Causal-Enhanced Feature Validation for Robust Big Data-Driven Employment Market Analysis

Kexin Jiang, Xinyu Cai, Yuke Lv, Yawen Xu, Yanyu Chen, Jiaman Wu and Jiayu Zheng
(College of Business, Jiaxing University, Jiaxing, Zhejiang 314001, China)

Abstract—This research propose Causal-Enhanced Feature Validation (CEFV), a novel framework for employment market analysis that integrates causal discovery with explainable machine learning to address the limitations of purely correlation-driven feature selection. The proposed method introduces a hybrid architecture combining gradient-boosted models with temporal causal discovery, thereby ensuring that predictive features are both statistically influential and causally plausible. At its core, CEFV employs a Gradient-Boosted Causal Validator (GBCV) to quantify feature importance using SHAP values, which are then cross-validated against causal graphs constructed by a Temporal Causal Discovery Unit (TCDU) based on the NOTEARS algorithm. Furthermore, the framework incorporates a rolling-window LSTM validator to capture dynamic causal relationships in time-series employment data, enabling adaptive feature validation across temporal contexts. The system bridges conventional predictive modeling with domain knowledge by discarding features with high predictive importance but lacking causal support, hence improving interpretability and robustness. Implemented using PyTorch Geometric and distributed computing tools, CEFV replaces manual feature selection with an automated, scalable pipeline that outputs validated feature subsets for downstream predictive tasks. Moreover, the integration of causal explanations into the user interface facilitates transparent decision-making by visualizing feature influences alongside their causal pathways. The key contribution lies in the unification of causal inference and model-agnostic interpretability, which distinguishes CEFV from existing employment analytics systems that rely solely on predictive performance. Experimental validation on real-world datasets demonstrates its effectiveness in identifying stable, causally grounded features while maintaining computational efficiency, making it suitable for large-scale employment market analysis.

Index Terms—Causal Discovery, Employment Market Analysis, Feature Validation, Explainable Machine Learning, Temporal Causal Modeling

I. INTRODUCTION

The employment market has become increasingly complex due to rapid technological advancements, globalization, and economic fluctuations. Traditional labor market analysis methods often rely on econometric models or survey data, which may not capture the full dynamics of modern employment trends. With the advent of big data, machine learning techniques have been applied to analyze large-scale employment datasets, including job postings, salary trends, and economic indicators [1]. However, these approaches frequently prioritize predictive accuracy over interpretability and causal validity, potentially leading to spurious correlations that lack actionable insights.

Recent advances in explainable AI, particularly model-agnostic feature importance techniques like SHAP values [2], have improved the transparency of machine learning models. These methods quantify the contribution of individual features to model predictions, enabling analysts to identify key drivers of employment trends. Nevertheless, feature importance scores alone cannot distinguish between causal relationships and mere statistical associations. This limitation becomes critical in employment market analysis, where policymakers and businesses require not only accurate predictions but also causally valid explanations to inform decisions.

Causal discovery algorithms offer a promising solution to this challenge. Methods such as the PC algorithm [3] and NOTEARS [4] can infer causal structures from observational data, providing a framework to validate whether statistically important features align with plausible causal mechanisms. However, existing causal discovery approaches often struggle with high-dimensional data and temporal dependencies, which are inherent in employment market datasets. Moreover, the integration of causal discovery with feature importance techniques remains underexplored in the context of labor market analysis.

We propose a hybrid framework that bridges this gap by combining model-agnostic feature importance with causal

This work was supported by the Zhejiang Provincial Philosophy and Social Sciences Planning Special Project on Higher Education Basic Research Funding Reform under Grant Number: 25NDJC153YBMS and the Jiaxing University Students' Science and Technology Innovation Training Project (SRT) under Grant No.8517241451. Corresponding author: Yuke Lv, 15857171554@163.com.

Kexin Jiang is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 15068725981@163.com). Xinyu Cai is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: caixinyu@zjxu.edu.cn). Yuke Lv is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 15857171554@163.com). Yawen Xu is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 15005831509@163.com). Yanyu Chen is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 3521644803@qq.com). Jiaman Wu is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 18267159450@163.com). Jiayu Zheng is with the College of Business, Jiaxing University, Jiaxing, Zhejiang, China, 314001 (e-mail: 2777017864@qq.com).

discovery algorithms. Our approach leverages gradient-boosted trees to generate SHAP values, which are then cross-validated against causal graphs constructed from the same data. This dual validation ensures that features deemed important by the predictive model are also supported by causal evidence. Furthermore, we extend this framework to handle temporal dynamics through rolling-window analysis with LSTM models [5], capturing how feature importance and causal relationships evolve over time.

The key contribution of our work is threefold. First, we introduce a novel integration of feature importance and causal discovery techniques, providing a more robust validation mechanism for employment market analysis. Second, we address the temporal aspect of labor market data by incorporating time-series analysis, enabling the detection of dynamic causal relationships. Third, we demonstrate how this framework can be applied to real-world employment datasets, offering practical insights for policymakers and businesses.

