Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 10, No. 2, May, Pp. 215-228

Development of lung cancer risk screening tool with causal discovery model evaluation approach

Sandi Wibowo[1], Jatniko Nur Mutaqin[1], Ari Apriansyah[1], Amirul Iqbal[1], Muhamad Koyimatu[1], Gusti Ayu Putri Saptawati Soekidjo[1]
Bandung Institute of Technology, Indonesia[1]

Article Info
Keywords:
Structural Intervention Distance, Structural Hamming Distance, Matthews Correlation Coefficient, Frobenius Norm, Screening

Article history:
Received: December 12, 2024
Accepted: February 25, 2025
Published: May 31, 2025

Cite:
S. Wibowo, J. N. Mutaqin, A. Apriansyah, M. Komiyatu, and G. A. P. S. Soekidjo, "Development of Lung Cancer Risk Screening Tool with Causal Discovery Model Evaluation Approach", *KINETIK*, vol. 10, no. 2, May 2025.
https://doi.org/10.22219/kinetik.v10i2.2188

*Corresponding author.
Sandi Wibowo
E-mail address:
23523308@mahasiswa.itb.ac.id

Abstract
Causal graph discovery approaches in healthcare for detecting high-risk diseases have been more widely applied in the last decade. The main challenge in causal graph discovery in healthcare data is the complexity of big data, which requires appropriate algorithms to reveal causal relationships between variables. This study focuses on evaluating the performance of seven causal discovery models—Peter-Clark (PC), Greedy Equivalent Search (GES), Direct LiNGAM, Directed Acyclic Graph-Graph Neural Network (DAG-GNN), Greedy Sparsest Permutation (GraSP), and Recursive Causal Discovery (RCD)—on opensource healthcare datasets. The model performance was evaluated using the Structural Intervention Distance (SID), Structural Hamming Distance (SHD), Matthews Correlation Coefficient (MCC), and Fobernius Norm (FN) metrics. The evaluation results conclusively show that the GES model performs best on low-complexity datasets. Meanwhile, the DAG-GNN model offers consistent performance on high-complexity data with MCC values ranging from 0.77 to 0.88. The application of the GES model for lung cancer risk screening, based on user question responses, demonstrated effectiveness by measuring MCC, SID, and SHD scores between the reference adjacency metrics and the resulting screening metrics.

1. Introduction

Lung cancer detection has gained significant attention in recent years, particularly with the emergence of machine-learning techniques that improve early diagnosis capabilities and intervention strategies. Early detection is critical as lung cancer is the leading cause of death in cancer cases worldwide. Various studies have explored applying different machine learning models for lung cancer detection, emphasizing the importance of comparative analysis among these methods to identify the most effective approach. A comprehensive review of several studies on using machine learning algorithms in lung cancer prediction emphasizes the need for diagnosis for early treatment intervention to increase the probability of patient cure [1]. This review serves as an essential reference for understanding the application landscape of ML in lung cancer detection. Another study showed that comparing various ML algorithms, including decision tree and ensemble methods, demonstrated accuracy in detecting lung cancer [2]. These results show that different algorithms produce different levels of accuracy.

Utilizing causal discovery algorithms in healthcare has shown significant advantages over traditional machine learning methods. The main reason is the ability of causal discovery algorithms to identify causal relationships more accurately, which is often disregarded by machine learning models that are more correlation-based. Understanding causality is crucial in interpreting medical results, especially in revealing the risk of disease, where misinterpretation can have serious consequences [3]. According to this study, whereas machine learning methods can provide reliable predictions, they are frequently faced with problems of measurement bias and selection of factors that might influence the results [4].

Causal discovery methods have performed significant results in identifying causal relationships in the healthcare domain. Several studies addressed causal discovery techniques, which can analyze observed data to discover cause-and-effect relationships among observed variables. This approach can enhance data interpretability and the development of more accurate predictive models [5]. Implementing causal models, such as the PC (Peter-Clark) and GES (Greedy Equivalence Search) algorithms, enables the construction of causal graphs that can explain the relationship between risk factors in the lung cancer detection domain. However, applying these causal models in lung cancer detection requires further investigation and validation [6]. In addition to these approaches, several lung cancer risk prediction methods utilize feature selection to simplify the input variables, thereby improving the accuracy and applicability of the model in health monitoring [7].

Early intervention in lung cancer can be significantly improved by developing a risk screening tool and implementing causal discovery algorithms. Developing these tools requires determining the right causal discovery

model for performance and capability to handle high-complexity data. The consideration of the causal inference algorithm is essential to achieve a high-performance candidate model in the healthcare domain. The comparative analysis of these causal discovery algorithms would assist not only in identifying the most effective models but also in enhancing the understanding of the various causal and risk factors of lung cancer that are quantifiable numerically.

Condition-based methods employ a probability approach to discover causal relationships. By quantifying the mutual information between variables, this method enables the determination of causal relations based on the probability scores [9]. The basic concept in the context of applying this method is the transfer entropy [10]. Condition-based methods are divided into two sub-methods: constraint-based and score-based. Constraint-based causal discovery algorithms are Peter-Clark (PC) and MVPC (Missing value peter-clark). Meanwhile, algorithms with score-based methods include GraSP and GES. The basic formula of the PC algorithm is based on the conditional independence test between pairs of variables.

The modeling-based methods of discovering causality use a structural equation modeling (SEM) approach. The SEM technique integrates causal factor analysis and regression in one single analysis. Model analysis involves observed and latent variables to determine the causal relationship between variables in the graph structure [11]. Some algorithms involved in this method are Direct LiNGAM and RCD (Recursive Causal Discovery).

Deep learning-based methods discover causal relationships through deep learning techniques to automatically analyze large-scale data and detect the hidden variables that affect causal relationships. DAG-GNN is one of the deep learning-based methods; this algorithm combines Graph Neural Networks with a score-based approach to obtain directed acyclic graphs from observational data, then utilizes GNN for node embeddings and defines scores to evaluate causal structures [12].

This research contributes to the field of causal discovery in healthcare by systematically evaluating multiple causal discovery algorithms for lung cancer risk screening. While previous studies primarily focused on machine learning models that emphasize predictive accuracy, this study addresses a critical gap by analyzing the causal relationships between risk factors, enabling more interpretable and actionable insights. The key contribution of this study lies in the comparative assessment of seven causal discovery models such as PC, MVPC, GES, GraSP, Direct LiNGAM, DAG-GNN, and RCD across various complexity levels of medical datasets.

Furthermore, the objective of this research is not only to develop a lung cancer risk screening tool but also to enhance understanding of causal inference methodologies in the healthcare domain. By integrating causal discovery with a structured evaluation framework, this study provides healthcare practitioners with a scientifically grounded approach to identifying high-risk individuals based on causal relationships rather than mere correlations.

