EXPLORING CAUSALITY AND EXPLAINABILITY IN TIME SERIES MODELS

MASTER'S THESIS

Ankan Kar*

Thesis Advisor: **Prof. K.V. Subrahmanyam**†

June 15, 2025

ABSTRACT

Causal inference is a powerful tool that allows us to move beyond surface-level correlations and uncover the true cause-and-effect relationships between variables. Whether we're working with static (non-time series) datasets or dynamic, time-dependent data, understanding *why* something happens is often more valuable than knowing *what* happens. In non-temporal datasets, causal inference helps identify directional influences between variables, often using statistical asymmetries, intervention models, or graphical structures to distinguish genuine causes from mere associations.

In this thesis, we examine the foundational principles of causal inference and how they apply across both non-time series and time series contexts. We explore a range of methodologies for identifying causal links and estimating their strength, with a focus on tools such as directed acyclic graphs (DAGs), structural equation models, and data-driven learning techniques.

A key part of our study focuses on applying these causal inference methods to time series data, where the order of events plays a crucial role. Time-dependent datasets—such as physiological signals of ECG—introduce new challenges, including temporal lags, feedback loops, and hidden confounders. To address these, we incorporate temporal structure into our causal models to better capture the dynamics at play.

Our future work will include implementing these causal techniques on real-world health data to identify significant events, such as abnormal spikes in physiological activity, and to determine the underlying causes. We also aim to build short-term predictive models to forecast metrics like heart rate, and to develop classification systems based on our causal findings. These insights will

*Computer Science, Chennai Mathematical Institute, ankank.mcs2023@cmi.ac.in

†Chennai Mathematical Institute, kv@cmi.ac.in

support early warnings, diagnostics, and informed decision-making. Ultimately, we plan to build machine learning models—including neural networks and random forests—that are enhanced by causal reasoning, making them both more interpretable and more effective for analyzing both static and time-based data.

The full implementation done as part of this thesis is available at this Github Link

Contents

1 Introduction

Understanding cause-and-effect relationships is at the heart of many scientific and practical questions. Whether we are trying to figure out what influences stock prices, how treatments affect patient health, or why certain trends appear in social data, we are ultimately asking: *what causes what?* This is where **causal inference** comes in—a powerful framework that enables us to move beyond surface-level correlations and uncover the true generative mechanisms behind observed data. It is widely used across Statistics, Mathematics, and Computer Science, providing tools to model interventions, simulate outcomes, and reason about alternative scenarios.

When dealing with variables that do not change over time—such as static features in clinical records—we can explore potential causal directions using statistical asymmetries. For example, if the relationship $X \to Y$ explains the data better than the reverse $Y \to X$, this may hint at a causal link. However, real-world systems often exhibit dynamics: physiological signals like ECGs, economic indicators, or climate variables evolve over time and interact in complex ways. This makes causal inference in **time series data** more challenging, as it requires accounting for *temporal precedence*, feedback loops, and potentially hidden confounding variables.

To rigorously model such systems, researchers rely on **graphical models**, particularly **Directed Acyclic Graphs (DAGs)**, where edges represent direct causal relationships. In the time series setting, these graphs are extended to capture not just contemporaneous interactions but also lagged dependencies. **Time Series Causal Models (TSCMs)** provide a principled framework to represent such temporal structures by incorporating assumptions like stationarity, finite memory, and time-invariant causal mechanisms. They can be viewed as structured extensions of **Vector Autoregressive (VAR)** models, where each variable is influenced by its own history and that of other variables, but with the added ability to learn a sparse causal graph rather than estimating all possible interactions.

In such frameworks, tools like **Granger Causality** serve as useful diagnostics for identifying predictive relationships, although they do not imply structural causation. On the other hand, models based on **Structural Equation Models (SEMs)** enable us to interpret data through functional causal relations and allow intervention-based reasoning. These models benefit from additional assumptions and statistical criteria such as the **Bayesian Information Criterion (BIC)** for graph selection, helping avoid overfitting by penalizing model complexity.

With high-dimensional data such as multivariate physiological signals, feature discovery becomes important. Here, methods like **Mutual Information** and **Transfer Entropy** can uncover both linear and nonlinear associations, enriching the process of causal feature selection. In machine learning pipelines, these causally-informed features can significantly enhance interpretability and robustness. For example, in our study, we use these to train classifiers such as the **Random Forest**, and also incorporate causal structure directly into deep learning architectures through our proposed **Causal Aware Convolutional Neural Network (CA-CNN)**, which augments standard CNNs with a causality map that captures inter-channel dependencies in feature space.

Ultimately, this work aims to demonstrate how causal discovery and reasoning can be integrated into time series analysis—especially in critical domains like healthcare, where understanding why a condition occurs is as vital as predicting when it will. Through a combination of theoretical modeling, data-driven learning, and causal feature selection, we construct interpretable and effective models capable of detecting and explaining key events in physiological signals such as ECGs. This causal lens not only enhances model performance but also supports actionable insights, making it an essential component in the development of next-generation time series analysis tools.

2 Causal Inference and Causal Models

Causal inference is concerned with identifying and reasoning about cause-and-effect relationships. Unlike statistical associations, which describe observed dependencies between variables, causal inference aims to uncover mechanisms that generate these dependencies. This distinction allows us to answer questions such as "What would happen if we intervene on a variable?" or "Was this outcome caused by a specific action?".

To formalize causality, Judea Pearl introduced a framework built upon probability theory, graphical models, and structural equations. These tools form the foundation of modern causal inference.

2.1 Basic Concepts in Probability Theory

Let X, Y, and Z be random variables. The probability distribution $P(X)$ captures our uncertainty about X. The conditional probability $P(Y|X)$ expresses our belief about Y given that X is known. Two variables X and Y are said to be *conditionally independent* given Z, denoted $X \perp\!\!\!\perp Y \mid Z$, if:

$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z).$$

Conditional independence plays a crucial role in both statistical reasoning and graphical modeling.

2.2 Graphical Models and d-Separation

A *Bayesian Network* is a directed acyclic graph (DAG) where nodes represent random variables and edges encode direct probabilistic dependencies. Each node X_i is associated with a conditional probability distribution $P(X_i \mid \text{Parents}(X_i))$.

Definition (d-Separation): Let G be a DAG and let X, Y, and Z be disjoint subsets of nodes. Then Z *d-separates* X and Y in G, written $(X \perp_d Y \mid Z)$, if every path from a node in X to a node in Y is blocked by Z. This implies conditional independence: $X \perp\!\!\!\perp Y \mid Z$ in all distributions compatible with G.

2.3 Causal Models

A *Causal Bayesian Network* extends a Bayesian Network by interpreting the edges as representing direct causal influences. The key difference is that causal models support **interventions**, allowing us to answer questions about what happens under manipulation.

Definition (Intervention): The operation $do(X = x)$ represents an intervention that sets variable X to a fixed value x, irrespective of its usual causes. The resulting interventional distribution is denoted $P(Y \mid do(X = x))$, which differs from the observational $P(Y \mid X = x)$ when confounding exists.

Proposition (Truncated Factorization Formula): Given a causal DAG G and a distribution P consistent with it, the post-intervention distribution after $do(X_i = x_i)$ is:

$$P(x_1, \ldots, x_n \mid do(X_i = x_i)) = \prod_{j \neq i} P(x_j \mid \text{Parents}(X_j)) \cdot \delta(x_i),$$

where $\delta(x_i)$ is the indicator fixing $X_i = x_i$.

2.4 Functional Causal Models

A more general framework uses *structural equation models (SEMs)*, where each variable X_i is a deterministic function of its parents and a noise term U_i:

$$X_i = f_i(\text{Parents}(X_i), U_i).$$

These models allow for defining counterfactuals and reasoning about hypothetical scenarios.

Definition (Counterfactual): The counterfactual Y_x is the value that variable Y would attain if variable X were set to x, possibly contrary to fact. Counterfactuals are central to causal questions like attribution and explanation.

2.5 Causal Discovery

Causal discovery is the process of learning the underlying causal structure from data. Under assumptions like the Causal Markov Condition and Faithfulness, algorithms like PC or FCI can recover parts of the causal graph.

Causal Markov Condition: In a causal DAG, each variable is conditionally independent of its non-descendants given its parents.

Faithfulness: All and only the conditional independencies in the data are entailed by the d-separations in the graph. From [9] we describe more formally it is described below:

Definition 2.1 (Faithfulness)**.** *A probability distribution P is said to be* faithful *to a directed acyclic graph (DAG) G if* ***all and only*** *the conditional independence relations that hold in P are exactly those that are entailed by the* Markov condition *applied to G.*

In other words, no additional conditional independence relations hold in P beyond those implied by the structure of G, and all such implied independencies do hold in P.