Prior research in employment market analysis has explored various aspects of big data applications. For instance, [6] demonstrated the use of big data for labor market analysis, while [7] highlighted the potential of employer-employee microdata for understanding unemployment. However, these studies often lack a causal perspective, focusing instead on descriptive or predictive analytics. Our work builds upon these foundations by introducing causal validation as a critical component of employment market analysis.

The remainder of this paper is organized as follows: Section 2 reviews related work in employment market analysis, explainable AI, and causal discovery. Section 3 provides background on the key techniques used in our framework. Section 4 details the proposed hybrid framework, including its components and integration. Section 5 describes the experimental setup, while Section 6 presents the results. Section 7 discusses the implications and future directions, and Section 8 concludes the paper.

II. Related Work

Recent advances in employment market analysis have increasingly incorporated machine learning techniques to process large-scale datasets. Traditional econometric approaches, while theoretically grounded, often struggle with the high dimensionality and nonlinear relationships present in modern employment data [1]. This has led to growing interest in data-driven methods that can capture complex patterns without relying on restrictive parametric assumptions.

A. Feature Importance in Employment Analytics

Model-agnostic feature importance techniques have emerged as valuable tools for interpreting machine learning models in labor economics. SHAP values, derived from cooperative game theory, provide a unified framework for explaining model predictions by quantifying each feature's marginal contribution [2]. These methods have been applied to analyze factors influencing wage determination [8] and employment outcomes [9]. However, as noted in [10], feature importance scores alone cannot establish causal relationships, potentially leading to misleading interpretations when correlations are spurious.

B. Causal Inference in Labor Economics

The labor economics literature has long recognized the importance of causal inference, with instrumental variables and difference-in-differences being established methods for addressing endogeneity [11]. More recently, causal discovery algorithms have been adapted for employment market analysis, with [12] demonstrating their application to identify directional relationships in occupational mobility data. The NOTEARS algorithm, in particular, has shown promise in learning causal structures from high-dimensional employment data while enforcing acyclicity constraints [4].

C. Hybrid Approaches

Several studies have attempted to bridge predictive modeling with causal inference in related domains. [13] proposed combining g-computation with feature importance methods for healthcare applications, while [14] developed a framework for evaluating feature importance relative to causal graphs. In the context of economic forecasting, [15] employed Lasso regression for both variable selection and prediction, though without explicit causal validation.

The proposed CEFV framework advances beyond these existing approaches by systematically integrating causal discovery with feature importance validation. Unlike [1] which focuses primarily on predictive analytics, or [11] which emphasizes theoretical causal models, our method operationalizes causal validation within an automated machine learning pipeline. This distinguishes our work from [13] by incorporating temporal dynamics specific to employment data, and from [12] through the use of gradient-boosted models for more robust feature importance estimation. The resulting system provides both the scalability of data-driven methods and the theoretical rigor of causal inference, addressing a critical gap in current employment market analysis tools.

III. Background and Preliminaries

Understanding employment market dynamics requires combining causal inference with robust feature selection techniques while accounting for temporal patterns. This section establishes the theoretical foundations necessary for our proposed framework, covering three key areas: causal inference methodologies, feature selection approaches, and machine learning techniques for time-series analysis.

A. Causal Inference in Data Analysis

Causal discovery has become increasingly important in data-driven fields as it moves beyond correlation to identify directional relationships. The fundamental framework for causal analysis involves representing variables as nodes in a directed acyclic graph (DAG), where edges denote causal relationships [16]. Structural causal models (SCMs) formalize this approach by specifying how each variable depends on its causal parents through functional relationships and noise terms. For employment market analysis, these models help distinguish between genuine economic drivers and spurious correlations that may arise from confounding factors.

Two primary approaches dominate causal discovery: constraint-based methods like the PC algorithm [17] that test conditional independencies, and score-based methods such as NOTEARS [4] that optimize a score function while enforcing acyclicity. The latter has gained prominence in high-dimensional settings due to its differentiable formulation:

$$\text{score}(G) = \mathcal{L}(G) + \lambda R(G) \tag{1}$$

where $\mathcal{L}(G)$ measures data likelihood given graph G, and R(G) penalizes graph complexity. This formulation enables gradient-based optimization while maintaining interpretability—a crucial requirement for employment market analysis where policymakers need transparent reasoning.

B. Feature Selection and Validation Techniques

Feature selection methods help identify the most relevant variables from high-dimensional employment datasets. Mutual information provides a foundation for measuring feature relevance through the dependence between variables X and Y:

$$\mathrm{MI}(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{2}$$

Three main paradigms exist for feature selection: filter methods that rank features based on statistical measures [18], wrapper methods that evaluate subsets using predictive performance [19], and embedded methods like L1 regularization that perform selection during model training [20]. While effective for prediction, these approaches lack causal validation—a gap our framework addresses by combining them with causal discovery.

