CauTsFool: Causality-Driven Imperceptible Attacks for Time Series Classification

Anonymous ICME submission

***Abstract*—In recent years, significant progress has been made in applying RNNs to time series classification (TSC). Yet, most adversarial attacks to date have targeted feedforward models and image tasks, leaving RNN-based TSC underexplored and vulnerable. Moreover, the cyclical nature of RNNs and the high temporal sensitivity of time series complicate direct model differentiation and local optimization of adversarial samples. In this paper, we mainly introduce CauTsFool, a causality-driven framework that focuses perturbations on the most influential subsequences. This approach achieves near-perfect attack success rates (often exceeding 98%) while substantially reducing perturbation magnitudes. By guiding adversarial noise toward causally critical regions, CauTsFool ensures greater efficiency and imperceptibility than conventional baselines. Experiments on seven UCR datasets demonstrate that CauTsFool outperforms existing white-box and black-box methods in effectiveness, speed, and subtlety.**

***Index Terms*—Causal Inference, Adversarial Attack, Time Series Classification**

I. Introduction

Deep Neural Networks (DNNs) have become indispensable across numerous domains, yet their susceptibility to adversarial examples remains a critical concern [1]. Classical adversarial attacks, such as FGSM [2], C&W [3], and PGD [4], highlight that, given sufficient information about the target network, adversaries can generate highly effective perturbations.

In practice, acquiring full knowledge of a DNN is often infeasible, making white-box attacks difficult. Black-box methods [5]–[7], which approximate gradients via queries, can circumvent this, but frequently at the expense of stealth and efficiency. Grey-box attacks leverage a generative model trained with some (but not complete) access to the target, enabling swifter deployment while avoiding extensive querying. However, directly applying these techniques to time series data proves challenging due to its sequential nature: small perturbations can significantly disrupt temporal dynamics, degrading classification performance more severely than in image domains.

While progress has been made, most adversarial methods assume readily differentiable models or do not address the unique temporal sensitivity of time series. Moreover, existing strategies often rely on localized, sample-level optimization, potentially overlooking global structures that could yield more stable and imperceptible perturbations.

Despite the critical importance of Time Series Classification (TSC) and the prevalent use of RNNs [8], [9], there is limited research on crafting effective adversarial attacks tailored for these models. Approaches like introducing differentiability to RNNs [10] or turning to black-box solutions often scale poorly or achieve limited performance. Additionally, DNNs' internal semantic overlaps suggest that robust adversarial perturbations could be engineered to transfer across architectures, yet such insights remain underexplored in time series contexts.

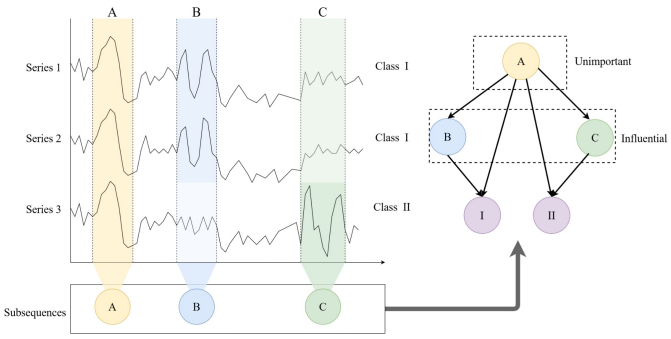

Figure 1. An illustrative example of three multivariate time series (Series 1, 2, and 3) and their corresponding subsequences: A (present in all three), B (in Series 1 and 2), and C (in Series 3). Series 1 and 2 belong to Class I, and Series 3 to Class II. After constructing a causal graph via causal inference and pruning non-influential connections, subsequences B and C emerge as critical factors influencing the final classification outcomes (I and II, respectively). In contrast, subsequence A, despite appearing in all three series, is deemed non-influential.

To address these limitations, we propose **CauTsFool**, an enhanced gray-box adversarial attack framework that builds upon the TSFool [11] methodology by integrating causal inference [12] for RNN-based time series classification. Unlike conventional methods that often rely on attention-based mechanisms—prone to misleading and spurious correlations [13], [14]—CauTsFool systematically identifies causally influential subsequences. By guiding perturbations toward these truly critical regions, our approach ensures more effective and stealthy adversarial attacks.