1.1 Greedy Equivalent Search (GES)

Greedy Equivalent Search (GES) is a standard algorithm used to study the structure of Bayesian networks. The GES algorithm is implemented by identifying a fitting model to describe the dependencies between variables. The basic principle of GES is to optimize the causality scoring function using the Bayesian Information Criterion (BIC) score [13][14]. This method generates a graph structure and then iteratively eliminates edges based on statistical validation [15]. This method is particularly effective in scenarios where the causal relationships between variables are complex. In several studies, a comparison of the GES algorithm was demonstrated to perform well in the context of graph accuracy when compared to other causal discovery algorithms [16]. The GES algorithm considers changes in the graph to maximize the likelihood or fitness of the model. In Equation 1, X and Y are the variables under test, while Z is the set of conditions of a variable, if X and Y are not conditionally independent of Z, then the edge is added or changed in the graph.

$$(add|remove)edge\ if\ X \perp Y \mid Z \tag{1}$$

Description:
X, Y = Variable under test
Z = Set of conditions

1.2 The Peter-Clark (PC)

The Peter-Clark (PC) algorithm is one of the algorithms in causality discovery designed to identify cause-and-effect relationships from observational data. It is a graph-based algorithm that utilizes variable independence to generate causal structures. The PC algorithm is implemented by examining the conditional independence between variables to determine causal relationships in the data [17][18]. The generation of the causality graph in the PC algorithm is initiated by revealing the network between variables and then determining the direction of the edge to the node through statistical variable independence testing [19]. In some studies, the PC algorithm can handle high-dimensional data [20]. The conditional independence relationship between two variables X and Y given a set of other variables Z is mathematically represented as Equation 2.

$$X \perp Y \mid Z \quad (2)$$

1.3 Missing Value Peter-Clark (MVPC)

The Missing Value PC (MVPC) algorithm is a method designed to solve the problem of revealing causality when datasets have missing values. MVPC is the latest development of the PC algorithm by integrating additional corrections to handle various missing data mechanisms consisting of missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [21]. In this context, MVPC contributes to restoring causal structure to unobserved variables [22][21]. These modifications enable MVPC to perform better with data limitations than the original PC algorithm [19]. Studies have demonstrated that MVPC can outperform traditional methods in specific scenarios, especially when dealing with complex causal structures that include latent variables that have not been clearly defined [5]. The conditional independence relationship between two variables X and Y given a set of other variables Z, including error terms ϵ, is mathematically represented as Equation 3.

$$X \perp Y \mid Z + \epsilon \quad (3)$$

Description:
X, Y = Variable under test
Z = Set of conditions
ϵ = Error terms

1.4 GrASP (Graphical Structure Learning via Sparse Penalty)

GraSP (Graphical Structure Learning via Sparse Penalty) is an innovative algorithm for causal discovery that uses a sparse penalty approach to improve the identification of causal relationships between variables. It utilizes sparsity in the learning process to handle large datasets by reducing the problem's dimensionality and focusing on the most significant variables [18]. The non-convex nature of the penalty values used in GraSP allows causality modeling to be applied flexibly [23][24]. The utilization of GraSP performs well on high-dimensional data, namely data characterized by the number of significant variables exceeding the number of observational data [5]. GraSP substantially generates causality structures from data with considerable noise [25]. The objective function for GraSP, including the sparse penalty term, is mathematically represented as Equation 4.

$$\lim_{G \in \varrho} \mathcal{L}(G; D) + \lambda \|G\| \quad (4)$$

Description:
G = Estimated graph structure
$\mathcal{L}(G; D)$ = Negative log-likelihood of data D
λ = Spare setting parameter

1.5 Direct LINGAM

Direct LiNGAM (Linear Non-Gaussian Acyclic Model) is a method of expressing causal relationships that utilizes the properties of Gaussian distribution and linear relationships between variables. The basic principle of Direct LiNGAM is to presume the existence of a directed acyclic graph (DAG) structure that represents the causal relationship for each variable. This characteristic allows the identification of causality direction based on statistical principles, and the elimination of causality direction is determined based on the standard distribution assumption score [26]. This approach efficiently identifies causal relationships in observational data with many complex assumptions [27]. The Direct LiNGAM algorithm has been applied in various domains, such as genetics and epidemiology, demonstrating its effectiveness in the health domain [28]. The basic linear model for the relationship between X and Y is shown in Equation 5. β and α are causality coefficients, while ϵY and ϵX are error values, Z is the variable that affects X.

$$\begin{aligned} Y &= \beta X + \epsilon_Y \\ X &= \alpha Z + \epsilon_X \end{aligned} \quad (5)$$

Description:
αZ = Causallity coefficients of *Z* variables
βX = Causallity coefficients of *X* variables
ϵ_Y, ϵ_X = Error values of variables *X* dan *Y*

1.6 DAG-GNN

Directed acyclic graph neural network (DAG-GNN) is a causal discovery algorithm that integrates graph theory with neural networks to model variable relationships represented by nodes and edges in a directed acyclic graph (DAG). DAG-GNN is a generative model that utilizes an autoencoder framework for graph structure learning [29]. It captures complex interrelationships among essential variables, such as climate causality studies and biological systems [30]. It also leverages deep learning capabilities to model complex relationships in observational data while maintaining the interpretability of causal graphs [5]. Some studies place the DAG-GNN algorithm as having high performance in causal inference, especially in high-dimensional datasets [25]. The basic formula of DAG-GNN uses neural network-based graph updating, represented as Equation 6.

$$h_v^{(t+1)} = \sigma\left(W.\sum_{u\in N(v)} h_u^{(t)} + b\right) \quad (6)$$

Description:

$h_v^{(t+1)}$ = Representatif of node *v* on step *t*
$N(v)$ = Set likelihood of node v
σ = Activation functions (ReLU or Sigmoid)
b = Bias

1.7 Recursive Casual Discovery

Recursive Causal Discovery (RCD) is an algorithm designed to identify causal relationships in complex datasets, primarily when implemented on relational data. The limitations of traditional causality models in capturing dependencies between variables in relational data prompted the development of the RCD causality discovery model [31]. The study indicates that the ability of the RCD algorithm is outstanding in handling relational data under certain assumption conditions [32]. Applying the RCD methodology to large datasets can effectively identify causal disclosure relationships in the health and environmental science domains [33]. RCD uses a recursion-based algorithm to detect the direction of causation between variables. The approach of this model is shown in Equation 7, where *f* and *g* are non-linear functions that describe the cause-and-effect relationship between variables.

$$Y = f(X,Z)\ or\ X = g(Y,Z) \quad (7)$$

Evaluation of causality disclosure models through comparative analysis of several algorithms requires the utilization of appropriate metrics to measure the effectiveness and accuracy of the algorithms. The metrics used in this study include Structural Hamming Distance, Structural Intervention Distance, Matthew Correlation Coefficient, and Frobenius Norm. Structural Hamming Distance (SID) and Structural Intervention Distance (SHD) are metrics frequently used in causality analysis to evaluate how well the model can explain the existing data structure. Hamming Distance measures the difference between two structures, while Intervention Distance assesses the impact of interventions in the model [34]. Matthew Correlation Coefficient (MCC) is also a metric that measures the quality of model predictions by considering all possible outcomes. MCC provides a more accurate overview of the model's performance than other metrics, focusing only on a single aspect of the outcome. In addition, the Frobenius Norm is utilized to measure the error between the matrix generated by the model and the reference matrix [35].

