inference. In our method, we use causal discovery to compute the relationships between variables in the dataset. It's important to understand the different methods available in this field in order to choose the one best suited to our specific case. Causal discovery methods are classified into three types: Constraint-based, Score-based, and Non-Gaussian [18]. Constraint-based methods, like the PC algorithm [20], use conditional independence tests and can handle hidden confounders with variants like FCI [20]. Score-based methods, such as Greedy Equivalence Search (GES), rank causal graphs using scores like BIC [22], with improvements like GIES [23] for scalability and GFIC for handling hidden confounders [25]. Non-Gaussian methods, including Independent Component Analysis (ICA) [26] and Additive Noise Models (ANMs) [27], identify causal links in nonlinear data. LiNGAM [28] deals with linear, non-Gaussian relationships without hidden confounders.

We use the constraint-based PC and FCI algorithms for their effectiveness in uncovering causal patterns and handling confounders, including hidden variables, in real-world data. These algorithms enable CFC-ATE to analyze causal links, essential for constructing effective features in machine learning models. To estimate the Average Treatment Effect (ATE) for feature construction. The existing methods are proposed for calculating or aiding to compute the ATE such as: Do-calculus [29], re-weighting methods [30], propensity score matching [31], tree-based methods [32], and regression adjustment [33]. In our work, we apply regression adjustment for its precision and ability to control confounding variables. This approach provides a straightforward and interpretable method, supporting our goal of building a robust and transparent framework for causal feature construction.

III. PROPOSED METHOD

We propose Causal Feature Construction via Average Treatment Effect (CFC-ATE), a method that creates meaningful features by preserving causal relationships and the average treatment effect. Using causal inference and graphical modeling, we reduce dimensionality while retaining key causal information. These features are then used for regression and classification models. Experiments on real and synthetic datasets validate CFC-ATE, demonstrating improved predictive performance and insights from complex data.

Algorithm 1 CFC-ATE Algorithm

Input: Dataset $D = \{X, Y\}$

Output: row: Constructed Feature set X'

- Construct the causal graph G from dataset D using causal discovery algorithms (PC and FCI).
- Generate the adjacency matrix A from graph G ,

$B = A$;

for each pair (X_i, X_j) **do**

if $a_{ij} = 1$ **then**

 calculate the $ATE(X_i, X_j)$;

$b_{ij} = ATE(X_i, X_j)$;

end

end

for each feature X_i **do**

if $X_i = Pa(X_j)$ **and** $b_{ij} \neq 0$ **then**

$X_j = X_j + b_{ij}$;

end

end

Initialize the transformed feature set $X' = \emptyset$;

for each feature X_j **do**

if $X_j = Pa(Y)$ **or** $X_j = child(Y)$ **then**

$X' = X' \cup \{X_j\}$;

end

end

Use the constructed features from X' as input for the regression or classification models to predict Y .

A. Causal Graph

In our approach, a causal graph $G = (V, E)$ is built, where V represents the features (nodes) in the dataset and E represents the relationships (edges) between them. We employ two causal discovery algorithms: the PC and FCI algorithms, both aimed at uncovering complex causal structures within the data. The PC algorithm is selected due to its scalability with large datasets and its ability to produce a Directed Acyclic Graph (DAG) under the assumption of no hidden confounders. The process begins with a fully connected undirected graph G that includes all variables V . For each variable pair (X, Y) , conditional independence is tested based on a conditioning set S . If X and Y are found to be conditionally independent given S , the edge between them is removed. The remaining edges are then directed to form a DAG that reflects the conditional independencies and dependencies observed in the data. The FCI algorithm is applied when hidden confounders or latent variables may exist. It extends the PC algorithm, allowing for a more thorough analysis by accounting for the presence of hidden variables [18].

FCI starts with a complete graph, removes edges based on conditional independence tests, and uses orientation rules to handle uncertainties from unobserved variables. This creates a Partial Ancestral Graph (PAG) that shows possible causal structures, including direct effects and confounders. FCI effectively identifies causal relationships in complex situations with hidden variables, providing a solid basis for the next steps of our approach.