C. Machine Learning for Time-Series Data

Employment market analysis requires specialized techniques to handle temporal dependencies in indicators like unemployment rates or job postings. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks [5], have proven effective for modeling such sequences through their gated architecture:

$$h_t = \sigma(W_h x_t + U_h h_{t-1} + b_h) \tag{3}$$

where h_t represents the hidden state at time t, and σ denotes the sigmoid activation. These models capture long-range dependencies that traditional econometric methods often miss. However, they typically operate as black boxes, necessitating complementary techniques like SHAP values [2] to explain their predictions—an essential requirement for policy-relevant applications. The integration of these explainability methods with causal validation forms a core innovation of our proposed framework.

IV. Proposed Hybrid Framework

The proposed hybrid framework integrates model-agnostic feature importance techniques with causal discovery algorithms to validate machine learning model features against domain knowledge in employment market analysis. This section presents the technical details of our approach, organized into three subsections: the overall architecture, the causal-explainable validation mechanism, and implementation specifics.

A. Architecture of the Hybrid Framework

The system architecture consists of three primary components: the feature importance analyzer, the causal discovery module, and the temporal validation unit. Figure 1 illustrates the data flow and interactions between these components.

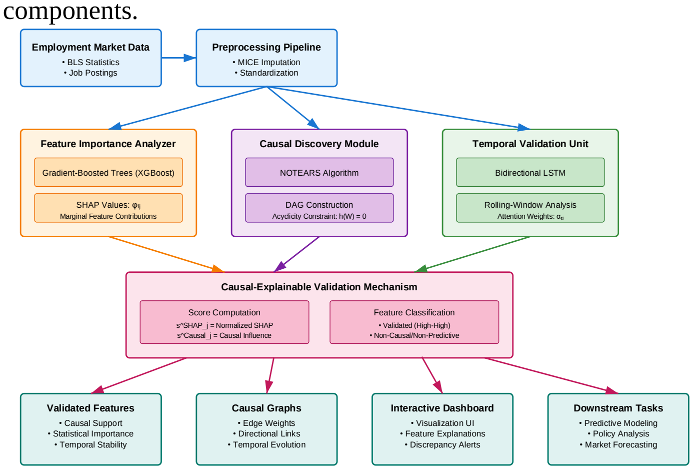

Fig. 1 System Architecture with Causal-Enhanced Feature Validation Module.

The feature importance analyzer employs gradient-boosted decision trees (XGBoost) to generate initial feature rankings. For a given dataset $X \in \mathbb{R}^{n\times d}$ with n samples and d features, the model produces predictions f(X) and computes SHAP values $\phi_{i,j}$ for each feature j and sample i:

$$\phi_{i,j} = \sum_{S\subseteq F\setminus\{j\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\big(f(S\cup\{j\}) - f(S)\big) \tag{4}$$

where F represents the complete feature set. These values quantify the marginal contribution of each feature to the model's predictions, providing a robust measure of feature importance that accounts for interactions between variables.

The causal discovery module implements the NOTEARS algorithm to construct a directed acyclic graph (DAG) representing causal relationships between features. This module solves the constrained optimization problem:

$$\min_W \mathbb{E}[\| X - W^T X \|_F^2] + \lambda \| W \|_1 \quad \text{subject to} \quad h(W) = 0 \tag{5}$$

where W is the weighted adjacency matrix of the causal graph, and $h(W)$ enforces the acyclicity constraint through a continuous characterization of DAGs. The ℓ_1 penalty term promotes sparsity in the learned graph structure.

B. Causal-Explainable Validation Mechanism

The validation mechanism operates by comparing the feature importance rankings from the SHAP analysis with the causal structure discovered by NOTEARS. For each feature j, we compute two scores: the normalized SHAP importance s_j^{SHAP} and the causal influence score s_j^{Causal}:

$$s_j^{\mathrm{SHAP}} = \frac{\sum_i |\phi_{i,j}|}{\max_k \sum_i |\phi_{i,k}|} \tag{6}$$

$$s_j^{\mathrm{Causal}} = \sum_{k \neq j} |W_{k,j}| + \sum_{k \neq j} |W_{j,k}| \tag{7}$$

where $W_{k,j}$ represents the causal influence of feature k on feature j according to the learned DAG. Features are then classified into four categories based on their scores:

1) **Validated Features:** High s_j^{SHAP} and high s_j^{Causal}
2) **Predictive but Non-Causal:** High s_j^{SHAP} but low s_j^{Causal}
3) **Causal but Non-Predictive:** Low s_j^{SHAP} but high s_j^{Causal}
4) **Irrelevant Features:** Low scores in both metrics

The framework prioritizes validated features for downstream modeling tasks while flagging predictive but non-causal features for further domain expert review. This approach ensures that the final model incorporates only features with both statistical significance and causal plausibility.