Formally, let $\mathcal{I}(P)$ denote the set of conditional independence relations true in P, and let $\mathcal{I}(G)$ denote those implied by the Markov condition on G. Then P is faithful to G if and only if

$$\mathcal{I}(P) = \mathcal{I}(G).$$

A distribution P is called faithful *(without reference to a specific graph) if there exists some DAG G such that P is faithful to G.*

Proposition (Identifiability): Under the Markov and Faithfulness assumptions, and in the absence of hidden confounders, the causal DAG is identifiable up to its Markov equivalence class from observational data.

2.6 Causal Inference vs Statistical Inference

Statistical models allow us to predict and estimate associations, while causal models allow us to simulate the effects of actions. Pearl summarizes this distinction in the "Ladder of Causation":

1. **Association:** Seeing, e.g., $P(Y \mid X)$
2. **Intervention:** Doing, e.g., $P(Y \mid do(X))$
3. **Counterfactuals:** Imagining, e.g., "Would Y have changed if X had been different?"

2.7 Conclusion

Causal inference provides a formal and practical framework to reason about interventions, counterfactuals, and explanations. By combining graphical models, probability theory, and structural equations, it becomes possible to uncover and validate causal claims in both observational and experimental settings. This foundation enables progress in fields ranging from medicine to economics to artificial intelligence, where understanding *why* something happens is as important as knowing *what* happens.

3 Causal Graphical Models

Causal graphical models (CGMs) provide a powerful framework for representing and reasoning about cause-and-effect relationships between variables. These models combine graph theory and probability to capture the structure of causal systems using directed acyclic graphs (DAGs), where:

- Nodes represent random variables.
- Directed edges ($X \to Y$) indicate that X is a direct cause of Y.

Each variable X_i in the system is modeled as a function of its parent variables and an exogenous (independent) noise term:

$$X_i := f_i(\mathrm{PA}_i, U_i)$$

where PA_i denotes the parents of X_i in the DAG, and U_i are jointly independent.

Key concepts in this theory involve defining causal structures and models as follows:

Definition 3.1 (Causal Structure in Pearl (2000) p.44)**.** *A causal structure of a set of variables V is represented as a directed acyclic graph (DAG) where each node corresponds to a distinct variable in V, and each link indicates a direct functional relationship among the corresponding variables.*

Definition 3.2 (Causal Model in Pearl (2000) p.44)**.** *A causal model is defined as a pair $M = \langle D, \Theta \rangle$, consisting of a causal structure D and a set of parameters Θ_D that are compatible with D. The parameters Θ_D assign a function $x_i = f_i(pa_i, u_i)$ to each variable $X_i \in V$ and a probability measure $P(u_i)$ to each random disturbance u_i, where P_{A_i} denotes the parents of X_i in D and each U_i is independently distributed according to $P(u_i)$.*

3.1 Blocked and Unblocked Paths

Definition 3.3 (Path)**.** *A **path** between two variables X and Y in a graph is any sequence of connected edges (regardless of direction) that starts at X and ends at Y. This includes both directed and undirected connections.*

Definition 3.4 (Blocked and Unblocked Paths (d-separation))**.** *A path between X and Y is said to be **blocked** (or inactive) by a set of nodes Z if one of the following conditions holds:*

- *The path contains a chain or fork: $A \to B \to C$ or $A \leftarrow B \to C$, where the middle node $B \in Z$ (i.e., B is conditioned on).*
- *The path contains a collider: $A \to B \leftarrow C$, and neither B nor any of its descendants are in Z (i.e., not conditioned on).*

*If none of these conditions hold, the path is said to be **unblocked** (or active) given Z.*

Definition 3.5 (d-separation)**.** *Two variables X and Y are said to be **d-separated** by a set of variables Z if all paths between X and Y are blocked given Z. Otherwise, they are **d-connected** (i.e., there exists at least one unblocked path).*

Important Note: In a causal DAG, even if there is **no directed path** from X to Y, it is still possible that X and Y are dependent due to an *indirect path* through:

- A **confounder**: e.g., $X \leftarrow Z \rightarrow Y$
- A **collider**: e.g., $X \rightarrow Z \leftarrow Y$

These are not directed causal paths but can carry statistical association unless they are properly blocked by conditioning.

For example, in the graph

$$X \leftarrow Z \rightarrow Y$$

there is no directed path from X to Y, yet they are statistically dependent due to the common cause Z (confounding). Likewise, in

$$X \rightarrow Z \leftarrow Y$$

X and Y are marginally independent, but conditioning on the collider Z opens a path of dependence.

Understanding blocked and unblocked paths is crucial to determining conditional independence via d-separation.

A **Causal Graphical Model (CGM)** is a general framework that uses graphs to represent causal relationships, which may or may not include probabilistic information. A **Causal Bayesian Network (CBN)** is a specific type of CGM that uses a directed acyclic graph (DAG) along with a joint probability distribution to model both causality and uncertainty. In essence, every CBN is a CGM, but not all CGMs are CBNs.

3.2 Basic Causal Structures

In a causal model, variables play different roles depending on their position in the underlying graph. Common patterns include chains, forks, and colliders, which help us reason about causal influence and statistical dependencies.

Definition 3.6 (Mediator)**.** *We say that a variable Z is a **mediator** between variables X and Y if the causal graph contains the structure:*

$$X \rightarrow Z \rightarrow Y.$$

Here, the effect of X on Y is transmitted indirectly through Z.

Definition 3.7 (Confounder)**.** *A variable Z is called a **confounder** for the relationship between X and Y if it causally influences both:*

$$Z \rightarrow X, \quad Z \rightarrow Y.$$

In this case, Z is a common cause of X and Y, potentially inducing spurious association between them.

Definition 3.8 (Collider)**.** *A variable Z is a **collider** on the path between X and Y if both X and Y causally influence Z:*

$$X \rightarrow Z \leftarrow Y.$$

Unlike mediators and confounders, conditioning on a collider (or its descendants) opens a path of dependence between X and Y.

Definition 3.9 (Marginal Dependence)**.** *Two variables X and Y are said to be **marginally dependent** if they are statistically dependent without conditioning on any other variables. That is,*

$$P(X, Y) \neq P(X)P(Y), \quad \textit{or equivalently,} \quad P(X|Y) \neq P(X).$$

If $P(X, Y) = P(X)P(Y)$, we say X and Y are marginally independent.

1. Chain (Mediation) In a Causal model of chain structure represents a sequence of causal influence:

$$X \to Z \to Y$$

Here, X affects Y indirectly through an intermediate variable Z (a *mediator*). In such a structure:

- X and Y are dependent marginally.
- Conditioning on Z blocks the path, making $X \perp\!\!\!\perp Y \mid Z$ (conditional independence).

2. Fork (Confounding) This structure represents a common cause:

$$Z \to X, \quad Z \to Y$$

The variable Z is a *confounder* that influences both X and Y. In this case:

- X and Y are dependent marginally.
- Conditioning on Z removes the dependence: $X \perp\!\!\!\perp Y \mid Z$.

3. Collider (Selection Bias) In a collider structure:

$$X \to Z \leftarrow Y$$

The variable Z is a *collider* because it is influenced by both X and Y. In contrast to the fork:

- X and Y are marginally independent.
- Conditioning on Z or any of its descendants induces dependence: $X \not\perp\!\!\!\perp Y \mid Z$.

This is a key source of *selection bias* and explains phenomena like Berkson's paradox.

3.3 Composite Structures

4. Confounding and Adjustment In practice, confounding arises when we attempt to estimate the effect of X on Y, but a common cause Z also affects both:

$$Z \to X \to Y, \quad Z \to Y$$

To recover the true causal effect of X on Y, we must adjust for Z by conditioning on it in our analysis.

5. Mediation and Indirect Effects Causal mediation analysis seeks to decompose the total effect of a treatment X on an outcome Y into:

- **Direct effect:** the path $X \to Y$
- **Indirect effect:** the path $X \to M \to Y$, where M is a mediator

Understanding mediation helps isolate mechanisms behind observed effects.

6. Front-door and Back-door Criteria These are graphical criteria used to identify causal effects:

- **Back-door criterion:** A set of variables Z satisfies the back-door criterion relative to (X, Y) if it blocks all paths from X to Y that enter X through a back-door (i.e., confounding paths), and no member of Z is a descendant of X.
- **Front-door criterion:** A set Z satisfies the front-door criterion if Z mediates all causal paths from X to Y, and X does not cause Z through other unblocked paths.