B. Feature Construction

The CFC-ATE method is based on the estimation of the average treatment effect (ATE) between variables that are interconnected within a causal graph. In this context, the dataset D is divided into two components: the features X and the target Y , $X = \{X_1, X_2, \dots, X_n\}$ represents the set of features, with n denoting the total number of features. The adjacency matrix plays a crucial role in representing the causal graph created from the dataset $D = \{X, Y\}$. The adjacency matrix is a square matrix used to represent the connections between variables (n features of X and one target Y) in the graph. For a dataset with n features $X = \{X_1, X_2, \dots, X_n\}$ and target Y , the adjacency matrix A is an $(n+1) \times (n+1)$ matrix where each element a_{ij} indicates whether a direct causal link exists from feature X_i to X_j or feature X_i to Y .

The adjacency matrix A is structured by:

$$A = (a_{ij})_{i,j=1}^{n+1}, \quad (1)$$

$$\text{where } a_{ij} = \begin{cases} 1 & \text{if } X_i \text{ has direct causal effect to } X_j \\ 0 & \text{otherwise} \end{cases}.$$

The diagonal elements of A , a_{ii} , are typically set to 0, indicating that a variable does not have a causal effect on itself in the context of acyclic graphs. The Average Treatment Effect (ATE) is a fundamental concept in causal inference

[34], quantifying the expected difference in outcomes when a treatment or intervention is applied across a population. In the context of our method, ATE is employed to understand and quantify the causal impact of one feature X_i (treatment) on another X_j (outcome) within a dataset. The method systematically assesses the ATE between each feature pair linked in the causal graph to determine their causal influence on one another.

Applying the ATE to the adjacency matrix A in the context of the CFC-ATE method involves using the matrix to guide the calculation of ATE between connected features. For each identified pair (X_i, X_j) with $X_i = \text{Pa}(X_j)$ (X_i is the parent of X_j) and $X_j = \text{Child}(X_i)$ (X_j is the child of X_i), the ATE is determined to quantify the average impact of changing X_i on X_j . This involves comparing the outcomes of X_j when X_i assumes various values (e.g., treated vs. untreated), adjusting for the effects of other variables as necessary to isolate the causal impact. The average treatment effect is calculated by:

$$\text{ATE}(X_i, X_j) = E[X_j | X_i = 1] - E[X_j | X_i = 0]. \quad (2)$$

In our work, the Average Treatment Effect (ATE) is calculated using regression adjustment with the DoWhy library, where for each directed pair in the adjacency matrix A , a regression model is specified with the influencing feature as the treatment and the influenced feature as the outcome, adjusting for other variables as covariates. We construct a matrix B , similar in size to A , where each element represents the ATE of X_i on X_j . If $a_{ij}=1$. If $a_{ij}=0$ (no direct causal link), the corresponding entry in the ATE matrix would be 0, indicating no computed effect. The matrix B is calculated by:

$$B = (b_{ij})_{i,j=1}^{n+1}, \quad (3)$$

$$\text{where: } B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1(n+1)} \\ b_{21} & b_{22} & \cdots & b_{2(n+1)} \\ \vdots & \vdots & \ddots & \vdots \\ b_{(n+1)1} & b_{(n+1)2} & \cdots & b_{(n+1)(n+1)} \end{bmatrix} \text{ and}$$

$$b_{ij} = \begin{cases} \text{ATE}(X_i, X_j), & a_{ij} = 1 \\ 0, & a_{ij} = 0 \end{cases}.$$

The ATE values calculated are used for the construction or transformation of features. For instance, features could be combined based on their ATE to enhance their predictive relevance and reflect causal relationships. For each variable X_j , we add the ATE of $\text{Pa}(X_j)$ to X_j to the variable to X_j . The new dataset is calculated by:

$$X_j = F(X_i, X_j), \quad (4)$$

where : $X_i = \text{Pa}(X_j)$ if $b_{ij} \neq 0$, the function $F: F(X_i, X_j) = X_j + b_{ij}$ and $i,j = \{1, 2, \dots, n\}$. If X_j has multiple parent variables $X_{i1}, X_{i2}, \dots, X_{ik}$, we add each b_{ikj} (if $b_{ikj} \neq 0$) of each parent of X_j to X_j , then the overall transformation for X_j should account for the contributions from all parents.