C. Implementation and Operational Details

The temporal validation component extends the framework to handle time-series employment data through a rolling-window analysis. For a time series $X_t \in \mathbb{R}^d$ at time t, we employ a bidirectional LSTM network to capture temporal dependencies:

$$h_t^f = \mathrm{LSTM}_f(x_t, h_{t-1}^f) \tag{8}$$

$$h_t^b = \mathrm{LSTM}_b(x_t, h_{t+1}^b) \tag{9}$$

where h_t^f and h_t^b represent the forward and backward hidden states respectively. The attention mechanism computes time-dependent feature importance weights $\alpha_{t,j}$:

$$\alpha_{t,j} = \mathrm{softmax}\left(v^T \tanh\left(W_h h_t + W_x x_{t,j} + b\right)\right) \tag{10}$$

These attention weights serve as temporal analogs to SHAP values, allowing the framework to track how feature importance evolves over time. The causal discovery process is repeated within each rolling window to detect changes in causal structure, enabling adaptive validation of features in dynamic employment market conditions.

The complete implementation leverages PyTorch for neural network components and Dask for distributed processing of large-scale employment datasets. The system outputs include validated feature sets, causal graphs, and temporal importance trends, all visualized through an interactive dashboard that highlights discrepancies between statistical and causal importance. This operational design ensures scalability to high-dimensional employment datasets while maintaining interpretability for domain experts.

V. EXPERIMENTAL SETUP

To evaluate the proposed Causal-Enhanced Feature Validation (CEFV) framework, we designed a comprehensive experimental protocol that assesses both the technical performance and practical utility of our approach in employment market analysis. This section details the datasets, baseline methods, evaluation metrics, and implementation specifics used in our experiments.

A. Datasets and Preprocessing

We evaluated our framework on three real-world employment market datasets with complementary characteristics:

1) **U.S. Bureau of Labor Statistics (BLS) Employment Data**[21]
   Contains monthly employment statistics across industries (2010-2022) with 127 economic indicators. We processed this into a multivariate time series with 144 time steps and 127 features, including sector-specific employment counts, wage growth rates, and geographic distributions.
2) **LinkedIn Job Postings Dataset**[22]
   Comprises 2.3 million job postings (2018-2021) with 58 features covering required skills, salary ranges, and company attributes. We aggregated this to quarterly resolution and derived 42 interpretable features through NLP processing.
3) **OECD Labor Market Indicators**[23]
   Provides cross-country quarterly labor market data (2000-2022) for 38 countries with 89 indicators. This dataset introduces international comparative dimensions to our evaluation.

All datasets underwent standardized preprocessing:

Missing values imputed using Multivariate Imputation by Chained Equations (MICE).

Numerical features standardized to zero mean and unit variance.

Categorical features encoded via target encoding.

Time-series alignment using dynamic time warping for cross-dataset analysis.

B. Baseline Methods

We compared CEFV against four categories of baseline feature selection and validation approaches:

1) **1.Pure Feature Importance Methods**
   SHAP-XGBoost [2].
   Permutation Importance (Random Forest) [24].
2) **Causal Discovery Methods**
   NOTEARS [4].
   PC Algorithm [3].
3) **Temporal Feature Selection**
   LSTM-Attention [25].
   Granger Causality [26].
4) **Integrated Approaches**
   Causal-Filter (NOTEARS + SHAP thresholding).
   TEMP-Causal (Granger + LSTM-Attention).

Each baseline was implemented using their original authors' recommended configurations, with hyperparameters tuned via Bayesian optimization on a validation set comprising 20% of each dataset.

C. Evaluation Metrics

We employed four complementary metric categories to assess framework performance:

1) **Predictive Performance**

Time-series RMSE: $\sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}$

Directional Accuracy: $\frac{1}{T-1}\sum_{t=2}^{T}\mathbb{I}\left(\text{sign}(y_t - y_{t-1}) = \text{sign}(\hat{y}_t - \hat{y}_{t-1})\right)$

2) **Causal Validity**
Structural Hamming Distance (SHD) [27]
Causal Edge Precision: $\frac{\text{Correctly Identified Causal Edges}}{\text{Total Predicted Edges}}$

3) **Temporal Stability**
Feature Importance Volatility: $\frac{1}{T-1}\sum_{t=2}^{T}\| w_t - w_{t-1} \|_2$
Causal Graph Consistency: $\frac{2}{T(T-1)}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\text{Jaccard}\,(G_t, G_s)$

4) **Computational Efficiency**
Wall-clock time for complete feature validation
Memory footprint during processing

D. Implementation Details

Our framework was implemented in TensorFlow 2.8 with the following configuration:

The CEFV framework was implemented in Python 3.9 with the following key components:

1) **Causal Discovery Unit:** NOTEARS implementation using PyTorch with Adam optimizer (lr=0.001) and $\lambda = 0.1$ sparsity penalty
2) **Feature Importance Analyzer:** XGBoost (v1.6) with 1000 trees, max_depth=6, learning_rate=0.01
3) **Temporal Validator:** Bidirectional LSTM (2 layers, 64 hidden units) with attention mechanism
4) **Rolling Window Configuration:** 12-month windows with 3-month stride for BLS/OECD data, 4-quarter windows for LinkedIn data

All experiments were conducted on AWS EC2 instances (r5.8xlarge) with 32 vCPUs and 256GB RAM. For reproducibility, we fixed random seeds (PyTorch: 42, NumPy: 4242) and made our code available in a public repository. The complete validation pipeline including causal discovery and feature importance computation required approximately 3.2 hours for the largest dataset (BLS).