These tools enable identification of causal effects from observational data under the right graphical conditions.

3.4 Key Properties of Causal DAGs

- **Markov Property:** Each variable is conditionally independent of its non-descendants given its parents.
- **d-separation:** A graphical criterion for conditional independence. If a set of nodes Z d-separates X from Y, then $X \perp\!\!\!\perp Y \mid Z$.
- **Faithfulness:** A distribution is faithful to a graph if all and only the conditional independencies in the distribution correspond to d-separations in the graph.

3.5 Observational Equivalence

Proposition 3.1 (VermaPearl1990)**.** *Two directed acyclic graphs (DAGs) (models) are observationally equivalent if and only if they have the same skeletons and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow.*

- Statistical methods are limited in inferring causal directions in DAGs due to observationally equivalent models.

- Only v-structures or causal directions creating new v-structures or cycles are inferrable.

- Some arrow directions in a DAG cannot be uniquely determined from data.

3.6 Why Use Causal Graphical Models?

Causal graphical models offer several practical benefits:

- They make assumptions about the causal structure explicit.
- They allow us to predict the effects of interventions (using do-calculus).
- They help identify biases (e.g., confounding, selection bias) and suggest strategies to correct for them.
- They provide a formal language to communicate causal assumptions and derive estimands.

By capturing the causal architecture of a system, CGMs enable more robust reasoning under distribution shifts, better policy design, and a deeper understanding of the mechanisms at play.

4 Statistical to Causal Learning

Traditional statistical learning focuses on finding patterns and associations in data, assuming that the data points are independent and come from the same distribution (i.i.d.). In this approach, models learn to predict outcomes based on observed correlations within the data, and they tend to work well as long as the conditions during training and testing are similar.

However, these models face challenges when things change, such as:

- When the data distribution shifts (for example, if new data looks different from the old),
- When we want to understand what happens if we intervene or change something deliberately,
- Or when we want to reason about "what if" scenarios—what might have happened if things were different.

This is where causal learning comes in. Instead of just capturing correlations, causal learning tries to understand the underlying processes that generate the data. It does this through structural causal models (SCMs), where each variable is defined by a function of its direct causes plus some random noise:

$$X_i := f_i(\mathrm{PA}_i, U_i)$$

Here, PA_i stands for the parents of X_i in a causal graph, and the U_i are independent noise factors.

Causal models give us several important advantages:

- They capture the direction of cause and effect (so X causing Y is different from Y causing X),
- They let us calculate what would happen if we actually intervene on the system (using the do-operator),
- They tend to be more stable and reliable across different settings because the causal mechanisms themselves don't change arbitrarily.

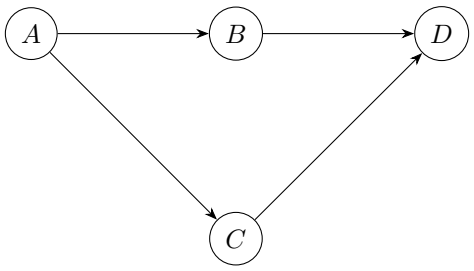
At the heart of causal reasoning is the idea that observed relationships in data come from real underlying causes. Reichenbach's Common Cause Principle sums this up nicely: if two things are statistically related, there's either a direct causal link or a shared cause behind it.

Moving from statistical to causal learning means shifting our focus—from just modeling what we see, to understanding how the system works underneath. This deeper insight allows us to predict how changes or interventions will affect outcomes, even when the data distribution changes.

5 Markov Properties on Causal Graph

Markov Property (Graphical Models). A probability distribution satisfies the *Markov property* with respect to a graph if each variable is conditionally independent of its non-descendants given its parents. This allows the joint distribution to factorize according to the graph's structure.

Here an example of an causal graphical structure:



Factorization of the joint distribution:

$$P(A, B, C, D) = P(A) \cdot P(B \mid A) \cdot P(C \mid A) \cdot P(D \mid B, C)$$

Local Markov Property: Each variable is conditionally independent of its non-descendants given its parents:

- $B \perp\!\!\!\perp C \mid A$
- $C \perp\!\!\!\perp B \mid A$
- $D \perp\!\!\!\perp A \mid B, C$

Global Markov Properties (via d-separation): Let's analyze conditional independencies:

- $A \perp\!\!\!\perp D \mid B, C$ (Once B and C are known, A gives no extra info about D)
- $B \perp\!\!\!\perp C \mid A$ (Given their common cause A, B and C are conditionally independent)
- $B \perp\!\!\!\perp C$ is **not** true unconditionally (they share common cause A)
- $A \perp\!\!\!\perp D$ is **not** true unconditionally (there is a directed path from A to D)
- $B \perp\!\!\!\perp D \mid C$ is **not** true (conditioning on C doesn't block $B \to D$)
- $C \perp\!\!\!\perp D \mid B$ is also **not** true

Summary: This example illustrates how a DAG imposes a rich set of conditional independencies, and how the local and global Markov properties guide us in understanding which variables are (in)dependent given others.

6 Search Algorithms

Inferring the underlying causal graph from data is known in the literature as *graph learning*. There are generally three main approaches to solving this problem. The first approach relies on a sequence of statistical tests involving *partial correlation coefficients*. These tests begin with low-order partial correlations—where only a few variables are conditioned on—and gradually progress to higher-order ones, involving more conditioning variables. Hoover (2005) provides an intuitive explanation of how this process works, while Spirtes et al. (2000) offer a more detailed and technical discussion of such algorithms. One of the most well-known methods in this category is the PC algorithm, and a simplified version of it is outlined below.

Algorithm 1 PC Algorithm

Input: Observations of a set of variables X generated from a DAG model.

Output: A pattern (DAG) compatible with the data generating DAG.

Start with a full undirected graph.

for each pair of variables $(X_i, X_j) \in X$ **do**

  Search a subset $S_{ij} \subseteq X \setminus \{X_i, X_j\}$ such that $X_i \perp X_j | S_{ij}$ holds.

  Delete the edge between X_i and X_j.

end for

for each pair of non-adjacent variables X_i and X_j with a common neighbor X_k **do**

  if $X_k \in S_{ij}$ **then**

    Continue.

  else

    Add arrowheads pointing as $X_k : (X_i \to X_k \leftarrow X_j)$.

  end if

end for

In the partially directed graph that results, orient as many of the undirected edges as possible subject to: (i) The orientation should not create a new v-structure, (ii) The orientation should not create a directed cycle.

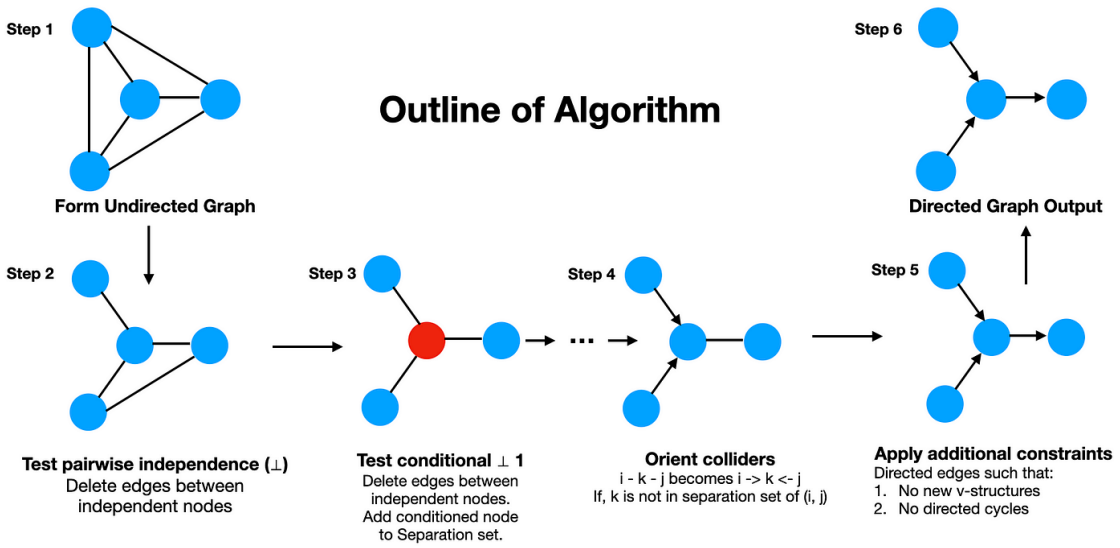


Figure 1: PC Algorithm for obtaining a compatible DAG

Image Source: *Causal Discovery Learning causation from data using Python*, https://towardsdatascience.com/causal-discovery-6858f9af6dcb

The PC algorithm's tests are designed to be consistent, meaning that as the number of observations increases and the significance level approaches zero, the likelihood of correctly identifying edges in the graph becomes nearly certain.