TABLE I: Summary of the datasets used in the evaluation.

Name of dataset	Sample size	Number of variables	Task
Housing	20640	10	Regression
Ailerons	13750	41	Regression
Pole	14998	27	Regression
Connect-4	67557	43	Classification
Coil2000	9822	86	Classification
Thyroid	7200	22	Classification

Thus, if X_j is influenced by several parent variables X_{ik} with non-zero coefficients b_{ikj} , the transformation becomes:

$$X_j = X_j + \sum_{ik: b_{ikj} \neq 0} b_{ikj}. \quad (5)$$

This evaluation guides the creation of new features or the transformation of existing ones to capture the causal impact of each variable X_i to X_j to remove X_i and create new feature that contain the value of X_j and the causal impact of all the parents variables in causal graph, with the goal of improving predictive modeling by offering a feature set that accurately represents the underlying causal structure affecting the target variable Y . The new features selected from the constructed dataset are the features linked directly to the target Y in causal graph, specifically, a feature X_i is selected if either $a_{i(n+1)} \neq 0$ or $a_{(n+1)i} \neq 0$.

The CFC-ATE algorithm (Algorithm 1) outlines the steps from constructing a causal graph to selecting features for model input, focusing on clear procedural steps.

IV. RESULTS AND DISCUSSION

A. Implementation Details

We evaluated our CFC-ATE method on synthetic and real-world datasets using four regression and two classification models. We compared the performance of models using CFC-ATE features with those using standard feature construction and selection methods. Metrics such as accuracy and R-squared were used to measure improvements. As illustrated in Figure 2, the process of our experiments to evaluate our method using regression and classification models.

Datasets. The synthetic dataset, generated using the Erdős-Rényi model with Gaussian noise, simulates causal structures to test CFC-ATE's ability to uncover and use causal relationships for feature construction. For real datasets, we selected six from the KEEL repository: "Housing" (predicts Boston home prices), "Ailerons" (predicts aircraft control actions), "Pole" (predicts telecommunication pole costs), "Connect-4" (predicts game outcomes), "Coil 2000" (predicts car insurance policy acquisitions), and "Thyroid" (classifies patients by thyroid disease). These datasets provide a comprehensive test of CFC-ATE across various domains and tasks (refer to Table 1).

Classification and regression models. We selected the best parameters for each model to ensure optimal performance, then applied the same input parameters and datasets across all models and methods used in the evaluation. For regression tasks, we applied Linear Regression (LR), Decision Trees

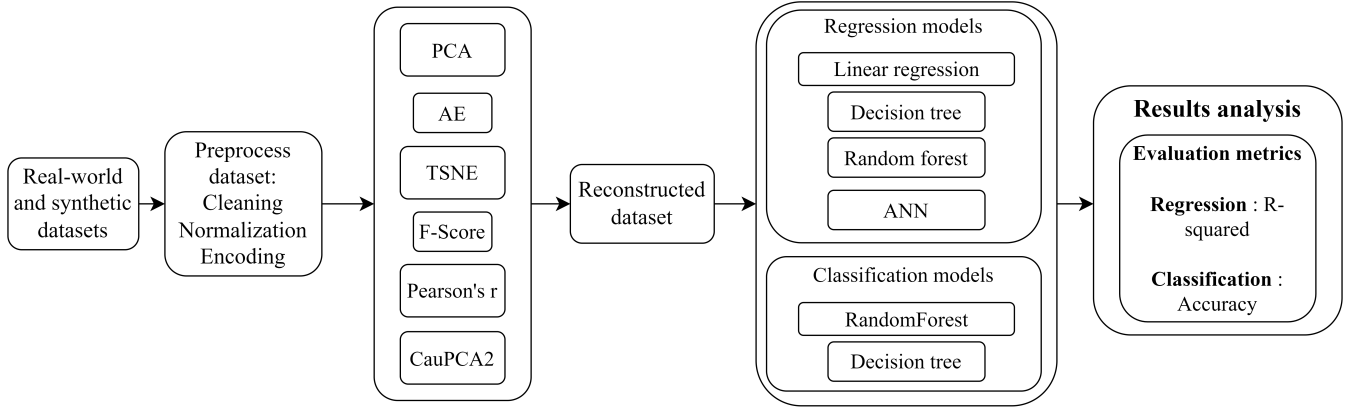


Fig. 2: The experimental process of the CFC-ATE method. First, the dataset is preprocessed, followed by the application of three dimensionality reduction techniques, three feature selection methods, and two causality-based feature construction methods. The reconstructed features are then used as inputs for regression and classification models to evaluate and compare the effectiveness of these methods using various evaluation metrics.