As illustrated in Figure 1, by focusing on these causally significant subsequences, CauTsFool substantially enhances the overall attack quality. Extensive experiments on 7 univariate and multivariate time series datasets from the UCR [15] archive confirm these benefits: CauTsFool attains near-perfect attack success rates (frequently exceeding 98%) and reduces the magnitude of perturbations compared to conventional baselines. These results validate the efficacy of a causality-driven strategy in producing imperceptible yet potent adversarial samples, thereby advancing the state of the art in adversarial time series classification.

Our contributions include:

- An enhanced use of causal inference to identify and precisely target the most influential segments in time series data, improving the localization of perturbations for adversarial attacks.
- An extension of the Camouflage Coefficient by incorporating causal insights, further refining the global optimization objective to achieve higher stealth and effectiveness of adversarial samples.
- The development of CauTsFool, an advanced gray-box adversarial method that builds upon TSFool by integrating causal inference, resulting in adversarial samples that are more effective, transferable, and imperceptible for RNN-based time series classification.

II. Related work

A. Causal Inference in Time Series.

There has been a propensity toward creating algorithms for causal inference on time series data. A mainstream of works is based on domain knowledge, artificially constructing causal graphs to solve time series problems in a particular field [16]. Despite the success of these works in their respective fields, they involved bringing in domain experts to build causality relations rather than automating the discovery of causality in time series. Moreover, for mining the causality in the time series, some works use Granger causality to analyze the time series [17]. Since these works use it to explore interactions inside time series, aside from the fact that it is actually investigating the causality between time series dimensions, Granger causality only means causality in the statistical sense, and it can not judge the internal mechanism between time series. To our knowledge, it is the first work on univariate time series causal discovery. We take causality between subsequences into account; in other words, we mine causal natural structures inside time series, which is the main novelty of our work.

B. Causality for Adversarial Attack

In recent years, significant progress has been made in applying causal inference to adversarial attacks. For instance, the CADE framework generates more realistic counterfactual adversarial examples by considering the causal generative process of the data. These approaches address fundamental questions of where to attack and how to attack, ensuring that generated adversarial examples are more applicable to real-world scenarios. However, such methods are primarily applied to image domains, where interventions are made on pixel-level features or latent variables to generate adversarial examples [18].

Despite the success of causal adversarial attacks in image-based tasks, there is limited research on applying causal inference to adversarial attacks in the time series domain. Time series data possess unique temporal dependencies and continuity, making causal-based adversarial attacks more challenging. Our work addresses this gap by introducing a method that leverages causal inference to identify pivotal subsequences in time series classification (TSC) tasks. This enables the generation of adversarial examples that are not only effective but also more transferable and imperceptible, advancing the field of adversarial attacks for RNN-based TSC.

III. Methodology

A. Mining Causal Natural Structures (MCNS)

1) Problem Definition: Given a time series $T = \{t_1, t_2, \ldots, t_n\}$ with optional label l_T, we aim to identify a causal natural structure $\mathbb{S}$ as a 4-tuple $< S_{sub}, l_T, \psi, C >$, where S_{sub} is a set of subsequences $T_{i,m}$, ψ represents causal relations, and C denotes causal strengths.

2) Finding Critical Data in Time Series: To identify critical data, we first determine subsequence length l based on the intrinsic period of T using the Fast Fourier Transform (FFT). The subsequence length l is set as $1/f$, where f is the dominant frequency. For consistency, we use the maximum l_T across the dataset.

Next, we extract k representative snippets s_T from each T using a domain-agnostic snippet discovery algorithm [19].

3) Constructing Inside Causal Graph: To construct the causal graph, the process consists of the following key steps:

Determine Factors: Cluster the subsequences into n classes using the k-shape algorithm [20]. Represent the time series T as a binary sequence, where the presence of a factor is denoted by 1.

Assemble Edges: Apply the GFCI algorithm [21] to infer causal relations and generate a Partial Ancestral Graph (PAG).