2. Research Method

This research methodology incorporates a comparative analysis approach of several causal discovery models to evaluate the performance in identifying causal relationships between variables in the dataset. The datasets are open-source data in the healthcare domain: LUCAS, ALARM, CHILD, SACHS, DIABETES, and LUNG CANCER. Model evaluation metrics are based on Structural Hamming Distance (SHD), Structural Intervention Distance (SID), and Matthews Correlation Coefficient (MCC) to measure the structural difference between the predicted graph and the reference graph. The model with the best performance is used as a candidate model for the lung cancer potential screening tool. The detailed framework of this research is shown in Figure 1.

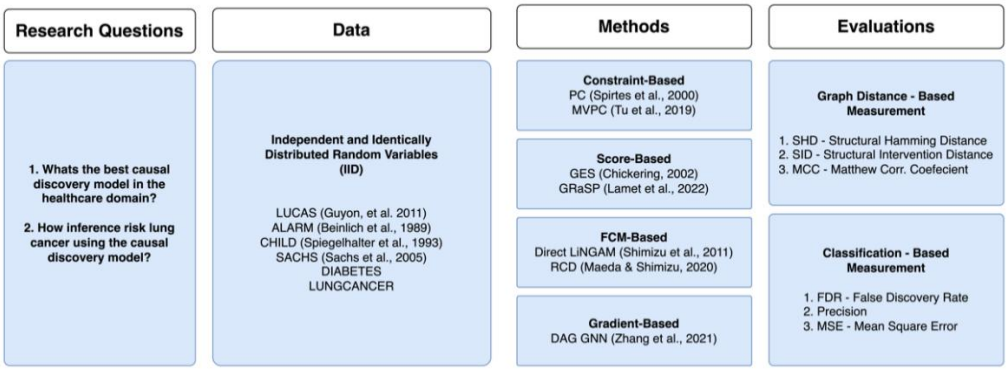

Figure 1. Research Framework

The model comparison process begins with using the identified benchmark datasets, and the algorithms are compared to understand the causal relationships between variables. After dataset identification and model definition, an iterative causality uncovering process enables continuous algorithm performance evaluation when faced with different training and testing data proportions. This process is repeated until adequate results are achieved and comprehensively evaluated based on the evaluation metrics. The process continues to the inference stage, where the best-performing candidate model is used to predict lung cancer risk based on the adjacency matrix generated from the screening form. At this stage, users need to fill in questions according to variables that have a causal relationship to the potential for lung cancer. The results of this inference are used to measure the SID, SHD, and MCC scores, which assess the similarity between the reference matrix and the screening matrix. These scores are then contextualized through a prompt into LLM to generate explanatory refinements so that users can understand the risk of lung cancer in the natural language. The complete flow chart of this research is shown in Figure 2.

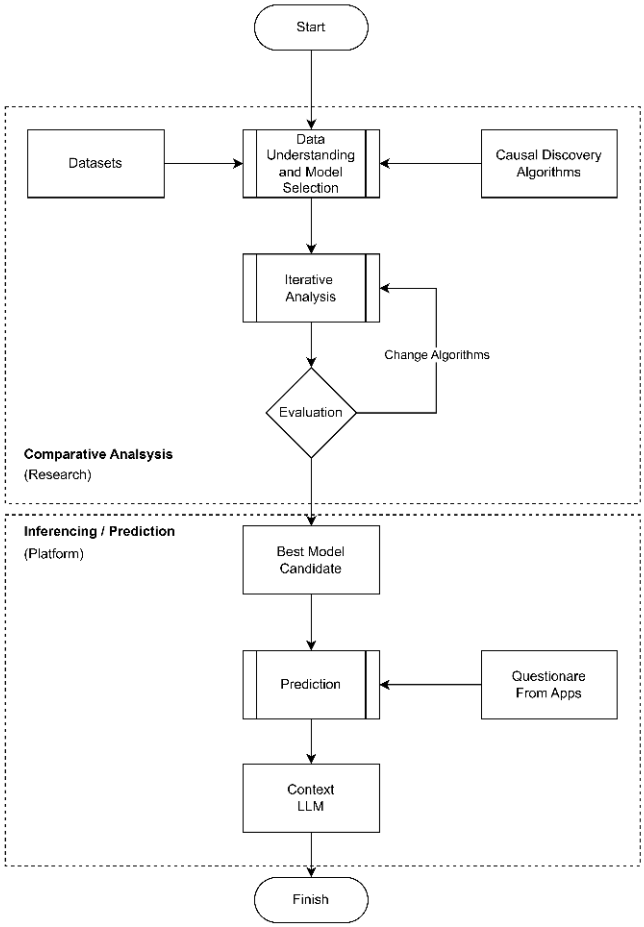

Figure 2. Research Flow Diagram

2.1 Datasets

The dataset for comparing algorithm performance uses data with independent and identically distributed (IID) types. Six data sources are used: LUCAS, CHILD, ALARM, SACHS, DIABETES, and LUNG CANCER. LUCAS data is a healthcare domain (medical) consisting of 2000 rows and 12 variables [36]. These datasets originate from different domains within healthcare and contain varying structures, including numerical and categorical variables. This dataset has a coefficient of variation (COV) characteristic on all variables ranging between 0.52 and 2.48, which indicates unbalanced characteristics. CHILD data is health domain data consisting of 5000 rows and 20 variables [37]. The characteristics of this data are that all variables are categorical types with a balanced distribution based on the Gini index value between 0.1 and 0.7. ALARM data is a dataset that represents a medical network that describes causal relationships in the context of patient monitoring. This data consists of 37 variables and 2000 rows. Ten numeric variables have COV values above 0.2, indicating an imbalanced distribution, while there are 27 categorical variables with a balanced distribution (chi-square = 1.0). The SACHS dataset is a flow cytometry experiment dataset in the biological domain that measures protein and phosphoprotein expression in human cells, consisting of 11 variables representing various proteins. This dataset contains 853 numeric samples with COV values between 0.2 and 0.5 (imbalance). DIABETES data is a causality dataset in the health sector consisting of 9 variables and 2768 data samples. This data is numeric, with COV between 0.3 and 1.4, indicating an imbalance in each variable. The LUNGSCANCER health dataset has 17 variables consisting of 11 that all show a balanced status (chi-square p-value = 1.0) and six numeric variables that show variations in balance.