(DT), Random Forest (RF), and Artificial Neural Networks (ANN). For classification, we used Decision Trees (DT), Random Forest (RF). Each dataset was split 70% for training, 15% for testing and 15% for validation. We optimised our models by manually tuning parameters. For Decision Trees (DT) and Random Forest (RF), we set the maximum tree depth to 3 for DT and used 1,000 estimators with a maximum depth of 5 for RF to balance accuracy and prevent overfitting. For the Artificial Neural Network (ANN), we used the Adam optimiser with a learning rate of 0.001 and early stopping with a patience of 50 epochs to prevent overfitting.

Evaluation Metrics. We used R^2 to measure how well regression models explain variance in the target and accuracy to assess the proportion of correct predictions in classification tasks. These metrics provide a clear and reliable evaluation of CFC-ATE’s performance.

B. Results and Discussion

We tested CFC-ATE using PC (CFC-PC) and FCI algorithms (CFC-FCI) on various datasets, comparing them with dimensionality reduction techniques like AE, PCA, TSNE, CAU-PCA2 [8], and feature selection methods like F-score and Pearson’s r. For regression, we used the Housing, Ailerons, Pole, and a synthetic dataset with LR, DT, RF, and ANN models. For classification, we used Connect-4, Coil2000, and Thyroid datasets with DT and RF models. Datasets were split 70% for training, 15% for testing, and 15% for validation. Results are shown in Table 2 for regression and Table 3 for classification. The results are shown in Table 2 for regression tasks and Table 3 for classification tasks.

The CFC-ATE method consistently improved performance across most datasets and models, especially in the Housing and Ailerons datasets. In the Housing dataset, RF with CFC-FCI achieved an R-squared of 0.61, and in the Ailerons dataset, ANN with CFC-FCI reached an R-squared of 0.76. While results for the synthetic dataset were modest, CFC-ATE methods still outperformed PCA and TSNE, highlighting

TABLE II: Comparative results between proposed approach and other dimensionality reduction approaches in regression problems. Metric: R^2

Dataset	Models	CFC-PC	CFC-FCI	AE	PCA	TSNE	F-score	Pearson's r	CAU-PCA2
Housing	LR	0.49	0.52	0.30	0.07	0.11	0.45	0.45	0.25
	DT	0.47	0.49	0.16	0.09	0.008	0.41	0.41	0.03
	RF	0.60	0.61	0.38	0.05	0.49	0.55	0.55	0.27
	ANN	0.57	0.56	0.43	0.08	0.22	0.53	0.53	0.40
Ailerons	LR	0.71	0.74	0.52	0.61	0.51	0.66	0.66	0.34
	DT	0.58	0.61	0.52	0.49	0.55	0.61	0.60	0.32
	RF	0.70	0.71	0.32	0.58	0.74	0.65	0.65	0.23
	ANN	0.72	0.76	0.53	0.63	0.73	0.64	0.64	0.32
Pole	LR	0.65	0.23	0.22	0.20	0.004	0.20	0.20	0.14
	DT	0.94	0.28	0.49	0.49	0.01	0.26	0.26	0.16
	RF	0.97	0.32	0.91	0.90	0.14	0.24	0.24	0.07
	ANN	0.97	0.35	0.93	0.94	0.01	0.29	0.29	0.20
Synthetic dataset	LR	0.42	0.47	0.06	0.08	0.08	0.43	0.43	0.002
	DT	0.38	0.38	0.05	0.07	0.09	0.34	0.34	0.01
	RF	0.40	0.40	0.04	0.01	0.20	0.36	0.36	0.10
	ANN	0.47	0.47	0.04	0.05	0.15	0.43	0.43	0.001