Proposition 6.1. *Under the assumption of faithfulness, the PC-algorithm can consistently identify the inferrable causal directions, i.e. for $T \to \infty$ the probability of recovering the inferrable causal structure of the data-generating causal model converges to one.*

The second approach to learning causal graphs is based on *Bayesian model averaging*. This method, detailed by Heckerman, combines *prior knowledge* with *observed data* to infer causal relationships among variables. Algorithms in this category differ mainly in two aspects: the *scoring criteria* used to evaluate a graph's fit to the data, and the *search strategy* employed to explore the space of possible graphs. Since the search space is *NP-hard*, exact solutions are often infeasible, and researchers rely on *heuristic algorithms* such as greedy search, greedy search with restarts, best-fit search, and Monte Carlo methods.

The third approach draws from *classical model selection* techniques. While it shares similarities with the Bayesian approach in terms of implementation and search strategies, it differs by *not incorporating any prior information*. Instead, it evaluates graphs based on *information-theoretic criteria* like the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). Like the Bayesian approach, it often uses greedy-based search strategies. A basic version of the *greedy search algorithm* is presented below.

Algorithm 2 Greedy Search Algorithm

Input: Observations of a set of variables X generated from a DAG model.

Output: A pattern (DAG) compatible with the data generating DAG.

Step 1: Start with a DAG A_0.

Step 2: Calculate the score of the DAG according to BIC/AIC/likelihood criterion.

Step 3: Generate the local neighbor DAGs by either adding, removing, or reversing an edge of the network A_0.

Step 4: Calculate the scores for the local neighbor DAGs. Choose the one with the highest score as A_n.

if the highest score is larger than that of A_0 **then**

Update A_0 with A_n and go to Step 2.

else

Stop and output A_0.

end if

It's important to recognize that a causal model functions as a statistical model. When the score used in the greedy search algorithm is based on a consistent model selection criterion, such as the Bayesian Information Criterion (BIC), the algorithm will reliably recover the inferable causal directions, assuming that the search space encompasses the true directed acyclic graph (DAG). BIC is a method that helps to identify the best-fitting model while penalizing for the number of parameters to avoid overfitting.

7 Time Series Causal Models

7.1 Background: DAGs and Structural Equation Models

When a multivariate random variable $X \in \mathbb{R}^n$ follows a joint Gaussian distribution, its causal structure can be modeled using a *linear causal model*. This is equivalent to a *linear recursive structural equation model* (SEM), where each component x_j is expressed as a linear function of its causal parents:

$$x_j = \sum_{k=1}^{j-1} a_{jk} x_k + u_j, \quad j = 1, 2, \ldots, n, \tag{1}$$

where a_{jk} quantifies the causal influence of variable x_k on x_j, and $u_j \sim \mathcal{N}(0, \sigma_j^2)$ is a noise term representing unobserved effects.

7.2 Matrix Representation of Recursive SEMs

This causal model can be represented compactly using a lower triangular matrix A, with ones on the diagonal:

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -a_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & \cdots & 1 \end{pmatrix}.$$

The SEM then satisfies the equation:

$$AX = U,$$

where $U \sim \mathcal{N}(0, \Lambda)$ with Λ diagonal. This implies:

$$A\Sigma A^\top = \Lambda,$$

which connects the SEM structure to the covariance matrix Σ of the jointly Gaussian variables.

7.3 Extension to Time Series

While SEMs are traditionally applied to independent data points, time series data are sequential and temporally dependent. Suppose we observe N variables over T time steps; we can treat the resulting NT variables as jointly distributed and embed them in a larger SEM that respects temporal order.

Assuming that each $X_{it} \sim \mathcal{N}(0, \sigma^2)$, the Gaussian assumption allows us to extend SEM theory to time series data.

We summarize this equivalence in the following proposition.

Proposition 7.1. *If a set of variables X are jointly normal $X \sim N(0; \Sigma)$, a linear causal model for X can be equivalently formulated as a linear recursive structural equation model (SEM) that is represented by a lower triangular coefficient matrix A with ones on the principal diagonal. Any nonzero element in this coefficient matrix, say α_{jk}, corresponds to a directed edge from variable k to variable j.*

$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \alpha_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{n1} & \alpha_{n2} & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -a_{21} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & -a_{n2} & \cdots & 1 \end{pmatrix}$$

where A is the triangular decomposition matrix of Σ with $A\Sigma A' = \Lambda$ and Λ is a diagonal matrix.

7.4 Recursive SEMs for Time Series

To model temporal causality, we impose a block lower-triangular structure over time:

$$\begin{pmatrix} A_{11} & 0 & \cdots & 0 \\ A_{21} & A_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{T1} & A_{T2} & \cdots & A_{TT} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{pmatrix}, \tag{2}$$

where:

- $X_t \in \mathbb{R}^N$ is the multivariate observation at time t,
- A_{tt} describes contemporaneous (within-time) effects,
- A_{ts} for $s < t$ captures lagged (across-time) causal effects,
- $\epsilon_t \sim \mathcal{N}(0, D)$ are uncorrelated noise terms.

This formulation ensures that past values can influence the present and future, but not vice versa, thereby preserving causality over time.

7.5 Structural Assumptions for Identifiability

Time series present unique challenges, including the issue that each time point yields only one sample. To enable learning of the causal structure, we adopt the following assumptions:

1. **Temporal Causality**: Causal mechanisms operate consistently over time.
2. **Time-Invariance**: The structure of the causal graph does not change across time steps.
3. **Finite Memory**: Each variable at time t can influence only up to p future steps, for some fixed lag p.

These assumptions reduce the model complexity and make estimation feasible.

7.6 Time Series Causal Model (TSCM)

Under the above assumptions, the full system can be expressed using a block Toeplitz structure for lag $p = 2$:

$$\begin{pmatrix} A_0 & 0 & \cdots & & \cdots & 0 \\ A_1 & A_0 & 0 & \cdots & & 0 \\ A_2 & A_1 & A_0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & A_2 & A_1 & A_0 & 0 \\ 0 & \cdots & 0 & A_2 & A_1 & A_0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{pmatrix} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{pmatrix}.$$

Here:

- A_0 captures contemporaneous effects,
- A_1, A_2 represent lag-1 and lag-2 dependencies,
- ϵ_t remains i.i.d. Gaussian noise across time.

This structured model is referred to as a **Time Series Causal Model (TSCM)**.

7.7 Illustration: Time Series Causal Graph

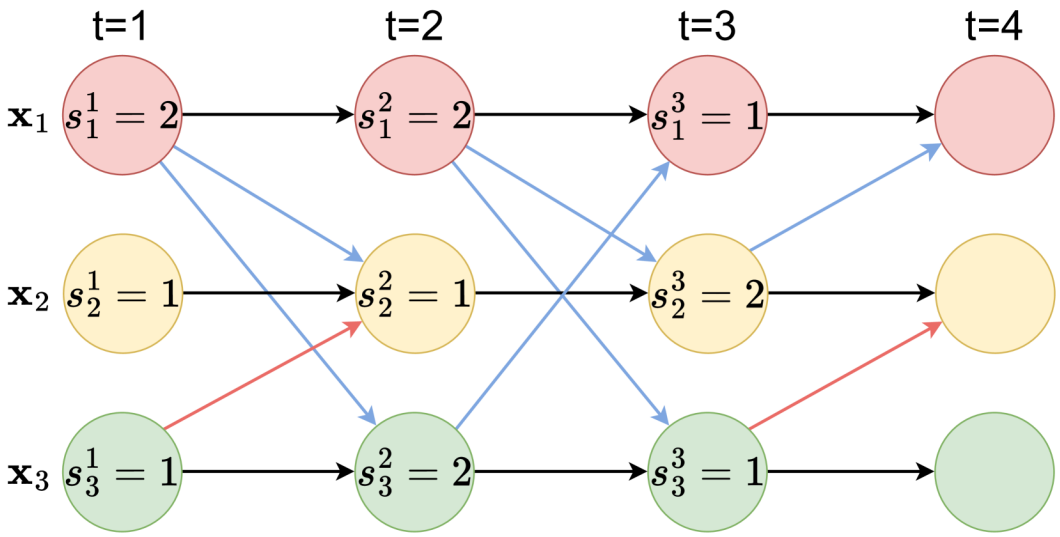


Figure 2: Example of a time series causal graph illustrating both contemporaneous and lagged causal connections.