To ensure the quality and consistency of these datasets before applying causal discovery algorithms, preprocessing steps were implemented. Data cleaning involved handling missing values using imputation techniques such as mean/mode substitution for numerical and categorical data, respectively, to prevent significant information loss. Normalization and standardization were applied to scale numerical variables, ensuring uniformity across datasets, particularly for algorithms sensitive to value ranges. Categorical encoding was performed by converting categorical variables into numerical representations using one-hot encoding or label encoding, depending on the algorithm's requirements. Additionally, the application of iterative analysis on the dataset was conducted by augmenting the dataset proportion between training and validation data, with proportions ranging from 10% training and 90% validation to 90% training and 10% validation. The iterative analysis on the dataset is perfomed by augmentating the dataset proportion between training and validation data. The combination of proportions begins with a composition of 10% training and 90% validation, up to 90% training and 10% validation.

The use of multiple datasets is justified by the need to evaluate the generalizability of causal discovery algorithms across different healthcare data structures. Each dataset represents distinct medical conditions, ensuring that the models are robust in diverse settings. The variations in variable types and distributions provide insight into how well each algorithm performs under different data complexities. This approach aligns with best practices in causal discovery research, where performance across varied datasets enhances model reliability and applicability in real-world scenarios.

2.2 Causal Discovery Model

The algorithms are classified into five approaches: Constraint-based, which includes PC and MVPC models that work by limiting statistical dependencies between variables to identify causal relationships; Score-based, such as GES and GraSP, which uses a score function to select the causal structure that best fits the data; modeling-based, such as the DirectLiNGAM and RCD algorithms that utilize causal function modeling to parse linear and non-linear relationships between variables; and Gradient-based, with the DAG-GNN algorithm that relies on gradient-based learning methods to detect causal structures. The analytical implementation procedure is executed within a Python programming framework utilizing the Causallearn and Castle libraries.

2.3 Evaluation Metrics

The presented performance measurement evaluation metrics cover a variety of approaches used to evaluate the performance of causality and graph structure models. These metrics are divided into two main categories: distance-based graph measures and classification-based measures, each of which plays an essential role in assessing the accuracy and fit of the model to the reference data.

2.3.1 Structural Hamming Distance

The edge discrepancies between the two graphs are computed by analyzing their corresponding adjacency matrices. A score of 0 in the Structural Hamming Distance (SHD) signifies that the two graphs are congruent. G1 and G2 are two networks that are being compared. E1 is a collection of edges in G1, E2 is a collection of edges in G2, E1 \ E2 is an edge that is in G1 but not in G2, E2 \ E1 is the edge that is in G2, but not in G1, RE is the number of edges whose direction is different between G1 and G2. The Structural Hamming Distance between the two graphs is mathematically represented as Equation 8.

$$SHD(G_1, G_2) = |E_1 \setminus E_2| + |E_2 \setminus E_1| + |RE| \quad (8)$$

2.3.2 Structural Intervention Distance

The Structural Intervention Distance (SID) highlights the importance of causal directionality alignment as established by the interventions executed, with a value of 0 denoting total unity between the target and the expected graphical depictions. G_1 and G_2 are the two networks being compared, n is the number of nodes in the network, i and j are the nodes being compared in the network, $P_G(i \rightarrow j)$ is the outcome of the effect of the intervention on i of j in the network G,$1(.)$ is an indicator function with the value one if its argument is valid and 0 if false. The Structural Intervention Distance between the two networks is mathematically represented as Equation 9.

$$SID(G_1, G_2) = \sum_{i=1}^{n} \sum_{j \neq 1} 1(P_{G_1}(i \rightarrow j) \neq (P_{G_2}(i \rightarrow j)) \quad (9)$$

2.3.3 Frobenius Norm

Frobenius norms (FN) assess all inconsistencies present within the graph structure, including the absence or presence of edges and differences in directionality, where a result of 0 indicates that the two graphs are identical. In Equation 10, A is an m×n, matrix a_{ij} is the element of matrix A in the row of i and j column. *m* and *n* are the number of rows and columns of A, respectively.

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} \quad (10)$$

2.3.4 Matthew Correlation Coefficient (MCC)

This metric measures how well a causal uncovering algorithm predicts the existence and direction of edges in a reference graph, with values ranging from -1 to +1, where +1 indicates perfect prediction. In Equation 11, TP is the number of positive detections that are true positive, TN is the number of negative detections that are true negative, FP is the number of negative cases that are detected as positive cases, and FN is the number of positive cases that are detected as negative.

$$MCC = \frac{TP.TN - TP.FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (11)$$

2.3.5 False Discovery Rate

False discovery rate (FDR) measures the proportion of causal relationships that the model incorrectly detects. FDR values range from 0 to 1, where 0 indicates no incorrect detection of causal relationships. Precision measures the proportion of correct causal relationships out of the total relationships predicted as positive by the model, with a value of 1 indicating a perfect prediction. TP is the number of positive detections that are confirmed positive, and FP is the number of negative cases that are detected as positive cases. The formula for FDR is mathematically represented as Equation 12.

$$FDR = \frac{FP}{FP + TP} \quad (12)$$

2.4 Model Inference and Tool Development

The causal discovery inference method for lung cancer screening involves identifying causal relationships between variables identified as causes of cancer based on the causal graph generated by the model. The variables based on the questionnaire answers are then represented as a screening graph that will be compared with the reference graph. Furthermore, MCC and FN-based evaluation metrics are used to measure the similarity of the screening graph pattern with the reference causal graph. By comparing these scores, we can calculate the suitability of the screening graph to the reference causal structure so that users get information on the potential risk of lung cancer numerically based on the MCC, SHD, and SID metrics. The development of the lung cancer risk screening tool is carried out in the Python programming environment. The Python framework used is Streamlit for interactive website management. The causal disclosure algorithm of the analysis results is used as reference knowledge to detect potential cancer risks. Transferring knowledge from the analysis environment to the tool environment (transfer knowledge) is carried out by saving the adjacency matrix of the analysis results in pickle format and then used in the application. The main

components of the tool consist of input, analysis, and output sections. The input section consists of a screening form adjusted to the variables that have a causal relationship to lung cancer. The process section consists of a causal disclosure algorithm and a large language model (LLM) to define the results of the graph suitability measurement. Then, the output section displays the results of the causality graph inferred by the user's screening results and the SID, SHD, and MCC suitability metric scores.

3. Results and Discussion

The results of the comparative analysis comparing seven causal discovery algorithms on six benchmark datasets were conducted to obtain candidate models for developing lung cancer risk screening tools. This comparison includes evaluating the performance of the PC, MVPC, GES, GraSP, DirectLiNGAM, RCD, and DAG-GNN algorithms tested on LUCAS, ALARM, SACHS, CHILD, DIABETES, and LUNGSCANCER data, considering aspects of accuracy, precision, and generalization ability in identifying causal relationships. The results of the analysis show significant performance variations between algorithms.