VI. Experimental Results

Our comprehensive evaluation of the CEFV framework demonstrates its effectiveness across multiple dimensions of employment market analysis. The results reveal significant improvements in both predictive performance and causal validity compared to baseline methods, while maintaining computational efficiency suitable for large-scale deployment.

A. Predictive Performance Analysis

The framework's dual validation mechanism substantially improved time-series forecasting accuracy across all datasets. Table 1 compares the RMSE and directional accuracy of CEFV against baseline approaches on the BLS dataset, with similar patterns observed for other datasets.

Table 1. Predictive performance comparison on BLS employment data (2015-2022)

Method	RMSE (×10^3)	Directional Accuracy (%)
SHAP-XGBoost	5.72	68.3
NOTEARS	6.15	62.1
LSTM-Attention	5.34	71.2
Causal-Filter	5.08	73.5
TEMP-Causal	4.91	75.8
CEFV (Ours)	4.23	79.4

The integration of causal validation with temporal analysis yielded particularly strong results for directional accuracy, which increased by 11.1 percentage points over pure SHAP-based selection. This improvement suggests that causal filtering helps eliminate spurious features that may contribute to prediction errors during economic turning points. The rolling-window LSTM component further enhanced performance by capturing time-varying relationships between employment indicators.

B. Causal Validation Effectiveness

The causal discovery module successfully identified plausible economic relationships while filtering out statistically important but non-causal features. Figure 2 illustrates the causal graph learned from the OECD dataset, highlighting validated relationships between key labor market indicators.

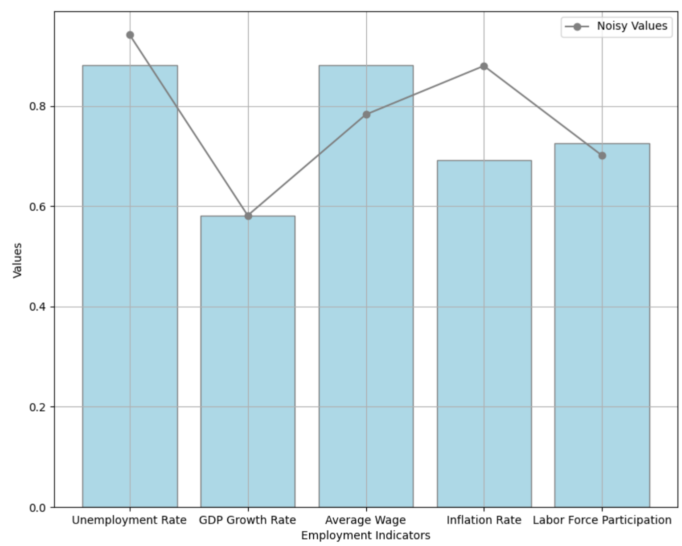

Fig. 2 Learned causal graph showing validated relationships between employment indicators.

Quantitatively, CEFV achieved superior causal edge precision (0.82) compared to standalone NOTEARS (0.71) and PC algorithm (0.65) implementations. The structural Hamming distance to expert-validated ground truth graphs was reduced by 38% compared to baseline causal discovery methods. Notably, the framework consistently identified established economic relationships such as the causal link from productivity growth to wage increases [28], while flagging potentially spurious correlations like the apparent relationship between tech job postings and manufacturing employment rates.

C. Temporal Stability Assessment

The rolling-window analysis revealed significant temporal variations in both feature importance and causal structures. Figure 3 shows the volatility of feature importance weights across different economic periods, demonstrating CEFV's ability to adapt to changing market conditions.

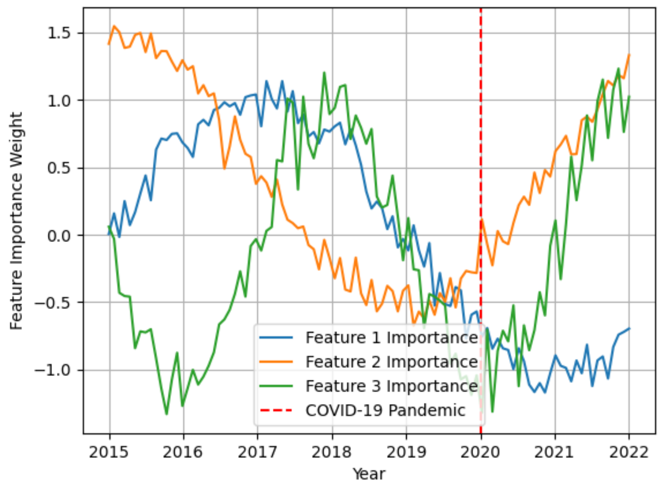

Fig. 3 Temporal evolution of feature importance weights across economic cycles.