Impose Constraints: Refine the causal graph by removing edges originating from label factors and enforcing temporal constraints to ensure that effects follow their causes.

Finalize Graph: Retain edges of type $X \to Y$ that denote direct causation, discard edges of type $X \leftrightarrow Y$ representing confounding variables, and resolve uncertain edges using bootstrapping and the Bayesian Information Criterion (BIC) [22].

4) Calculating Causal Strength: We compute causal strength $\phi_{T,Y}$ for each edge $T \to Y$ using propensity score matching:

$$\begin{aligned}\phi_{T,Y} &= E[Y \mid do(T=1)] - E[Y \mid do(T=0)] \\ &= \left[\sum_{t_i=1} \Delta_{i,j} - \sum_{t_i=0} \Delta_{i,j}\right] /N.\end{aligned} \tag{1}$$

Here, $do(T=1)$ indicates an intervention on T, and $\Delta_{i,j}$ is the outcome difference between similar instances i and j.

5) Impregnating DNN with MCNS: **Refine Attention with Causal Strength:** Enhance LSTM attention by adding a causal strength-guided loss H_{cau}:

$$H_{cau} = \sum_{i=1}^{n} |a_{ij} - \zeta_i|. \tag{2}$$

The updated loss function is:

$$L = \alpha H(p,q) + \beta H_{cau}. \tag{3}$$

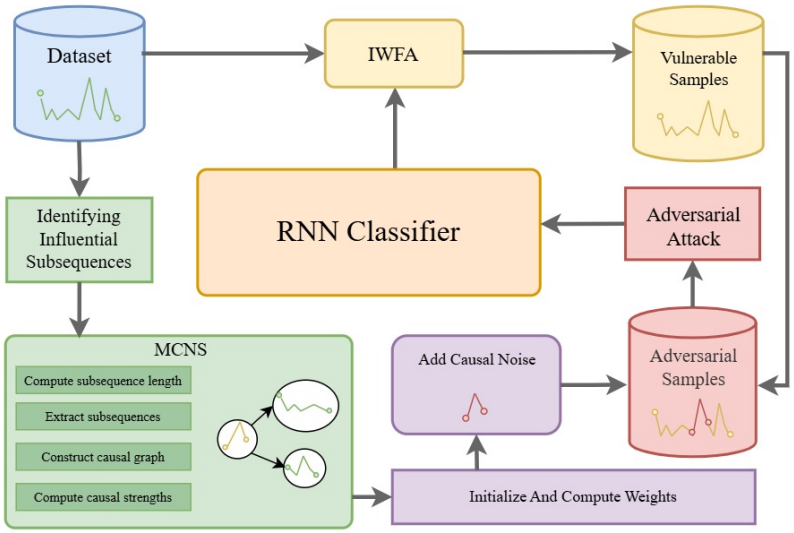

Figure 2. This figure illustrates the framework of CauTsFool, which integrates causal inference to generate adversarial samples for RNN-based time series classification. The process identifies influential subsequences, computes causal strengths, and adds targeted causal noise to craft imperceptible and effective adversarial attacks.

B. Integrating Causal Strength into Attack

In order to leverage the mined causal knowledge for more effective adversarial attacks, we first need to preprocess the causal weights to integrate them with the time series data.

The causal weights provide a relative measure of how influential each subsequence is to the overall time series classification. However, the raw weights may be on an arbitrary scale and contain zeros for non-causal subsequences. We need to transform these weights to be compatible with the time series and sampling process.

We propose the following processing steps:

1) Rescale the weights so non-causal subsequences have a small uniform weight instead of zero. This allows for a minimal sampling of less influential regions.

2) Boost the weights for highly influential subsequences for greater effect.

3) Apply softmax to normalize the weights to a probability distribution.

4) Ensure the weights for each time series row sum to 1, so they can be interpreted as a sampling distribution.

The algorithm below implements these steps:

This transforms the raw causal weights into a normalized distribution that can be integrated into the attack sampling process, focusing perturbations on highly influential subsequences for efficiency.