3.1 Model Performance

The results of iterative analysis on the PC, GES, Direct LINGAM, MVPC, GraSP, DAG-GNN, and RCD algorithms on the disclosure of causality graphs of health domain dataset produce varying evaluation metric values. The complexity of the dataset used significantly influences the evaluation metric value. Based on the measurement of distance-based evaluation metrics, namely SID and SHD, the information obtained is shown in Figure 3.

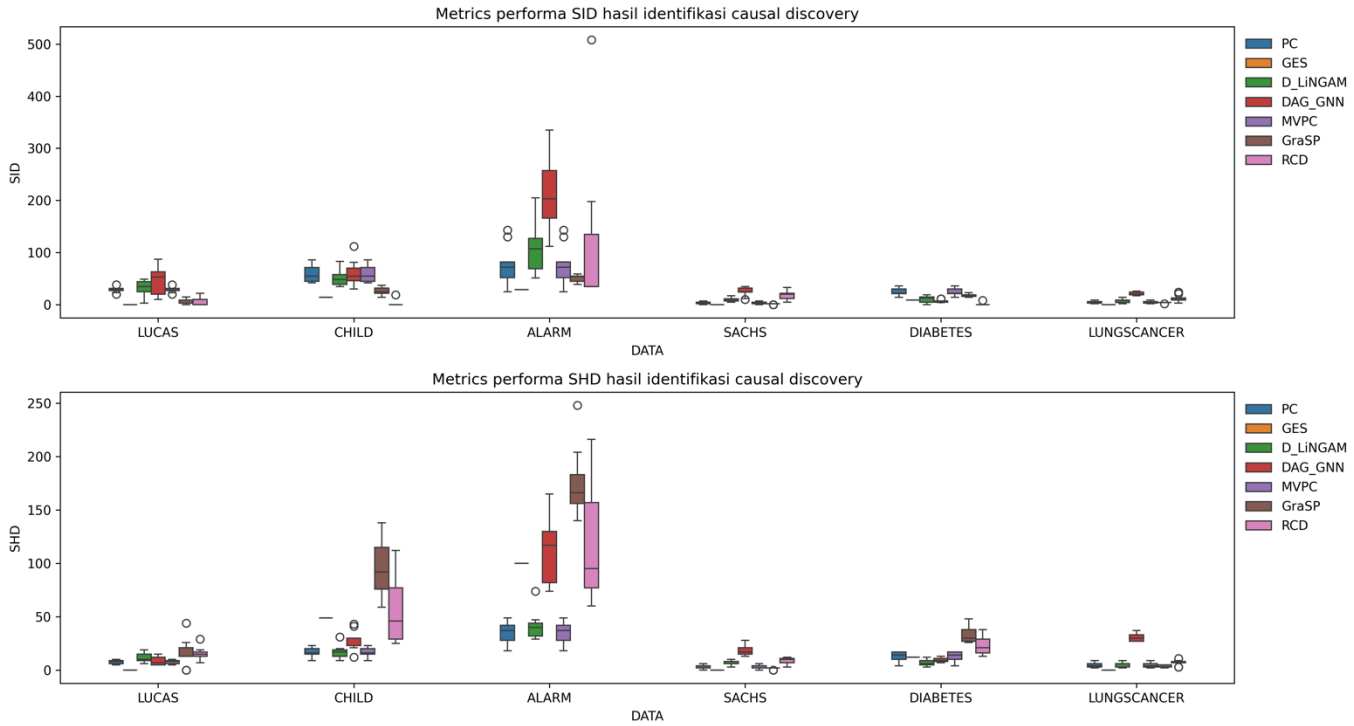

Figure 3. The Metrics Evaluations of (a). Structural Intervention Distance and (b). Structural Hamming Distance

The Structural Hamming Distance (SHD) and Structural Intervention Distance (SID) metrics show the level of difference between the actual graph and the predicted graph. From the performance displayed in the boxplot, it can be seen that data complexity (number of variables) significantly affects the graph generation error. In datasets with high complexity, such as ALARM, the average algorithm performance value of SID and SHD are 96 and 88, respectively. This shows that causal direction errors are directly proportional to the errors in the number of structures. However, the SID value is higher than SHD, indicating that the causal relationship error is more prominent than the mistake in predicting the number of nodes and edges. The performance of most algorithms on other datasets with low to moderate complexity shows that the algorithm's ability remains relatively stable, with SID and SHD scores below 100.

Meanwhile, based on boxplot size variation, several models demonstrate high detection consistency despite differences in training and testing proportions. The PC, MVPC, GES, and Direct LiNGAM algorithms exhibit higher detection consistency than DAG-GNN, GraSP, and RCD. The RCD and DAG-GNN algorithms are significantly influenced by dataset proportions, with the highest performance achieved at 40% to 60% training-to-testing data

proportions. Based on the SHD and SID performance, the GES (Greedy Equivalent Search) algorithm shows the highest performance, achieving SHD = 0 and SID = 0 on the LUCAS dataset, indicating its effectiveness in reconstructing correct causal structures, although with a risk of overfitting. Deep learning-based algorithms such as DAG-GNN exhibit moderate performance when implemented with an appropriate dataset proportion.

The evaluation of algorithm performance using classification-based measurements, namely Matthew Correlation Coefficient (MCC) and Frobenius Norm (FN), provides insight into prediction accuracy. MCC assesses causal structure classification quality, considering true positives, true negatives, false positives, and false negatives, with scores ranging from -1 to +1, where higher values indicate more accurate predictions. Meanwhile, Frobenius Norm (FN) measures differences between predicted adjacency matrices, where lower FN values indicate higher similarity between the predicted causal structure and the actual structure. The comparative analysis results highlight the performance variations among the seven tested algorithms, as illustrated in Figure 4.