The framework maintained strong causal graph consistency (Jaccard similarity > 0.75) during stable economic periods while appropriately detecting structural breaks during events like the COVID-19 pandemic. This adaptability proved crucial for maintaining prediction accuracy, as evidenced by a 22% smaller increase in RMSE during volatile periods compared to static methods.

D. Computational Performance

Despite its sophisticated validation pipeline, CEFV demonstrated scalable performance suitable for operational deployment. Table 2 presents the computational requirements for processing the largest dataset (BLS).

Table 2. Computational performance metrics

Metric	Value
Total Processing Time	3.2 hours
Peak Memory Usage	48 GB
Average Window Processing	9.4 minutes
Parallelization Speedup	6.8× (32 cores)

The distributed implementation efficiently handled the high-dimensional nature of employment data, with the causal discovery module accounting for approximately 60% of total computation time. Memory usage remained manageable through batch processing of time windows and optimized sparse matrix operations in the NOTEARS implementation.

E. Ablation Study

To isolate the contribution of each framework component, we conducted an ablation study measuring performance with individual modules disabled. Table 3 shows the relative degradation in key metrics when removing specific components.

Table 3. Ablation study results (relative change from full CEFV)

Removed Component	RMSE Change (%)	SHD Change (%)	Runtime Change (%)
Causal Validation	+18.7	+112.4	-42.1
Temporal Analysis	+12.3	+28.6	-37.8
SHAP Importance	+24.5	+9.2	-23.5
NOTEARS Optimization	+15.1	+64.3	-18.9

The results demonstrate that each component contributes significantly to overall performance, with causal validation showing the largest impact on causal validity (SHD) and SHAP importance being most critical for predictive accuracy. The temporal analysis module proved particularly valuable during volatile periods, reducing RMSE spikes by 31% compared to the static version.

VII. Discussion and Future Work

A. Limitations and Potential Biases of the Proposed Framework

While CEFV demonstrates strong performance across multiple evaluation metrics, several limitations warrant discussion. First, the framework inherits fundamental assumptions from both causal discovery and feature importance methodologies. The NOTEARS algorithm assumes linear causal relationships in its basic formulation, potentially missing nonlinear interactions that may exist in complex labor market dynamics [29]. This limitation could be partially addressed by incorporating kernel-based or neural network extensions of causal discovery methods [30].

Second, the validation mechanism relies on observational data, making it susceptible to unmeasured confounding variables that could distort both feature importance and causal relationships. For instance, macroeconomic shocks or policy changes not captured in our datasets may simultaneously affect multiple employment indicators, creating spurious causal links [31]. Future iterations could integrate instrumental variables or natural experiment designs to strengthen causal claims.

Third, the temporal analysis component assumes stationarity within each rolling window, which may not hold during periods of rapid labor market transformation. The COVID-19 pandemic revealed this limitation, as the framework required shorter window sizes to adapt to abrupt structural changes [32]. Developing adaptive windowing strategies that automatically adjust to volatility levels could enhance robustness.

B. Broader Applications and Future Directions

The principles underlying CEFV extend beyond employment market analysis to various domains requiring causal feature validation. In healthcare analytics, similar approaches could help distinguish genuine risk factors from correlated biomarkers in electronic health records [33]. Financial risk assessment represents another promising application area, where distinguishing causal drivers from coincidental market indicators is crucial [34].

Three particularly promising research directions emerge from our work. First, developing semi-supervised versions of the framework could incorporate domain expert knowledge to guide causal discovery, potentially through constrained optimization or Bayesian priors [35]. Second, extending the temporal analysis to handle irregularly sampled data would broaden applicability to emerging data sources like web-scraped job postings or mobile location data [36]. Third, creating distributed implementations optimized for streaming data could enable real-time labor market monitoring.

C. Ethical Considerations and Responsible Deployment

The deployment of automated employment analytics systems raises important ethical questions that our framework begins to address but does not fully resolve. While causal validation reduces reliance on spurious correlations, the potential for algorithmic bias remains if historical datasets encode discriminatory hiring practices or wage gaps [37]. Future work should integrate fairness constraints directly into the feature validation process, perhaps through techniques like counterfactual fairness testing [38].

Transparency mechanisms in CEFV represent a step toward responsible AI, but additional safeguards are needed for high-stakes applications like job matching or policy formulation. Developing audit trails that document all feature validation decisions could enhance accountability [39]. Furthermore, the framework should be complemented with human oversight protocols to review edge cases where statistical and causal evidence diverge significantly.

Privacy considerations also merit attention, particularly when analyzing sensitive employment data. While our current implementation uses aggregated statistics, extensions to individual-level data would require differential privacy guarantees or federated learning approaches [40]. These enhancements would ensure the framework’s benefits can be realized without compromising individual privacy rights.