the benefits of causal inference in feature construction. In the Pole dataset, TSNE’s low performance indicates it may not be suitable for extracting key information for accurate predictions, as it is designed for visualization rather than feature reduction for modeling. PCA and TSNE showed lower performance on the Housing and Synthetic datasets compared to CFC-PC and CFC-FCI, suggesting these methods may not effectively encode the information needed for accurate predictive modeling. PCA’s linear transformation may miss important data patterns, while TSNE’s transformation does not enhance data discriminability. Classification results on datasets like Connect-4 and Coil2000 further demonstrated the robustness of CFC-ATE, with RF and CFC-FCI showing strong outcomes, such as an accuracy of 93.19% in Coil2000. These results highlight the effectiveness of integrating causal discovery and inference in feature construction, showcasing the potential of the CFC-ATE method to enhance predictive modeling, regression and classification tasks across diverse applications.

V. CONCLUSION

This work presents the Causal Feature Construction via Average Treatment Effect (CFC-ATE) method, which improves

TABLE III: Comparative results between proposed approach and other dimensionality reduction approaches in classification problems. Metric: Accuracy(%)

Dataset	Models	CFC-PC	CFC-FCI	AE	PCA	TSNE	F-score	Pearson's r	CAU-PCA2
Connect-4	DT	66	68	56	55	44	66	66	65
	RF	66	68	64	65	50	67	67	65
Coil2000	DT	92.12	92.51	86	88	89	92.18	92.48	53.57
	RF	92.84	93.19	91	92.77	92.18	91.94	92.36	71.42
Thyroid	DT	96.98	97.05	90.67	88.95	85.9	97.96	97.96	92.91
	RF	98.24	98.17	93.40	92.98	91.64	97.89	97.89	92.84

model interpretability and performance by using causal discovery to create concise, high-level features while preserving key causal relationships. Unlike traditional methods, CFC-ATE enhances feature interpretability and discriminability. Experiments on synthetic and real-world datasets demonstrate that CFC-ATE outperforms standard feature engineering, significantly boosting model performance. This research highlights the importance of causality in predictive modeling and opens the door for further advancements in causality-driven machine learning techniques.

ACKNOWLEDGMENT

This work was funded by FCT/MEC through national funds and co-funded by the FEDER—PT2020 partnership agreement under the projects UIDB/50008/2020, POCI-01-0247-FEDER-033395, and by NOVA LINES (UIDB/04516/2020) with the financial support of FCT/IP.

REFERENCES

- [1] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, and X. Wu, "Causality-based feature selection: Methods and evaluations," *ACM Comput. Surv. (CSUR)*, vol. 53, no. 5, pp. 1–36, 2020.
- [2] K. Chalupka, T. Bischoff, P. Perona, and F. Eberhardt, "Unsupervised discovery of El Nino using causal feature learning on microlevel climate data," *arXiv preprint arXiv:1605.09370*, 2016.
- [3] J. Dubois, F. Eberhardt, L. K. Paul, and R. Adolphs, "Personality beyond taxonomy," *Nature Human Behaviour*, vol. 4, no. 11, pp. 1110–1117, 2020.
- [4] J. Dubois, H. Oya, J. M. Tyszka, M. Howard III, F. Eberhardt, and R. Adolphs, "Causal mapping of emotion networks in the human brain: Framework and initial findings," *Neuropsychologia*, vol. 145, p. 106571, 2020.
- [5] K. Chalupka, P. Perona, and F. Eberhardt, "Visual causal feature learning," *arXiv preprint arXiv:1412.2309*, 2014.
- [6] A. Geiger, K. Richardson, and C. Potts, "Neural natural language inference models partially embed theories of lexical entailment and negation," *arXiv preprint arXiv:2004.14623*, 2020.
- [7] A. Geiger, H. Lu, T. Icard, and C. Potts, "Causal abstractions of neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 9574–9586, 2021.
- [8] J. Liu, Y. Xu, and Y. Chen, "Causality-based PCA Methods for Condition Modeling of Mechatronic Systems," *IEEE Transactions on Industrial Informatics*, 2023.
- [9] P. Ray, S. S. Reddy, and T. Banerjee, "Various dimension reduction techniques for high dimensional data analysis: a review," *Artificial Intelligence Review*, vol. 54, no. 5, pp. 3473–3515, 2021.
- [10] M. Malekipirbazari, V. Aksakalli, W. Shafqat, and A. Eberhardt, "Performance comparison of feature selection and extraction methods with random instance selection," *Expert Systems with Applications*, vol. 179, p. 115072, 2021.
- [11] M. A. Kramer, "Autoassociative neural networks," *Computers & Chemical Engineering*, vol. 16, no. 4, pp. 313–328, 1992.
- [12] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [13] L. V. D. Maaten, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579, 2008.