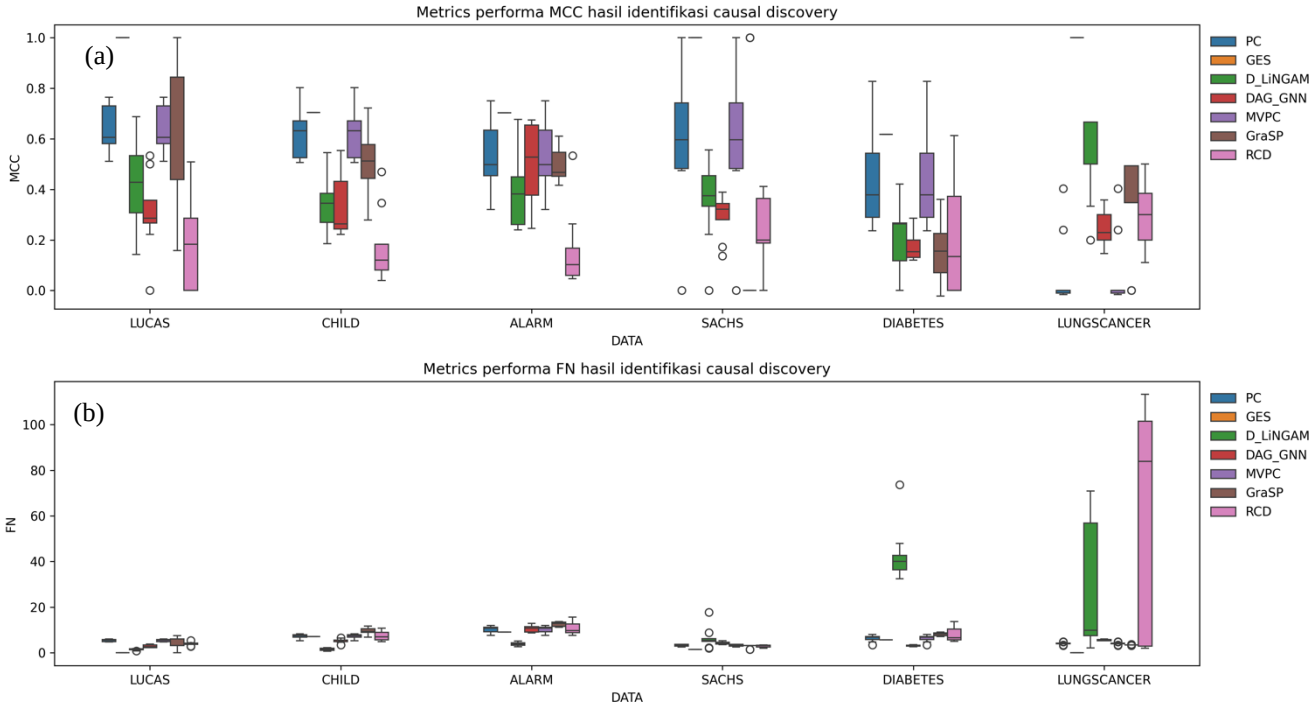

Figure 4. The Metrics Evaluations of (a). Mathew Coefficient of Correlation and (b). Frobenius Norm

The results of the MCC metric evaluation (Table 1) show variations in algorithm performance in predicting causality from six benchmark datasets. The algorithm with consistent MCC scores above 0.5 is GES, while other algorithms tend to be influenced by data complexity and variations in the proportion of training and testing. The RCD algorithm has low performance on all datasets. Meanwhile, the PC and MVPC algorithms perform well on several datasets but poorly on the LUNGSCANCER dataset, with MCC values ranged between -0.1 and 0.4.

Table 1. Median MCC Values for Causal Discovery Models Across Datasets

Data	DAG_GNN	D_LiNGAM	GES	GraSP	MVPC	PC	RCD
ALARM	0.527132	0.381818	0.702831	0.467022	0.498234	0.498234	0.102564
CHILD	0.263158	0.344828	0.704026	0.512007	0.631669	0.631669	0.120000
DIABETES	0.153846	0.263158	0.617964	0.155941	0.377964	0.377964	0.134615
LUCAS	0.285714	0.428571	1.000000	0.843345	0.606288	0.606288	0.183333
LUNGSCANC.	0.229167	0.500000	1.000000	0.347375	0.000000	0.000000	0.300000
SACHS	0.321429	0.375000	1.000000	0.000000	0.596557	0.596557	0.200000

The DAG-GNN algorithm, with a gradient-based approach, shows increased performance on data with high complexity, but the MCC score tends to be small on data with a small number of variables. The performance characteristics of DAG-GNN can reveal causal relationships in data with many variables. This is in contrast to the GES algorithm, which is identified as experiencing a decrease in accuracy on data with high complexity. The Direct LiNGAM algorithm based on FCM shows moderate performance on datasets with a small to medium number of columns. Decreased performance is seen on datasets with significant variables caused by limitations in handling non-linear

dependencies between variables. Evaluation of the Frobenius norm (FN) score on the compared algorithms shows relatively small variation results, where the average FN score is below 20. The FN value on the DIABETES and LUNGSCANCER datasets identified several algorithms, such as Direct LiNGAM and RCD, that obtained high score variations. The RCD algorithm on the LUNGCANCER dataset has a variation in FN values from 1.9 to 113.2, which shows that this algorithm is highly influenced by the proportion of data and characteristics of the dataset being analyzed. The RCD algorithm gets the best FN score on a data proportion of 80% training and 20% testing, while based on data characteristics, this algorithm performs poorly on data with high data type complexity and disproportionate data balance. The Direct LiNGAM algorithm has the same characteristics as RCD in its application on the LUNGS CANCER dataset, which has a variation in FN between 0.6 and 70.8, while the optimum proportion of the dataset is between 80% and 90% for the training process.

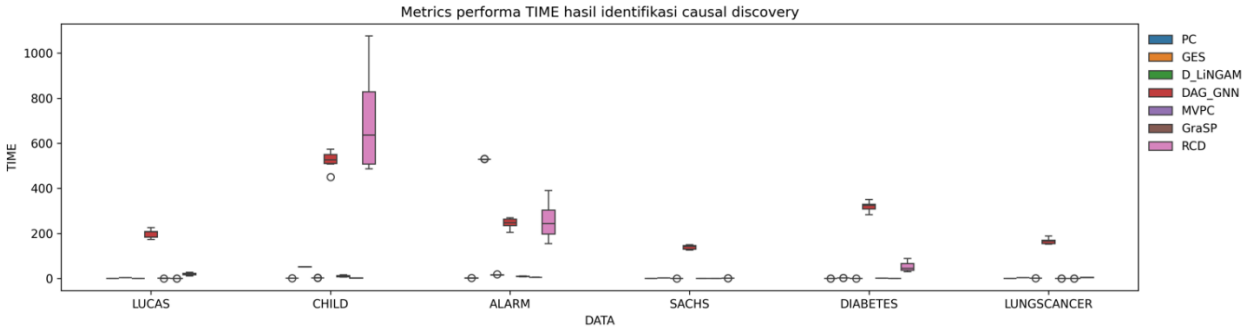

Figure 5. Time Requirements

The analysis time requirement parameters, as shown in Figure 5, show that the DAG-GNN and RCD algorithms requirelonger analysis time than other algorithms with more than 400 seconds. This time requirement is due to the analysis process using high computing resources to run the deep learning-based analysis process. On the other hand, the DAG-GNN and RCD algorithms can generate causal relationships in data with complex variables. Still, as compensation, they require longer analysis time than other algorithms. The overall evaluation results show that the GES algorithm has high performance in generating causal graphs on health domain datasets but has weaknesses in datasets with high complexity. This follows previous studies, which state that the GES algorithm consistently detects DAG graphs from non-parametric data and can handle specific test errors by minimizing nodes and edges [38]. The GES algorithm performs optimally on low-complexity large samples, guaranteeing maximum convergence of the generative structure class [39]. Overall, the evaluation results confirm that GES performs best in generating causal graphs on health domain datasets. However, its performance declines in high-complexity datasets. This finding aligns with previous studies, indicating that GES consistently detects Directed Acyclic Graph (DAG) structures from non-parametric data while minimizing false positive causal edges. The GES algorithm is optimal for low-complexity, large-sample datasets, ensuring maximum generative structure convergence.