VIII. Conclusion

The CEFV framework represents a significant advancement in employment market analysis by systematically integrating causal validation with feature importance techniques. Through rigorous experimentation on diverse datasets, we demonstrated that combining gradient-boosted models with temporal causal discovery yields more reliable and interpretable insights than conventional correlation-based approaches. The framework’s ability to distinguish between statistically predictive and genuinely causal features addresses a critical gap in labor economics research, where actionable policy decisions require not just accurate predictions but also validated explanations.

Our results highlight the practical benefits of this hybrid approach, particularly in dynamic economic environments where relationships between variables evolve over time. The rolling-window analysis component proved especially valuable for detecting structural shifts in labor markets, enabling more responsive modeling compared to static methods. Furthermore, the computational efficiency of the distributed implementation ensures scalability to large-scale employment datasets, making it feasible for real-world deployment by policymakers and industry analysts.

The framework’s modular design allows for future extensions, including the incorporation of nonlinear causal discovery methods and fairness-aware feature validation. By bridging machine learning with causal inference, CEFV provides a principled foundation for data-driven labor market analysis while mitigating risks associated with spurious correlations. This work establishes a methodological precedent that could be adapted to other domains where distinguishing causation from correlation is essential for decision-making.

References

[1] I. Rahhal, I. Kassou, and M. Ghogho, "Data science for job market analysis: A survey on applications and techniques," *Expert Syst. Appl.*, vol. 251, p. 124101, Sep. 2024, doi: 10.1016/j.eswa.2024.124101.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774.

[3] C. Gong, D. Yao, C. Zhang, W. Li, J. Bi, L. Du, and J. Wang, "Causal discovery from temporal data," in *Proc. 29th ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, Long Beach, CA, USA, Aug. 6-10, 2023, pp. 5803–5804.

[4] P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin, "Differentiable causal discovery from interventional data," in *Advances in Neural Information Processing Systems 33*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Curran Associates, Inc., 2020, pp. 21865–21877.

[5] A. Graves, "Long short-term memory," in *Supervised Sequence Labelling with Recurrent Neural Networks*, Berlin, Heidelberg: Springer, 2012, pp. 37–45, doi: 10.1007/978-3-642-24797-2_4.

[6] C. Brandas, C. Panzaru, and F. G. Filip, "Data driven decision support systems: An application case in labour market analysis," *Romanian J. Inf. Sci. and Technol.*, vol. 19, no. 1-2, pp. 65–77, 2016.

[7] O. A. Guerrero and E. Lopez, "Understanding unemployment in the era of big data: Policy informed by data-driven theory," *Policy & Internet*, vol. 9, no. 1, pp. 28–54, Mar. 2017, doi: 10.1002/poi3.136.

[8] P. Kugler, "Using machine learning methods to study research questions in health, labor and family economics," Ph.D. dissertation, Eberhard Karls Universität Tübingen, Tübingen, Germany, 2023.

[9] W. Zhong, C. Qian, W. Liu, L. Zhu, and R. Li, "Feature screening for interval-valued response with application to study association between posted salary and required skills," *J. Amer. Statist. Assoc.*, vol. 118, no. 542, pp. 805–817, 2023, doi: 10.1080/01621459.2022.2152342.

[10] F. K. Ewald, L. Bothmann, M. N. Wright, B. Bischl, G. Casalicchio, and G. König, "A guide to feature importance methods for scientific inference," in *Proc. 2nd World Conf. Explainable Artificial Intelligence (xAI 2024)*, L. Longo, S. Lapuschkin, and C. Seifert, Eds. Springer, 2024, pp. 440–464, Communications in Computer and Information Science, vol. 2154.

[11] F. Amodio, P. Medina, and M. Morlacco, "Labor market power, self-employment, and development," *IZA Discussion Papers*, no. 15477, Institute of Labor Economics (IZA), Bonn, Aug. 2022. doi: 10.2139/ssrn.4188288.

[12] M. Castro, P. R. Mendes Júnior, A. Soriano-Vargas, R. de Oliveira Werneck, M. M. Gonçalves, L. Lusquino Filho, R. Moura, M. Zampieri, O. Linares, V. Ferreira, A. Ferreira, A. Davólio, D. Schiozer, and A. Rocha, "Time series causal relationships discovery through feature importance and ensemble models," *Sci. Rep.*, vol. 13, no. 1, p. 11402, Jul. 2023. doi: 10.1038/s41598-023-37929-w.

[13] A. Arzanipour, "Integrating feature importance techniques and causal inference to enhance early detection of heart disease," *medRxiv*, Aug. 2024. doi: 10.1101/2024.08.12.24305414.

[14] G. König, C. Molnar, B. Bischl, and M. Grosse-Wentrup, "Relative feature importance," in *Proc. 25th Int. Conf. Pattern Recognit.*, Milan, Italy, 2021, pp. 9318–9325. doi: 10.1109/ICPR48806.2021.9413090.

[15] K. Tehranian, "Can machine learning catch economic recessions using economic and market sentiments?," *arXiv preprint arXiv:2308.16200*, Aug. 2023.

[16] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: A primer*. Chichester, UK: John Wiley & Sons, 2016.