Image Source: *Causal Discovery from Conditionally Stationary Time Series, arXiv:* https://arxiv.org/abs/2110.06257

The above figure shows directed edges across and within time steps, consistent with the assumptions of the TSCM framework.

7.8 Observational Equivalence in TSCMs

Proposition 7.2. *A partial Directed Acyclic Graph (DAG) has an observationally equivalent model if there are some arrows between elements of X_t that satisfy the following two conditions:*

- *The lagged parents of the connected elements of X_t are the same*
- *The change of the arrow directions will not lead to a new v-structure or a cycle in the partial DAG*

Corollary: If in a partial DAG all the elements of X_t have different lagged parents, the partial DAG does not have an observationally equivalent model.

8 Vector Autoregression (VAR) Model

8.1 Introduction

The **Vector Autoregression (VAR)** model is a fundamental statistical tool used to capture the linear interdependencies among multiple time series. Unlike univariate autoregressive models that examine the temporal dynamics of a single variable, the VAR model extends this idea to a multivariate context, allowing for the modeling of systems where multiple variables influence each other over time. This makes VAR particularly well-suited for analyzing macroeconomic systems, financial markets, and any domain where variables evolve jointly.

8.2 Model Structure

In a VAR model, all variables are treated as *endogenous*. That is, each variable in the system is explained by its own past values as well as the past values of all other variables in the model. Let $y_t = (y_{1,t}, y_{2,t}, \ldots, y_{k,t})^\top$ represent a k-dimensional vector of time series variables at time t. Then a VAR model of order p, denoted VAR(p), is specified as:

$$y_t = c + A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + e_t, \tag{3}$$

where:

- $c \in \mathbb{R}^k$ is a vector of intercept terms,
- $A_i \in \mathbb{R}^{k \times k}$ for $i = 1, \ldots, p$ are coefficient matrices capturing the linear influence of lagged variables,
- $e_t \in \mathbb{R}^k$ is a white noise error term representing unpredictable shocks at time t.

8.3 Properties of the Error Terms

The error terms e_t are assumed to satisfy the following conditions:

1. **Zero Mean**: $\mathbb{E}(e_t) = 0$. On average, the errors are centered around zero.
2. **Contemporaneous Correlation**: The covariance matrix $\mathbb{E}(e_t e_t^\top) = \Omega$ is positive semi-definite, allowing for correlation across the different components of e_t at a single time step.
3. **No Serial Correlation**: $\mathbb{E}(e_t e_{t-k}^\top) = 0$ for all $k \neq 0$, implying that the error terms are uncorrelated across time.

These assumptions are crucial for the statistical validity of inference and forecasting in VAR models.

8.4 Lag Order Selection and Stationarity

The order p in VAR(p) specifies the number of past time steps (lags) included in the model. Choosing the appropriate lag order is essential for accurately capturing the system's dynamics without overfitting. Information criteria such as AIC, BIC, or cross-validation are typically used for this purpose.

An important aspect of VAR modeling is the stationarity of the time series. A time series is said to be **stationary** if its statistical properties—such as mean, variance, and autocorrelation—do not change over time. In contrast, **non-stationary** series often exhibit trends, seasonality, or changing variance.

The concept of integration order helps formalize this:

- A series is $I(0)$ if it is stationary in levels.
- A series is $I(d)$ if it becomes stationary after differencing d times.

For example, an $I(1)$ series needs to be differenced once to achieve stationarity.

8.5 Implications for Model Specification

Stationarity is a critical assumption in time series analysis. If the variables in a VAR model are stationary ($I(0)$), the model can be estimated and interpreted directly. However, if the variables are non-stationary ($I(d)$, with $d > 0$), they require preprocessing through differencing or transformation. In particular, if the non-stationary variables are cointegrated—i.e., they share a long-term equilibrium relationship—then a **Vector Error Correction Model (VECM)** is more appropriate. A VECM augments the VAR framework by incorporating both short-run dynamics and long-run relationships among variables.

In practice, testing for stationarity (e.g., using the Augmented Dickey-Fuller or KPSS test) is a necessary first step before estimating a VAR model.

9 Relation of TSCM and VAR Model

Now we will discuss about the relation between TSCMs and the VAR models in time series econometrics via these two propositions.

Proposition 9.1. *A TSCM is a restricted structural VAR model identified by the inferred causal relations among* $\{X_t\}_{t=1}^{T}$*, and hence it corresponds to a restricted VAR model.*

Proposition 9.2. *An unconstrained VAR model corresponds to a full partial DAG such that the TSCM does not contain any inferable causal relations except the temporal causal orders.*

10 Granger Causality in Time Series Causal Models (TSCMs)

Granger causality is a widely used concept in time series analysis that assesses whether one variable can help forecast another. Originally proposed by Clive Granger in 1969, the central idea is that if the past values of one time series X contain information that improves the prediction of another time series Y, beyond what is possible with Y's own past values, then X is said to *Granger-cause* Y.

10.1 Conceptual Framework

The Granger causality framework is founded on two key principles:

1. The cause must precede the effect in time.
2. The cause must contain unique information about the effect's future.

In practical terms, this means that if incorporating past values of X improves forecasts of Y, then X is a Granger cause of Y.

10.2 Mathematical Formulation

Formally, Granger causality can be defined in terms of conditional probabilities. Let $\mathcal{I}(t)$ denote the information set available at time t, and let $\mathcal{I}_{-X}(t)$ be the same set excluding the history of X. Then X Granger-causes Y if:

$$\mathbb{P}[Y(t+1) \in A \mid \mathcal{I}(t)] \neq \mathbb{P}[Y(t+1) \in A \mid \mathcal{I}_{-X}(t)] \tag{4}$$

Here, A is an arbitrary non-empty set of values, and $\mathbb{P}$ denotes probability. If this inequality holds, then knowledge of X's past provides additional predictive power for $Y(t+1)$.

10.3 Testing Procedure

To empirically test Granger causality between two series x_t and y_t, we perform the following steps:

1. Estimate a univariate autoregressive model:

$$y_t = a_0 + a_1 y_{t-1} + \cdots + a_m y_{t-m} + \varepsilon_t$$

2. Estimate an augmented model that includes lagged values of x_t:

$$y_t = a_0 + a_1 y_{t-1} + \cdots + a_m y_{t-m} + b_p x_{t-p} + \cdots + b_q x_{t-q} + \varepsilon_t$$

3. Use an F-test (or similar test) to assess whether the coefficients on x's lags are jointly significant.

If the inclusion of lagged x_t terms improves the model's fit, we reject the null hypothesis that x_t does not Granger-cause y_t.

10.4 Multivariate Granger Causality

In a multivariate setting, Granger causality is tested within the framework of a Vector Autoregressive (VAR) model:

$$X(t) = \sum_{\tau=1}^{L} A_\tau X(t-\tau) + \varepsilon(t)$$

Here, $X(t)$ is a vector of time series variables, A_τ are coefficient matrices, and $\varepsilon(t)$ is a vector of white Gaussian noise. A variable X_i is said to Granger-cause X_j if at least one coefficient $A_\tau(j,i)$ is significantly different from zero for some lag τ.

10.5 Relation to Time Series Causal Models (TSCMs)

Granger causality plays an important role in understanding predictive relationships in TSCMs. While Granger causality focuses on statistical predictability, TSCMs aim to uncover causal structure by leveraging temporal and conditional dependencies.

The relationship between these two concepts is captured in the following proposition:

Proposition 10.1. *Let $X_{i,t}$ and $X_{j,t}$ be two time series variables in a TSCM. Then $X_{j,t}$ Granger-causes $X_{i,t}$, given the rest of the variables in the TSCM, if and only if there exists a directed path from some lagged variable $X_{j,t-s}$ to $X_{i,t}$ for $s > 0$ in the partial Directed Acyclic Graph (DAG) representing the TSCM.*

This shows that a Granger causal relationship corresponds to the presence of a directed path in the graphical representation of a TSCM.

10.6 Distinguishing Granger Causality from Structural Causality

It is important to emphasize that Granger causality and TSCMs address different types of questions:

- **Granger Causality** assesses whether the past values of one variable improve predictions of another, without requiring structural assumptions about the data-generating process.
- **TSCMs**, by contrast, aim to model the underlying causal mechanisms using a formal structure (e.g., a partial DAG), which may impose constraints that go beyond predictive association.

In summary, Granger causality is a valuable diagnostic tool for identifying potential causal relationships in time series data, while TSCMs offer a principled framework for modeling and interpreting such relationships in a structural and graphical way.