3.2 Inference Evaluation

The performance evaluation of the causal discovery algorithm identified that the Greedy Equivalent Search (GES) algorithm has good performance, with the MCC metric score reaching 1 and the SID and SHD scores being 0. The causal graph results of the GES algorithm are shown in Figure 6 and will be used as a reference causal graph in the questionnaire-based lung cancer risk screening method.

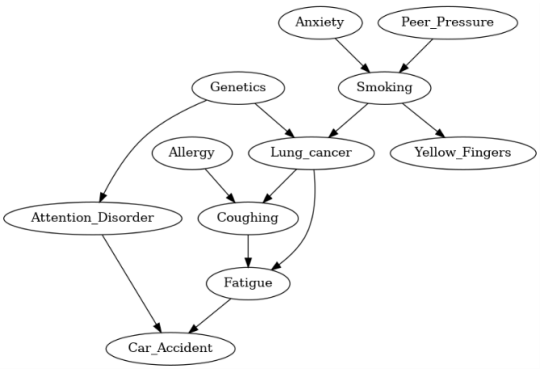

Figure 6. Graph Causal Discovery GES (LUCAS)

The lung cancer variable is a variable that is based on the causality graph as a result of the smoking and genetics variables. It is also a causal factor for coughing and fatigue. The two variables that are the result of lung cancer are symptom variables that indicate lung cancer. Based on the causal graph, we compile screening variables to detect the risk of lung cancer based on three categories, namely direct causes: smoking and genetics; direct symptoms: coughing and fatigue; and indirect causes: anxiety and peer pressure. The tool development process is carried out through the knowledge transfer process of the greedy equivalent search (GES) algorithm from the analysis environment to the application environment. This knowledge transfer process uses the Python pickle library to store the adjacency matrix parameters and the actual graph of the model detection results. The application framework used is streamlit, considering its ease of development and interactive features. The results of the tool development are installed on a virtual machine and can then be accessed using the URL https://hc.labdata.id. The input form is developed based on the identification of the causality variables that show the factors that cause and affect lung cancer. Integration with the large language model (Large Language Model, LLM) gemma:27b is carried out to produce explanatory information on the similarity score of the analysis. The appearance of the tool result interface is shown in Figure 7.

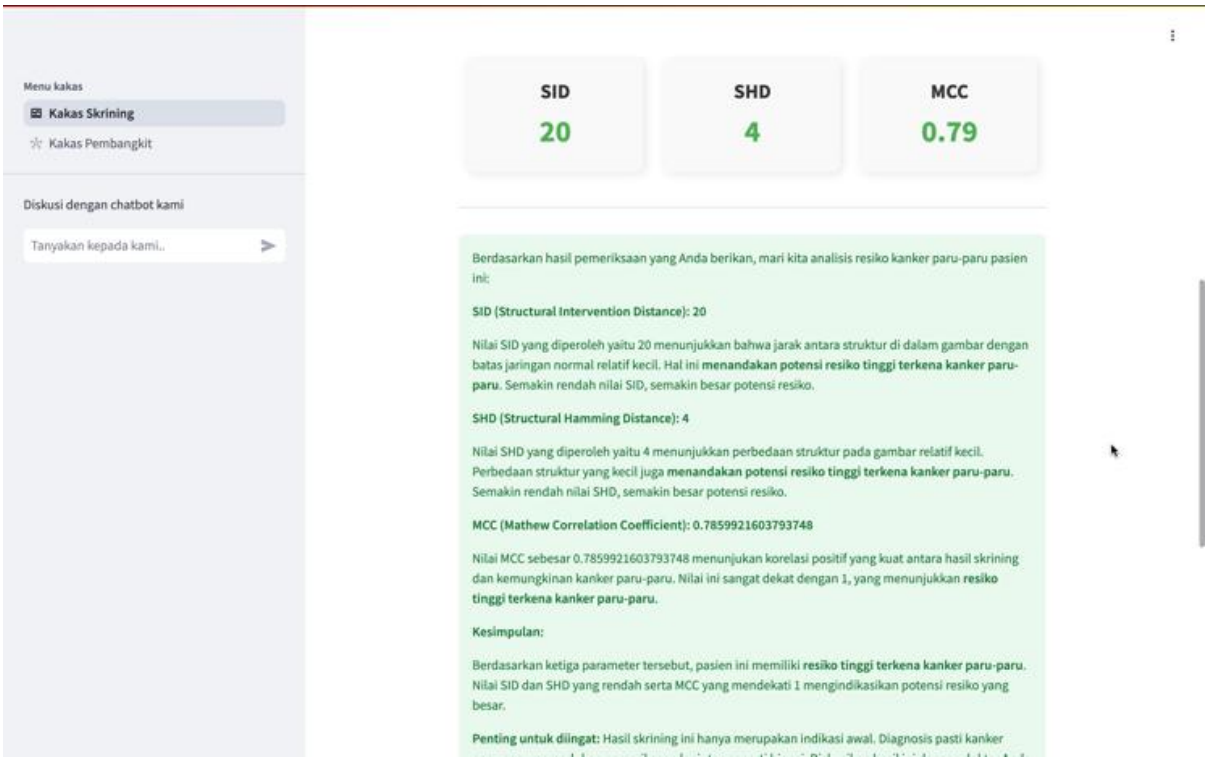

Figure 7. Output Interface Visualization of Lungs Cancer Screening

The lung cancer risk detection tool provides multiple outputs, including a causal graph, risk score, adjacency matrix comparison, and LLM-based explanation. The tool visualizes both the reference causal graph derived from the developed model and the user-inferred graph generated based on screening responses. To ensure that the system aligns with the developed model, Structural Hamming Distance (SHD), Structural Intervention Distance (SID), Matthews Correlation Coefficient (MCC), and Frobenius Norm (FN) are utilized to quantify the structural similarity between the reference and inferred graphs.