To leverage the mined causal knowledge for more effective and imperceptible adversarial attacks, we incorporate causal strength into the TSFool pipeline as follows:

2) **Interval sampling**: We initialize the sampling density between VNS and TPS proportionally to the normalized causal strength of each subsequence:

$$density(s_i) = \frac{strength(s_i)}{\sum_{j=1}^{n} strength(s_j)} \tag{4}$$

Algorithm 1 Time Series Adversarial Sample Generation Using MCNS and Causal Strength

Input: Time series dataset $\mathcal{D} = \{T_1, T_2, \ldots, T_m\}$
Labels $L = \{l_1, l_2, \ldots, l_m\}$
Total sampling steps M
Maximum noise limit ϵ_{max}
Output: Adversarial time series dataset $\mathcal{D}_{adv}$
$\mathcal{D}_{adv} \leftarrow \emptyset$
for each $T \in \mathcal{D}$ **do**
  Find dominant frequency f via FFT, set $l = 1/f$
  Select k key subsequences $S_{sub} = \{s_1, ..., s_k\}$ (e.g., from top patterns)
  Obtain causal strengths $C = \{\phi_{s_1}, ..., \phi_{s_k}\}$ (via causal analysis)
  Set $W = C$; let μ = mean of nonzero $W[i]$
  for $i = 1$ to k **do**
    if $W[i] = 0$ **then**
      $W[i] = \mu/2$
    else if $W[i] > \mu/2$ **then**
      $W[i] = 2 \times W[i]$
    end if
  end for
  Apply softmax to W and normalize so that $\sum W[i] = 1$
  for $i = 1$ to k **do**
    $\epsilon(s_i) = \epsilon_{max} \times \frac{1 - W[i]}{1 - \min(W)}$
    $s_i^{adv} = s_i + \mathcal{N}(0, \epsilon(s_i))$
  end for
  Replace s_i with s_i^{adv} in T to form T_{adv}
  Add T_{adv} to $\mathcal{D}_{adv}$
end for
return $\mathcal{D}_{adv}$

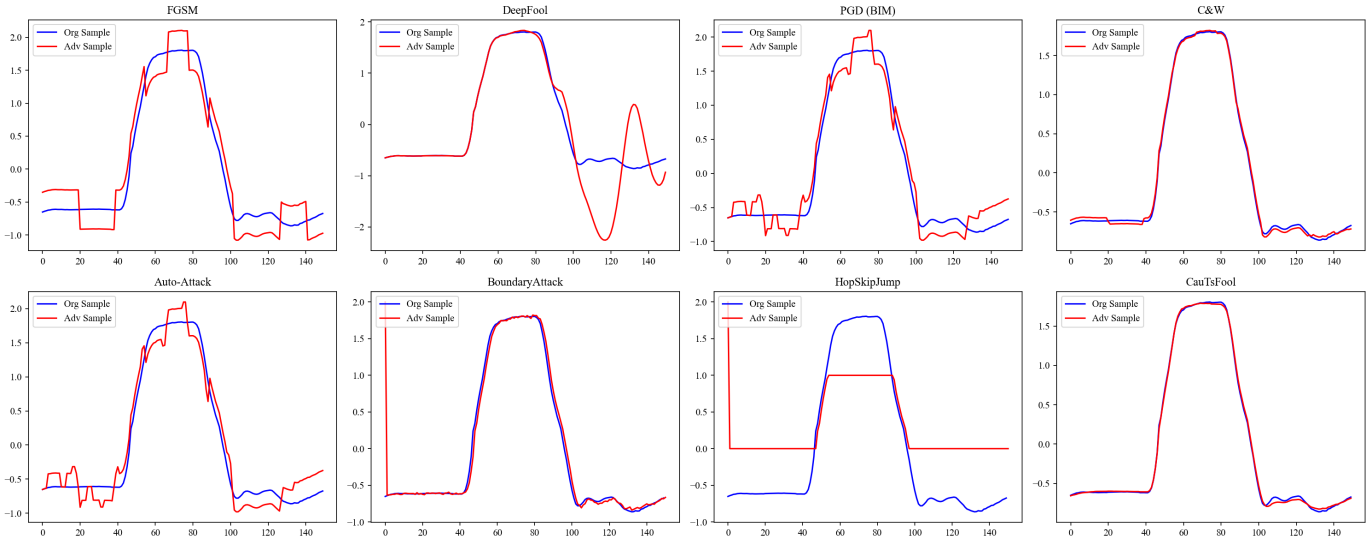

Figure 3. The figures illustrate the effects of adversarial attacks through seven benchmark methods and CauTsFool on UCR dataset. The results show the original sample (blue) and the adversarial sample (red) for each attack method.