11 Learning TSCMs

In a Time Series Causal Model (TSCM), we only need to learn a partial DAG with $(p+1)N$ nodes, instead of the full DAG with TN nodes.

Lemma 11.1. Given the assumption of a causal model, an information set (joint distribution) containing a node and its parent variables is sufficient for the PC algorithm to connect the node to its parents and exclude non-descendants from connecting to it.

Proposition 11.1. *To learn the partial DAG with arrows into X_t, the information set including $X_t, X_{t-1}, \ldots, X_{t-p}$ is sufficient.*

12 Search Algorithms for TSCMs

Here are two algorithms discussed for discovering the causal relation between variables in a Time Series Model.

12.1 PC Algorithm for TSCMs

Algorithm 3 PC Algorithm for a Partial DAG in TSCM

Input: Observations of a set of time series variables X generated from a TSCM.

Output: A partial DAG compatible with the data-generating DAG.

Step 1: Choose a reasonable $\hat{p}$.

Step 2: Calculate the correlation matrix $\Sigma = \text{corr}(X_t, X_{t-1}, \ldots, X_{t-\hat{p}})$.

Step 3: Use Σ as input to obtain a DAG for $(X_t, X_{t-1}, \ldots, X_{t-\hat{p}})$.

Step 4: Delete all arrows and edges that do not connect at least one element of X_t.

Step 5: Orient all edges between X_{t-i} and X_t with arrowheads at X_t.

Step 6: Orient all edges between elements of X_t using the rules in the PC algorithm.

12.2 Greedy Search Algorithm for TSCMs

- Evaluating graph scores is an alternative to uncovering the data-generating DAG model.
- For a partial DAG, the score can be based on the likelihood of the SVAR model.
- Since unconstrained models have higher likelihoods, a proper score includes a penalty term.
- The BIC criterion for a partial DAG of X_t is defined as:

$$\text{BIC} = \sum_{t=1}^{T} \log L(A_0, A_1, \ldots, A_p; X_t | X_{t-1}, \ldots, X_{t-p}) - (|E| + |V|) \log(T)$$

- Where $|E|$ is the number of arrows heading at X_t and $|V|$ is the number of elements in X_t.
- $(|E| + |V|)$ represents the number of free varying parameters in the TSCM.
- The BIC criterion is a sum of the log-likelihood function and a penalty factor.
- As $T \to \infty$, the BIC criterion becomes consistent for model selection.

Now we can summarize this details on greedy search into the following proposition:

Proposition 12.1. *Under the assumption of TSCM, the BIC criterion is a consistent score, such that the probability of identifying the true model converges to 1 as $T \to \infty$, assuming the search space covers the true model.*

13 Aim of Work

The primary aim of this work is to explore how causal discovery algorithms can be effectively applied to time-series health data, such as ECG and EEG recordings. By analyzing these signals, we hope to uncover meaningful causal relationships that are often hidden within complex temporal patterns. A major focus is on identifying key variables and understanding the factors that may lead to significant medical events, such as tissue damage, arterial blockages, or abnormal physiological responses. In doing so, we also aim to develop classification strategies that can help organize and interpret these events more clearly, ultimately supporting more accurate and insightful medical diagnoses.

14 Data Availability

The datasets are taken mainly from three sources. The real life ECG Data is provided by Madras Medical Mission. The ECG data we trained our Causal CNN model is from the open source dataset at `https://www.kaggle.com/datasets/evilspirit05/ecg-analysis`. The other dataset used for the random forest model is at `https://doi.org/10.1038/s41591-023-02396-3`.

15 Our Work and Results

15.1 Random Forest Model with Feature Selection Using Causality

This section describes the methodology used to identify causally relevant variables from multivariate time series data and assess their predictive value in classifying the target outcome, `Outcome_Occlusion_MI`. The process integrates statistical causality tests, information-theoretic measures, and supervised learning model of Random Forest Classifier to build a robust pipeline for feature discovery and model evaluation.

To identify candidate variables that have predictive influence on the target, we employed both *linear* and *nonlinear* causality measures.

15.1.1 Linear Granger Causality

We first applied the Granger causality test, a standard method to assess whether one time series is useful for forecasting another. For each variable X_i in the dataset, we tested whether past values of X_i improve the prediction of the target variable $Y = \texttt{Outcome\_Occlusion\_MI}$ beyond what is possible using the past values of Y alone.

The test was implemented using the `grangercausalitytests` function from the `statsmodels` library. For each pair, we evaluated multiple lags (up to `max_lag`) and extracted the **minimum p-value** across all lags. Variables with smaller p-values indicate stronger evidence of linear Granger causality. The 25 variables with the lowest p-values were selected as top linear causal candidates.

15.1.2 Nonlinear Causality Measures

To capture nonlinear dependencies that may not be identified by linear models, we incorporated two information-theoretic approaches:

- **Transfer Entropy (TE):** TE quantifies the directional information transfer from one time series to another. It is particularly useful for detecting nonlinear and time-asymmetric relationships. For each variable, TE was computed from the candidate variable to the target with a one-step lag ($k = 1$). Higher TE values indicate a stronger nonlinear causal effect.
- **Mutual Information (MI):** As an efficient and scalable alternative, we computed the mutual information between each variable and the target using `mutual_info_regression` from `scikit-learn`. Though MI is not directional, it effectively captures general nonlinear associations. The top 25 variables with the highest MI scores were selected.

Together, these methods provided a ranked list of potential causal drivers using both linear and nonlinear lenses.

15.1.3 Predictive Modeling and Evaluation

To evaluate the utility of the selected causal features, we trained a supervised learning model to predict the binary target `Outcome_Occlusion_MI`.

15.1.4 Dataset Preparation

We constructed two feature sets:

- One containing the top 25 variables identified through **linear Granger causality**, and
- Another with the top 25 variables based on **nonlinear scores** (TE or MI).

These features were extracted from the dataset and paired with the target variable to create the full training set. The data was then split into training and test subsets using an **80/20 stratified split**, ensuring balanced class distributions in both sets.

15.1.5 Model Training: Random Forest Classifier

We trained a **Random Forest Classifier** with 100 decision trees (`n_estimators=100`) and a fixed random seed for reproducibility. This ensemble method was chosen for its robustness, ability to handle mixed-type features, and built-in feature importance estimation. Predictions were generated on the test set.

15.1.6 Performance Evaluation

We evaluated the trained Random Forest model using multiple metrics:

- **Accuracy:** 94% of predictions on the test set were correct.
- **Classification Report:** Shown below, detailing precision, recall, and F1-score for each class.

Table 1: Classification Report of the Random Forest Model

Class	Precision	Recall	F1-score	Support
0 (Negative)	0.95	0.99	0.97	2453
1 (Positive)	0.71	0.30	0.42	186
Accuracy		0.94		
Macro Avg	0.83	0.65	0.70	2639
Weighted Avg	0.93	0.94	0.93	2639

These metrics highlight high performance on the majority class (0), but a notable drop in recall for the minority class (1), reflecting the challenge of class imbalance.

In addition to predictive accuracy, we analyzed the **feature importances** from the trained Random Forest model. This helped identify which of the selected causal variables were most influential in making predictions.

Such analysis provides further validation of the causal features' relevance—not only from a statistical dependency standpoint, but also in their actual predictive contribution.

15.1.7 Model Training and Performance Evaluation

To predict the binary outcome `Outcome_Occlusion_MI`, we employed a Random Forest Classifier trained on features selected using Granger causality and mutual information methods. The dataset was split into training and testing subsets using an 80/20 ratio with stratified sampling to maintain class balance.

Initial training was performed using a default Random Forest with 100 estimators. The model demonstrated good performance on the test set, and predictions were evaluated using standard classification metrics.

To further improve performance, we conducted a comprehensive hyperparameter tuning process using grid search with 5-fold cross-validation. The grid included variations over the number of trees (`n_estimators`), tree depth (`max_depth`), minimum samples required to split and at leaves, and feature subset selection strategies:

- `n_estimators`: 50, 100, 200
- `max_depth`: None, 10, 20
- `min_samples_split`: 2, 5, 10
- `min_samples_leaf`: 1, 2, 4
- `max_features`: sqrt, log2

The best-performing model from this search achieved an AUROC (Area Under the Receiver Operating Characteristic Curve) of 0.864 on the test set, as shown in Figure 3. This indicates a strong ability to discriminate between the two classes.