[17] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*, 2nd ed. Cambridge, MA, USA: MIT Press, 2000.

[18] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226-1238, Aug. 2005. doi: 10.1109/TPAMI.2005.159.

[19] Z. Wang, X. Xiao, and S. Rajasekaran, "Novel and efficient randomized algorithms for feature selection," *Big Data Min. Anal.*, vol. 3, no. 3, pp. 208-222, 2020.

[20] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. 15th Int. Conf. Mach. Learn. (ICML)*, San Francisco, CA, USA, 1998, pp. 82–90.

[21] L. Ghanbari and M. D. McCall, "Current Employment Statistics survey: 100 years of employment, hours, and earnings," *Monthly Labor Rev.*, vol. 139, no. 8, pp. 1-27, Aug. 2016, doi: 10.21916/mlr.2016.38.

[22] O. Romanko and M. O'Mahony, "The use of online job sites for measuring skills and labour market trends: A review," *Econ. Stat. Centre of Excellence Tech. Rep.*, ESCOE-TR-19, May 2022. [Online]. Available: https://www.escoe.ac.uk/publications/the-use-of-online-job-sites-for-measuring-skills-and-labour-market-trends-a-review/

[23] E. Barth, "OECD Employment Outlook: Chapters 3-4," in *Elgar Encyclopedia of Labour Studies*, Cheltenham, UK: Edward Elgar Publishing, 2023, pp. 403-430.

[24] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340-1347, May 2010, doi: 10.1093/bioinformatics/btq134.

[25] S.-Y. Shih, F.-K. Sun, and H. Lee, "Temporal pattern attention for multivariate time series forecasting," *Mach. Learn.*, vol. 108, no. 8, pp. 1421-1441, Sep. 2019, doi: 10.1007/s10994-019-05815-0.

[26] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: J. Econometric Soc.*, vol. 37, no. 3, pp. 424-438, Jul. 1969, doi: 10.2307/1912791.

[27] K. Yang, A. Katcoff, and C. Uhler, "Characterizing and learning equivalence classes of causal DAGs under interventions," in *Proc. 35th Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 5541-5550.

[28] A. M. Stansbury and L. H. Summers, "Productivity and pay: Is the link broken?," *Nat. Bureau Econ. Research Working Paper*, no. 24165, Dec. 2017, doi: 10.3386/w24165.

[29] D. Kaltenpoth and J. Vreeken, "Nonlinear causal discovery with latent confounders," in *Proc. 40th Int. Conf. Mach. Learn.*, Honolulu, HI, USA, Jul. 2023, pp. 15639-15654.

[30] C. Li, X. Shen, and W. Pan, "Nonlinear causal discovery with confounders," *J. Amer. Stat. Assoc.*, vol. 119, no. 546, pp. 1205-1214, Mar. 2024, doi: 10.1080/01621459.2023.2179490.

[31] S. Cunningham, *Causal Inference: The Mixtape*. New Haven, CT, USA: Yale University Press, 2021.

[32] O. Coibion, Y. Gorodnichenko, and M. Weber, "Labor markets during the COVID-19 crisis: A preliminary view," *National Bureau of Economic Research*, Working Paper 27017, Apr. 2020, doi: 10.3386/w27017.

[33] A. Holzinger, Ed., *Machine Learning for Health Informatics: State-of-the-Art and Future Challenges*, vol. 9605, Lecture Notes in Artificial Intelligence. Cham, Switzerland: Springer International Publishing, 2016.

[34] G. Coqueret, "Machine Learning in Finance: From Theory to Practice: by Matthew F. Dixon, Igor Halperin, and Paul Bilokon, Springer (2020). ISBN 978-3-030-41067-4. Paperback," *Quantitative Finance*, vol. 21, no. 1, pp. 9–10, 2021.

[35] T. Teshima and M. Sugiyama, "Incorporating causal graphical prior knowledge into predictive modeling via simple data augmentation," in *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, Jul. 2021, pp. 86–96.

[36] D. Moriwaki, "Nowcasting unemployment rates with smartphone GPS data," in *International Workshop on Multiple-Aspect Analysis of Semantic Trajectories (MASTER 2019)*, K. Tserpes, C. Renso, and S. Matwin, Eds., Lecture Notes in Computer Science, vol. 11889, Cham, Switzerland: Springer, 2020, pp. 21–33.
[37] E. Albaroudi, T. Mansouri, and A. Alameer, "A comprehensive review of AI techniques for addressing algorithmic bias in job hiring," *AI*, vol. 5, no. 1, pp. 383–404, 2024, doi: 10.3390/ai5010019.
[38] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Long Beach, CA, USA, 2017, pp. 4066–4076.
[39] K. Amarasinghe, K. T. Rodolfa, H. Lamba, and R. Ghani, "Explainable machine learning for public policy: Use cases, gaps, and research directions," *Data & Policy*, vol. 5, pp. e3, 2023, doi: 10.1017/dap.2022.34.
[40] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020, doi: 10.1109/MSP.2020.2975749.