and

$$numSteps(s_i) = density(s_i) * M, \tag{5}$$

where $strength(s_i)$ is the causal strength of subsequence s_i, n is the total number of subsequences, and M is the total sampling steps.

When moving the TPS towards VNS guidance through interval sampling, more fine-grained sampling points are allocated to subsequences with higher causal strength. This allows efficient crossing of classification boundaries by focusing perturbations on influential regions with minimal changes.

3) **Add noise**: The maximum magnitude of noise $\epsilon(s_i)$ added to each subsequence s_i is constrained as:

$$\epsilon(s_i) = \epsilon_{max} * \frac{1 - strength(s_i)}{1 - strength_{min}}, \tag{6}$$

where ϵ_{max} is the overall noise limit and $strength_{min}$ is the minimum strength among all subsequences. Hence, subsequences with higher causal strength will be assigned smaller maximum noise. This prevents detectable perturbations from being added to influential subsequences.

By tailoring the sampling density and noise magnitude based on causal significance, the attack is more focused on key subsequences to induce misclassification with imperceptible perturbations efficiently.

IV. Experiment

A. Experimental Settings and Details

Dataset. We use 7 datasets from the UCR time-series archive [23], including PowerCons (PC), GunPoint(GP), ECG200 (ECG), ECGFiveDays (EFD), LargeKitchenAppliances (LKA), DistalPhalanxOutlineCorrect (DPOC), and MiddlePhalanxOutlineCorrect (MPOC). These datasets span binary and multi-class classification tasks with varying data sizes and domains, ensuring a comprehensive evaluation of our method's robustness against adversarial attacks.

B. Attack Success Rate Analysis

We evaluate the Attack Success Rate (ASR) of CauTsFool against a range of baseline adversarial attacks, as shown in Table I. The ASR indicates the percentage of adversarial samples that successfully mislead the RNN classifier.

CauTsFool consistently demonstrates outstanding performance, achieving an overall ASR of **99.26%** on average—substantially higher than most baseline methods. Notably, on critical datasets such as PowerCons (PC), GunPoint (GP), and ECG200 (ECG), CauTsFool attains a **100%** ASR, outperforming standard approaches like FGSM, PGD, and BIM by margins exceeding 30–50 percentage points in some cases.

Even in those few datasets where CauTsFool does not achieve the absolute highest ASR—such as EFD (98.00% vs. 100.00%), LKA (98.00% vs. 100.00%), and MPOC (99.20% vs. 99.66%)—its results remain very close to the top performer and consistently secure a position among the top two methods.

These consistently high ASR results highlight not only CauTsFool's overall dominance but also its robustness across diverse time series scenarios. The method's ability to reliably produce near-perfect success rates, regardless of dataset variability, underscores its effectiveness as a powerful adversarial attack strategy for RNN-based time series classification.

C. Efficiency and Detectability of Adversarial Attacks

We evaluate the performance of adversarial attacks using two metrics: the Time Efficiency Score (TES) and the Camouflage Coefficient (CC). The TES balances the attack success rate (ASR) and runtime, defined as:

$$TES = ASR \times e^{-\alpha \cdot Time}, \tag{7}$$

where α controls time sensitivity (set to 1). Higher TES indicates better efficiency.

Table I
PERFORMANCE COMPARISON ACROSS DIFFERENT DATASETS AND ATTACK METHODS. BEST RESULTS ARE SHOWN IN **BOLD**, AND SECOND-BEST ARE MARKED WITH *. THE AVERAGES (AVG) ARE COMPUTED ACROSS ALL DATASETS FOR EACH METHOD.