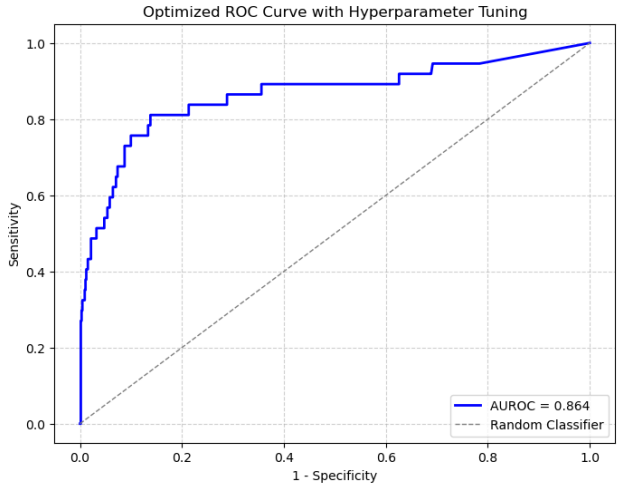


Figure 3: Optimized ROC Curve of Random Forest Classifier (AUROC = 0.864)

15.1.8 Comparison with Benchmark Models

To contextualize these results, we compared our classifier to the performance of ECG-SMART, a deep learning-based ECG interpretation model reported in the literature. As illustrated in Figure 4, ECG-SMART achieved an AUROC of 0.91 (95% CI: 0.87–0.96) on its internal test set, outperforming both traditional commercial ECG systems and clinical experts. This reuslt is taken from [8].

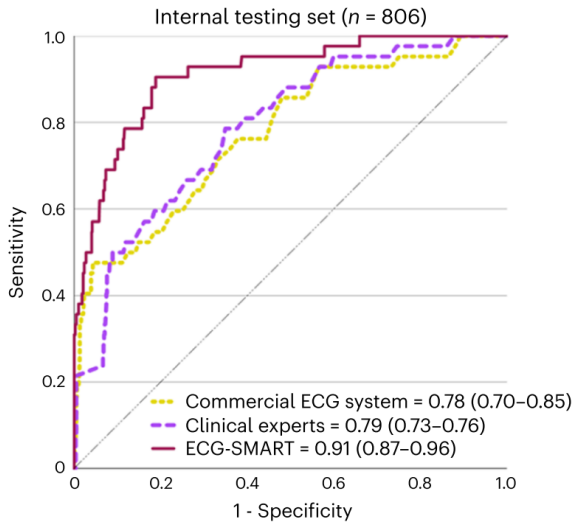


Figure 4: ROC Performance Comparison with ECG-SMART, Clinical Experts, and Commercial ECG Systems

Although the ECG-SMART model slightly outperforms our tuned Random Forest classifier, the gap is relatively narrow. This suggests that with informed feature selection and proper tuning, interpretable machine learning models can approach state-of-the-art performance in clinical prediction tasks.

15.2 Causal DAG with the BIC Criterion

To understand the causal relationships between variables of the previous Random Forest Model, we create Directed Acyclic Graphs (DAGs) based on two sets of features: one using only the selected features from our analysis, and another using all available features. In both cases, we use the Bayesian Information Criterion (BIC) to guide the model selection.

The BIC helps us find a balance between how well the model fits the data and how complex it is. This means it favors simpler models that still explain the data well, avoiding unnecessary complexity.

Figure 5 shows the DAG when all features are included, providing a detailed picture but also including more noise.

On the other hand, Figure 6 shows the causal graph we get when using just the selected features. This graph highlights the key causal links that matter most.

15.3 Convolutional Neural Network (CNN) Model for ECG Image Classification

Here we implemented our final CNN model on ECG data.

In this study, a convolutional neural network (CNN) was designed and trained to classify ECG images into four distinct categories. The dataset consisted of labeled ECG images organized into training and testing directories.

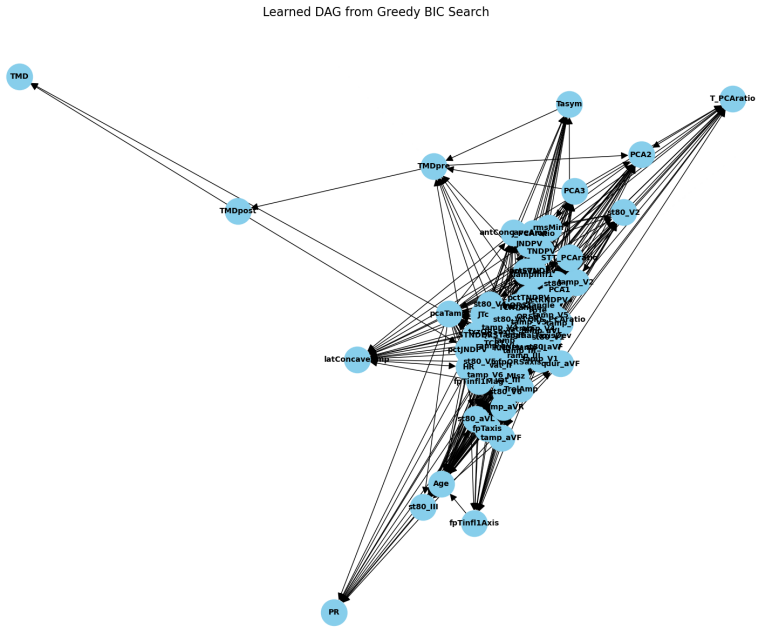


Figure 5: Causal DAG using BIC criterion based on all features.

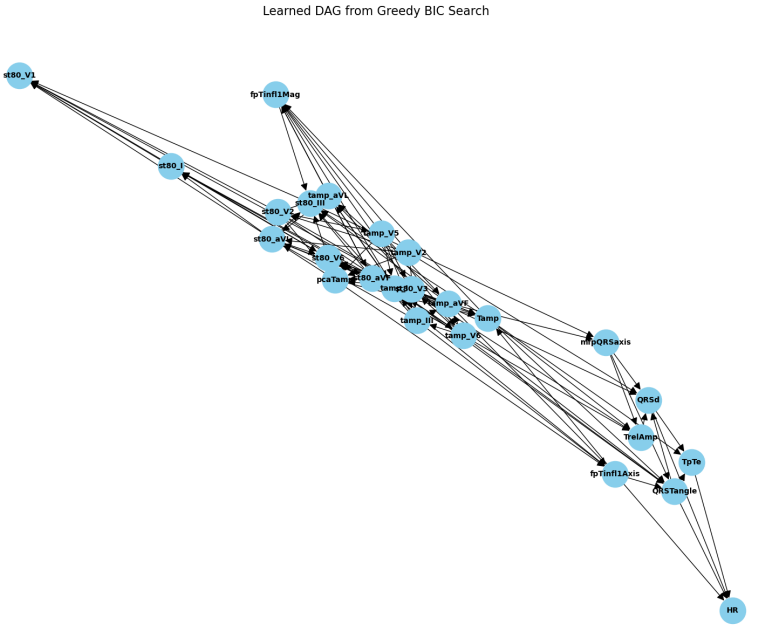


Figure 6: Causal DAG using BIC criterion based on the selected features.

15.3.1 Data Preprocessing and Augmentation

To enhance model generalization, the training images were preprocessed and augmented using the `ImageDataGenerator` utility from TensorFlow Keras. Specifically, the pixel values were rescaled to the $[0, 1]$ range, and data augmentation techniques including random shear transformations, zooming, and horizontal flipping

were applied. The test images were only rescaled without augmentation to maintain evaluation consistency. Both training and test images were resized to 150×150 pixels and processed in batches of 32.

15.3.2 Model Architecture

The CNN model comprises three convolutional blocks followed by fully connected layers:

- **Convolutional Blocks:** Each block consists of a convolutional layer with ReLU activation and "same" padding, followed by batch normalization, max pooling with a 2×2 window, and dropout regularization. The number of filters in these blocks increases progressively from 32 to 64 and then 128 to capture hierarchical image features effectively.
- **Fully Connected Layers:** The convolutional outputs are flattened and fed into a dense layer with 256 neurons and ReLU activation. A dropout layer with a rate of 0.5 is applied to reduce overfitting.
- **Output Layer:** The final dense layer contains four neurons with softmax activation to output class probabilities corresponding to the four ECG categories.

Figure 7 shows the basic structure of an ordinary CNN.

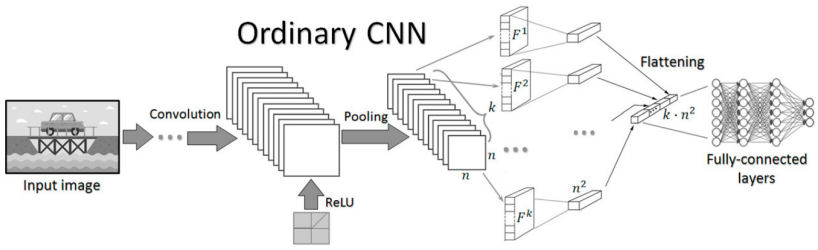


Figure 7

Image is taken from 6

15.3.3 Model Training and Optimization

The model was compiled using the Adam optimizer with a learning rate of 0.001 and optimized using categorical cross-entropy loss appropriate for multi-class classification. Accuracy was monitored as the primary metric.