- [14] P. J. Leary, "Causality, correlation, and cardiac disease: Does smoking cause cardiac hypertrophy and diastolic dysfunction?," *Circulation: Cardiovascular Imaging*, vol. 9, no. 9, p. e005441, 2016.
- [15] Y. Zhou, P. X.-K. Song, and X. Wen, "Structural factor equation models for causal network construction via directed acyclic mixed graphs," *Biometrics*, vol. 77, no. 2, pp. 573–586, 2021.
- [16] K. Kuang, L. Li, Z. Geng, L. Xu, K. Zhang, B. Liao, H. Huang, P. Ding, W. Miao, and Z. Jiang, "Causal inference," *Engineering*, vol. 6, no. 3, pp. 253–263, 2020.
- [17] K. Yu, L. Liu, J. Li, W. Ding, and T. D. Le, "Multi-source causal feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2240–2256, 2019.
- [18] G. Gendron, M. Witbrock, and G. Dobbie, "A survey of methods, challenges and perspectives in causality," *arXiv preprint arXiv:2302.00293*, 2023.
- [19] A. R. Nogueira, A. Pugnana, S. Ruggieri, D. Pedreschi, and J. Gama, "Methods and tools for causal discovery and causal inference," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 12, no. 2, p. e1449, 2022.
- [20] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT Press, 2001.
- [21] J. Pearl, *Causality*. Cambridge University Press, 2009.
- [22] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of Machine Learning Research*, vol. 3, no. Nov, pp. 507–554, 2002.
- [23] A. Hauser and P. Bühlmann, "Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2409–2464, 2012.
- [24] J. Ramsey, M. Glymour, R. Sanchez-Romero, and C. Glymour, "A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images," *International Journal of Data Science and Analytics*, vol. 3, pp. 121–129, 2017.
- [25] J. M. Ogarrio, P. Spirtes, and J. Ramsey, "A hybrid causal search algorithm for latent variable models," in *Proc. Conf. Probabilistic Graph. Models*, 2016, pp. 368–379.
- [26] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4–5, pp. 411–430, 2000.
- [27] P. O. Hoyer, A. Hyvärinen, R. Scheines, P. L. Spirtes, J. Ramsey, G. Lacerda, and S. Shimizu, "Causal discovery of linear acyclic models with arbitrary distributions," *arXiv preprint arXiv:1206.3260*, 2012.
- [28] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-Gaussian acyclic model for causal discovery," *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.
- [29] J. Pearl, "Causal inference in statistics: An overview," 2009.
- [30] G. W. Imbens and D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.
- [31] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.
- [32] W.-Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.
- [33] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Comput. Surv. (CSUR)*, vol. 53, no. 4, pp. 1–37, 2020.
- [34] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of Educational Psychology*, vol. 66, no. 5, p. 688, 1974.
- [35] I.-D. Borlea, "Centroid Update Approach to K-Means Clustering," *Advances in Electrical and Computer Engineering*, vol. 17, no. 4, pp. 3–10, Nov. 2017.
- [36] I.-D. Borlea, R.-E. Precup, and A.-B. Borlea, "Improvement of K-means Cluster Quality by Post Processing Resulted Clusters," *Procedia Computer Science*, vol. 199, pp. 63–70, 2022, doi: <https://doi.org/10.1016/j.procs.2022.01.009>.
- [37] U. Kilic, Esra SARAC ESSIZ, and Mumine KAYA KELES, "Binary Anarchic Society Optimization for Feature Selection," *Romanian Journal of Information Science and Technology*, vol. 2023, no. 3–4, pp. 351–364, Sep. 2023.