Attack	Dataset	ASR	Time (s)	TES	CC
FGSM	PC	50.00%	**0.0072**	0.4963*	1.9586
	GP	50.67%	**0.0067**	0.5033*	1.8911
	ECG	52.00%	**0.0066**	0.5165	1.2350
	EFD	49.71%	**0.0067**	0.4937*	1.2357
	LKA	65.07%	**0.0397**	**0.6253**	0.7677*
	MPOC	57.04%	**0.0038**	0.5682	0.9831
	DPOC	47.83%	**0.0052**	0.4758	0.8955*
	AVG	53.19%	**0.0108**	0.5256*	1.2810
PGD	PC	50.00%	0.0802*	0.4614	1.9546
	GP	50.67%	0.0812	0.4671	1.8900
	ECG	62.00%	0.0644	0.5812	1.2550
	EFD	49.71%	0.3074	0.3655	1.2368
	LKA	65.07%	0.3577*	0.4549	**0.7674**
	MPOC	57.04%	0.0383	0.5489	0.9825*
	DPOC	52.54%	0.0419	0.5038*	0.9077
	AVG	55.29%	0.1388	0.4833	1.4277
BIM	PC	50.00%	0.1179	0.4443	1.9491
	GP	19.33%	0.0696*	0.1803	1.2743
	ECG	18.00%	0.0604	0.1694	1.3130
	EFD	15.80%	0.0604*	0.1487	1.1021
	LKA	65.87%	0.3643	0.4575*	0.7963
	MPOC	39.18%	0.0401	0.3763	0.9961
	DPOC	36.96%	0.0363	0.3564	0.9761
	AVG	35.02%	0.1070*	0.3047	1.2010
DeepFool	PC	50.56%	0.5058	0.3048	1.6860
	GP	54.67%*	0.3825	0.3729	1.7227
	ECG	91.00%*	0.0538	0.8623*	1.2713
	EFD	52.61%	0.3750	0.3615	1.1741
	LKA	77.07%	1.0470	0.2705	0.8375
	MPOC	85.57%	0.1572	0.7312*	0.9954
	DPOC	79.35%	1.0337	0.2822	1.0277
	AVG	70.12%	0.5079	0.4551	1.2450
HotSkipJump	PC	48.89%	299.1868	5.68E-131	1.4707
	GP	48.00%	37.6851	2.06E-17	1.4086
	ECG	**100.00%**	49.0404	5.04E-22	1.0000*
	EFD	41.46%	16.3885	3.16E-08	1.0924
	LKA	**100.00%**	316.6665	2.97E-138	0.8745
	MPOC	96.22%	38.8788	1.25E-17	1.0000
	DPOC	**100.00%**	40.6164	2.29E-18	1.0000
	AVG	76.37%	114.066	≈ 0	1.1209*
BoundaryAttack	PC	51.11%*	5.8544	0.0014	1.1554*
	GP	**100.00%**	26.0962	4.64E-12	1.0247*
	ECG	**100.00%**	6.7921	0.0011	1.0000*
	EFD	**100.00%**	4.4649	0.0115	**0.8677**
	LKA	62.13%	30.6556	3.02E-14	0.9502
	MPOC	**99.66%**	5.8343	0.0029	1.0000
	DPOC	**100.00%**	5.9697	0.0025	1.0000
	AVG	87.56%*	12.2382	0.0028	1.1426
CauTsFool	PC	**100.00%**	0.2054	**0.8142**	**0.7096**
	GP	**100.00%**	0.2140	**0.8072**	**0.7412**
	ECG	**100.00%**	0.0445*	**0.9564**	**0.7435**
	EFD	98.00%*	0.0705	**0.9132**	0.9757*
	LKA	98.00%*	0.8552	0.4166	0.7970
	MPOC	99.20%*	0.0103*	**0.9818**	**0.9004**
	DPOC	99.65%*	0.0121*	**0.9844**	**0.8674**
	AVG	**99.26%**	0.2017	**0.8391**	**0.8193**

The CC measures the imperceptibility of adversarial perturbations:

$$CC(\mathbf{x}^*) = \frac{D_{\text{orig}}(\mathbf{x}^*)}{D_{\text{adv}}(\mathbf{x}^*)}, \tag{8}$$

where $D_{\text{orig}}(\mathbf{x}^*)$ and $D_{\text{adv}}(\mathbf{x}^*)$ represent distances to the original and adversarial class distributions, respectively.