To prevent overfitting and improve training efficiency, early stopping was employed to halt training if the validation loss did not improve for five consecutive epochs, and the best model weights were restored. Additionally, model checkpoints were saved during training to retain the best-performing model based on validation loss.

The model was trained for a maximum of 20 epochs, with batch-wise steps determined by the size of the training and test sets.

15.3.4 Performance Evaluation

The CNN model achieved a test accuracy of approximately 43.5%, indicating moderate performance in classifying the ECG images into four categories. While the training accuracy gradually improved over epochs, the validation accuracy showed fluctuations, suggesting that the model may benefit from further hyperparameter tuning or enhanced data augmentation to improve generalization.

15.4 Causal Aware CNN Model (CA-CNN) for ECG Image Classification

Here we introduced new methodology to upgrade the previous CNN model discussed which is explained in [6].

15.4.1 Modifications Introduced in the Causal CNN Model

The Causal Aware CNN (CA-CNN) model was developed to allow for more complex operations beyond the sequential stacking of layers. A novel *causality map* layer was incorporated, which computes a channel-wise causality matrix from the convolutional feature maps by reshaping and multiplying them to capture inter-channel dependencies. This causality map is normalized and then flattened.

The flattened causality map is concatenated with the flattened convolutional feature maps, effectively combining spatial features with causal relationships. This combined representation is then passed through fully connected layers for classification.

Other architectural components such as convolutional blocks, batch normalization, max pooling, dropout, and training configurations remain consistent with the previous CNN model. These modifications aim to enhance the model's ability to learn and utilize causal relationships within ECG image features, potentially improving classification performance.

15.4.2 Model Architecture

The Causal Aware CNN model comprises of the same ordinary CNN architecture with an extra Causal Feature Map layer that is given as input to the fully connected layer. Figure 8 shows the basic structure of an oridnary CNN.

15.4.3 Model Training and Optimization

The model training and optimization is similar to ordinary CNN.

15.4.4 Performance Evaluation

We checked the performance using two methods explained below:

1. Input to the Fully Connected Layer is both the Causal Feature Map and the data samples. Then the CA CNN model achieved a test accuracy of approximately 77.48%, indicating way better performance than the ordinary CNN.

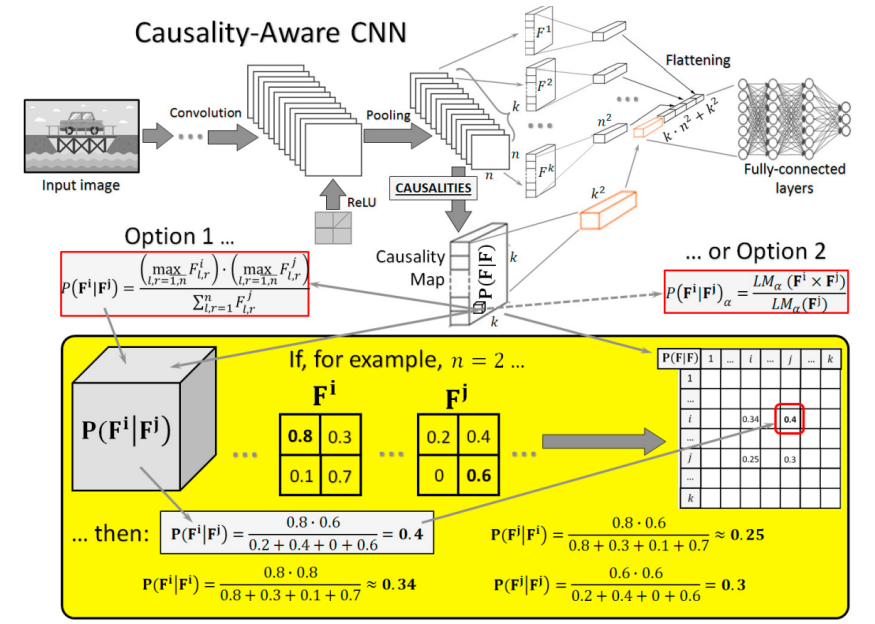


Figure 8

Image is taken from 6

2. Input to the Fully Connected Layer is only the Causal Feature Map and not the data samples. Then the CA CNN model achieved a test accuracy of approximately 54.31%, still indicating better performance than the ordinary CNN.

Further improvements are still possible as not much work is done in the architecture of the CNN mode and not many ways are checked for the causal map generation.

16 Conclusion

This study explored the potential of integrating causal reasoning into both traditional and deep learning models for ECG image classification. We first employed a Random Forest classifier trained on features selected through causal inference, allowing the model to focus on inputs with genuine causal relevance to the target class. This approach proved effective, with the classifier achieving strong performance despite its simplicity. Although it was slightly outperformed by the ECG-SMART model, the margin was narrow—highlighting that causally-informed, interpretable models can come remarkably close to the carefully designed and tuned model.

To further leverage the power of causality, we developed a Causal Convolutional Neural Network (CA-CNN) that augments a standard CNN with a causality map layer. This component captures inter-channel dependencies within feature maps, enriching the model's ability to learn structured, causally meaningful representations. Over the training epochs, the CA-CNN showed steady improvements in classification accuracy, ultimately surpassing baseline models.

These results underscore the value of embedding causal structure directly into deep learning architectures, especially for complex biomedical signals such as ECGs.

Overall, this work demonstrates that combining causal inference with machine learning—whether through feature selection or architectural design—can lead to models that are both more interpretable and more effective. Further directions may include exploring richer causal mechanisms, incorporating temporal dependencies, and applying these techniques to other domains of medical imaging for classification and future prediction as well.

17 Acknowledgements

I am deeply grateful to my thesis advisor, Prof. K. V. Subrahmanyam, for his steady guidance, thoughtful feedback, and constant encouragement throughout the course of this thesis. I would also like to sincerely thank Dr. Mullasari Ajit Sankardas, whose insights into ECG interpretation—particularly in the context of Occluded Myocardial Infarction—played a crucial role in shaping the direction of this project. A special thanks to Ms. Suchitra Karunakaran for her kind support in helping me preprocess and structure the ECG data, which was essential for building and training the model. This project would not have been possible without their support and expertise.

References

[1] Chen, Pu (2010). “A time series causal model." MPRA Paper 24841, University Library of Munich, Germany. Available at: https://ideas.repec.org/p/pra/mprapa/24841.html.

[2] Pearl, J. (2000). *Causality*. Cambridge University Press, 1st edition.

[3] Pearl, J., & Verma, T. (1991). A theory of inferred causation. In J.A. Allen, R. Fikes, and E. Sandewall (Eds.), *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference*, San Mateo, CA: Morgan Kaufmann, pp. 441–452.

[4] Del Tatto, V., Fortunato, G., Bueti, D., & Laio, A. (2024). Robust inference of causality in high-dimensional dynamical processes from the Information Imbalance of distance ranks. *Proc. Natl. Acad. Sci. U.S.A.*, **121**(19), e2317256121. Available at: https://doi.org/10.1073/pnas.2317256121.

[5] Ni, H., et al. (2024). Time Series Modeling for Heart Rate Prediction: From ARIMA to Transformers, *2024 6th International Conference on Electronic Engineering and Informatics (EEI)*, Chongqing, China, pp. 584-589. doi: 10.1109/EEI63073.2024.10695966.

[6] Vagan Terziyan, Oleksandra Vitko. (2023). Causality-Aware Convolutional Neural Networks for Advanced Image Classification and Generation, *Procedia Computer Science*, Volume 217, pp. 495-506, ISSN 1877-0509, doi: https://doi.org/10.1016/j.procs.2022.12.245.

[7] Bernhard Schölkopf and Julius von Kügelgen (2022), Bernhard Schölkopf and Julius von Kügelgen, arXiv: https://arxiv.org/abs/2204.00607

[8] Al-Zaiti, S.S., Martin-Gill, C., Zègre-Hemsey, J.K. et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. Nat Med 29, 1804–1813 (2023). doi: https://doi.org/10.1038/s41591-023-02396-3

[9] Spirtes, P., C. Glymour, and R. Scheines (1993). Causation, Prediction, and Search. New York: Springer Verlag.