The Camouflage Coefficient (CC) quantifies how closely the adversarial sample resembles the original class distribution compared to the adversarial class distribution. Specifically, $D_{\text{orig}}(\mathbf{x}^*)$ indicates the distance between the adversarial sample and the original class distribution, while $D_{\text{adv}}(\mathbf{x}^*)$ represents the distance to the target adversarial class distribution. The ratio $\frac{D_{\text{orig}}(\mathbf{x}^*)}{D_{\text{adv}}(\mathbf{x}^*)}$ captures this relationship: a lower CC value suggests that the adversarial sample remains closer to the original class distribution and farther from the adversarial class distribution. This implies that the sample maintains the overall statistical characteristics of the original class, making the perturbation harder to detect on a global feature level. Therefore, a lower CC indicates better global camouflage.

Result Analysis. As illustrated in Table I, although CauTsFool does not achieve the absolute lowest average runtime, its overall time cost remains competitively low, with only a marginal difference of approximately 0.2 seconds from the fastest methods. This minor time gap becomes negligible when viewed in the context of CauTsFool's significantly higher Attack Success Rate (ASR) compared to FGSM, PGD, and BIM. In addition, the Time Efficiency Score (TES) metrics indicate that CauTsFool ranks as the top-performing or second-best method in most datasets. These findings highlight that CauTsFool provides highly efficient adversarial attacks, striking an admirable balance between runtime and overall performance.

In terms of imperceptibility, the Camouflage Coefficient (CC) presented in Table I demonstrates that CauTsFool significantly outperforms the baseline methods. Its average CC is consistently lower than that of the other attacks, and in the majority of datasets, it secures a top-two position. Such superior concealment capabilities are further illustrated by Figure 3, which displays adversarial samples generated on UCR datasets. Visual inspection confirms that CauTsFool's perturbations remain more subtle and harder to detect than those produced by competing approaches.

Overall Comparison. Taken together, these results underscore that CauTsFool not only outperforms existing baselines in terms of both efficiency and stealthiness, but it also achieves this dual excellence with minimal compromise in runtime. By attaining near-best efficiency scores while maintaining superior imperceptibility, CauTsFool stands as a robust and versatile adversarial attack strategy. This combination of rapid execution and high-quality, inconspicuous perturbations effectively distinguishes CauTsFool from current state-of-the-art baseline methods.

D. Ablation Study

Table II
COMPREHENSIVE COMPARISON OF CAUSAL AND NON-CAUSAL

Attack Method	ASR	Time	TES	CC
TsFool	90.49%	0.2513	0.7038	0.8944
CauTsFool	99.26%	0.2017	0.8391	0.8193
Boosting	**+8.77%**	**-19.74%**	**+19.24%**	**-8.40%**

To further assess the impact of causal guidance, we compare **CauTsFool** to **TsFool**, which does not incorporate causal strengths and relies on uniform weighting. As shown in Table II, integrating causal information leads to substantial improvements across all key metrics. Specifically, CauTsFool achieves a higher ASR (99.26% vs. 90.49%), attains a shorter running time (0.2017 s vs. 0.2513 s), and produces adversarial samples with a higher TES (0.8391 vs. 0.7038) and a lower CC (0.8193 vs. 0.8944). Compared to TsFool, these enhancements correspond to approximately an 8.77% increase in ASR, a 19.74% reduction in execution time, a 19.24% boost in TES, and an 8.40% decrease in CC. Together, these results highlight that incorporating causal structure significantly elevates effectiveness, efficiency, and stealthiness, underscoring the critical role of causal information in guiding adversarial perturbations for time series classification tasks.

V. CONCLUSION

In this paper, we propose CauTsFool, a framework that integrates causal inference into adversarial machine learning for time series classification. By incorporating causal insights into the traditional TsFool method, CauTsFool precisely targets critical subsequences, enhancing the efficiency and stealth of adversarial attacks.

Our approach minimizes perturbations to irrelevant regions, achieving over 90% attack success rates with low perturbation magnitudes. Extensive experiments on real-world datasets validate the effectiveness of CauTsFool in generating imperceptible adversarial samples.

This research marks a step forward in combining causal analysis and adversarial learning, offering a foundation for more reliable and interpretable time series classification models. Future work will explore extensions to multivariate time series and dynamic causal relationships.

REFERENCES

[1] Naveed Akhtar and Ajmal Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *Ieee Access*, vol. 6, pp. 14410–14430, 2018.

[2] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[3] Nicholas Carlini and David Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. Ieee, 2017, pp. 39–57.

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[5] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu, "Improving black-box adversarial attacks with a transfer-based prior," *Advances in neural information processing systems*, vol. 32, 2019.

[6] Andrew Ilyas, Logan Engstrom, and Aleksander Madry, "Prior convictions: Black-box adversarial attacks with bandits and priors," *arXiv preprint arXiv:1807.07978*, 2018.

[7] Yiwen Guo, Ziang Yan, and Changshui Zhang, "Subspace attack: Exploiting promising subspaces for query-efficient black-box attacks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] Daizong Ding, Mi Zhang, Fuli Feng, Yuanmin Huang, Erling Jiang, and Min Yang, "Black-box adversarial attack on time series classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, vol. 37, pp. 7358–7368.

[9] Daizong Ding, Mi Zhang, Yuanmin Huang, Xudong Pan, Fuli Feng, Erling Jiang, and Min Yang, "Towards backdoor attack on deep learning based time series classification," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 2022, pp. 1274–1287.

[10] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[11] Yanyun Wang, Dehui Du, Haibo Hu, Zi Liang, and Yuanhao Liu, "Tsfool: Crafting highly-imperceptible adversarial time series through multi-objective attack," in *ECAI*, 2024.

[12] Judea Pearl, "Causal inference in statistics: An overview," 2009.

[13] Sarthak Jain and Byron C Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.

[14] Sofia Serrano and Noah A Smith, "Is attention interpretable?," *arXiv preprint arXiv:1906.03731*, 2019.

[15] Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh, "The ucr time series archive," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1293–1305, 2019.

[16] Saurabh Mathur, Athresh Karanam, Predrag Radivojac, David M Haas, Kristian Kersting, and Sriraam Natarajan, "Exploiting domain knowledge as causal independencies in modeling gestational diabetes," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2023: Kohala Coast, Hawaii, USA, 3–7 January 2023*. World Scientific, 2022, pp. 359–370.

[17] Atalanti A Mastakouri, Bernhard Schölkopf, and Dominik Janzing, "Necessary and sufficient conditions for causal feature selection in time series with latent common causes," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7502–7511.

[18] Chao-Han Huck Yang, I-Te Hung, Yi-Chieh Liu, and Pin-Yu Chen, "Treatment learning causal transformer for noisy image classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6139–6150.

[19] Shima Imani, Frank Madrid, Wei Ding, Scott Crouter, and Eamonn Keogh, "Matrix profile xiii: Time series snippets: a new primitive for time series data mining," in *2018 IEEE international conference on big knowledge (ICBK)*. IEEE, 2018, pp. 382–389.

[20] John Paparrizos and Luis Gravano, "k-shape: Efficient and accurate clustering of time series," in *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, 2015, pp. 1855–1870.

[21] Juan Miguel Ogarrio, Peter Spirtes, and Joe Ramsey, "A hybrid causal search algorithm for latent variable models," in *Conference on probabilistic graphical models*. PMLR, 2016, pp. 368–379.

[22] Gideon Schwarz, "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978.

[23] Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, Gustavo Batista, and Hexagon-ML, "The ucr time series classification archive," October 2018, https://www.cs.ucr.edu/eamonn/time_series_data_2018/.