Additionally, the usability of the system is enhanced through an interactive interface and an LLM-based explanation mechanism, which translates numerical similarity scores into interpretable insights. The system's readiness has been evaluated through performance benchmarking, ensuring real-time response and computational efficiency. These validation steps confirm the accuracy, usability, and robustness of the lung cancer screening tool. The tool output graph consists of a reference graph (knowledge) inferred by the user with red lines in the form of edges and nodes. This color indicates that the causality in that section is validated based on the user's input in the screening form. Then, the following tool output is an adjacency matrix consisting of a reference matrix and screening result metrics. The adjacency matrix in Figure 10 visualizes user responses when filling out the questionnaire based on screening variables. The two matrices will be measured for their level of similarity using the Matthew Correlation Coefficient (MCC) and Frobenius Norm (FN) measurements, indicating a high risk if the MCC value approaches 1 and FN approaches 0. The results of user inference on the reference graph are displayed in Figure 8 to provide visual information in the form of a

causality graph, which shows that the nodes significantly affect the risk of lung cancer. With this description, users will get visual information about which variables cause direct and indirect effects on lung cancer nodes.

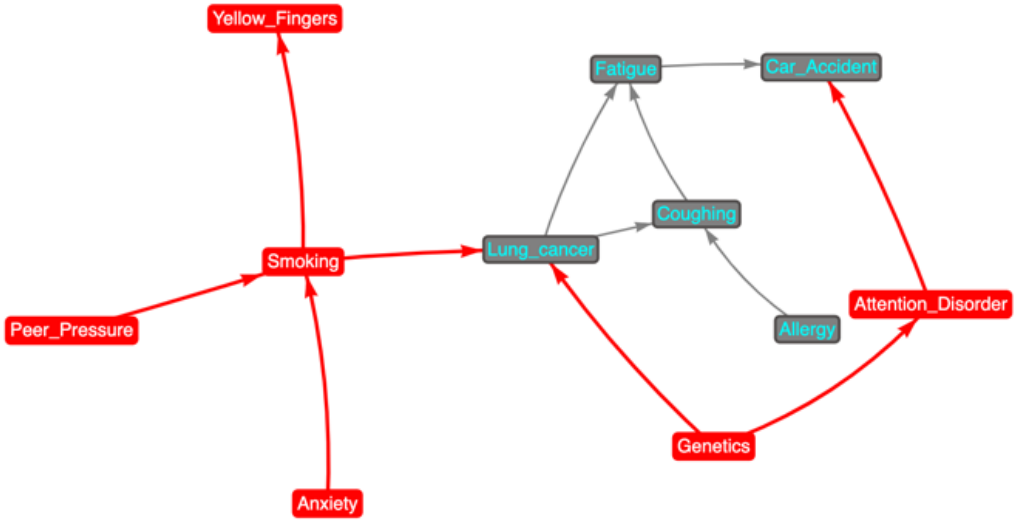

Figure 8. Inferred Graph from Questionnaire User

4. Conclusion

The comparative results of causal discovery algorithms on health domain data (medical) produce information that the Greedy Equivalent Search (GES) algorithm performs best. The GES algorithm is relatively stable in revealing causality with imbalanced data types and low complexity. Meanwhile, the trend of increasing algorithm performance in applying high-complexity data is shown in the DAG-GNN machine learning-based algorithm, which has the characteristic of growing accuracy with many variables. However, applying DAG-GNN requires more excellent computing resources and processing time than the GES algorithm. The implementation of the GES algorithm in detecting lung cancer risk has been carried out by measuring the suitability metric by comparing the reference adjacency metric and screening results. Further research can focus on increasing the number of datasets and adding variables that cause potential lung cancer. In additional studies, datasets with higher complexity can be used to compare the performance between models.

Acknowledgment

I want to express my sincere gratitude to the faculty members and lecturers from the Bandung Institute of Technology for their valuable support and guidance throughout this study.

References

[1] Krishna, R. S. (2023). Machine Learning Approaches in Early Lung Cancer Prediction: A Comprehensive Review. INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, 07(09). https://doi.org/10.55041/IJSREM25584
[2] P.R., R., Nair, R. A. S., & G., V. (2019). A Comparative Study of Lung Cancer Detection using Machine Learning Algorithms. 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 1–4. https://doi.org/10.1109/ICECCT.2019.8869001
[3] Castro, D. C. de, Walker, I. D., & Glocker, B. (2020). Causality Matters in Medical Imaging. In Nature Communications. https://doi.org/10.1038/s41467-020-17478-w
[4] Doupé, P., Faghmous, J. H., & Basu, S. (2019). Machine Learning for Health Services Researchers. In Value in Health. https://doi.org/10.1016/j.jval.2019.02.012
[5] Nauta, M., Bucur, D., & Seifert, C. (2019). Causal Discovery with Attention-Based Convolutional Neural Networks. Machine Learning and Knowledge Extraction, 1(1), 312–340. https://doi.org/10.3390/make1010019
[6] Runge, J. (2018). Causal network reconstruction from time series: From theoretical assumptions to practical estimation. Chaos: An Interdisciplinary Journal of Nonlinear Science, 28(7). https://doi.org/10.1063/1.5025050
[7] Xie, N.-N., Hu, L., & Li, T.-H. (2015). Lung Cancer Risk Prediction Method Based on Feature Selection and Artificial Neural Network. Asian Pacific Journal of Cancer Prevention, 15(23), 10539–10542. https://doi.org/10.7314/APJCP.2014.15.23.10539
[8] Niu, W., Gao, Z., Song, L., & Li, L. (2024). Comprehensive Review and Empirical Evaluation of Causal Discovery Algorithms for Numerical Data. http://arxiv.org/abs/2407.13054
[9] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinović, D. (2019). Detecting and Quantifying Causal Associations in Large Nonlinear Time Series Datasets. In Science Advances. https://doi.org/10.1126/sciadv.aau4996
[10] Schreiber, T. (2000). Measuring Information Transfer. In Physical Review Letters. https://doi.org/10.1103/physrevlett.85.461
[11] Shimizu, S., Hoyer, P. O., & Hyvärinen, A. (2009). Estimation of Linear Non-Gaussian Acyclic Models for Latent Factors. In Neurocomputing. https://doi.org/10.1016/j.neucom.2008.11.018

[12] Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., López-Paz, D., & Sebag, M. (2018). Learning Functional Causal Models With Generative Neural Networks. https://doi.org/10.1007/978-3-319-98131-4_3
[13] Bühlmann, P., Peters, J., & Ernest, J. (2014). CAM: Causal Additive Models, High-Dimensional Order Search and Penalized Regression. In The Annals of Statistics. https://doi.org/10.1214/14-aos1260
[14] Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. In Machine Learning. https://doi.org/10.1007/s10994-006-6889-7
[15] David Maxwell Chickering. (1996). Learning bayesian networks is np-complete. Learning from data: Artificial intelligence and statistics V: Vol. V. Springer.
[16] Singh, K., Gupta, G., Tewari, V., & Shroff, G. (2018). Comparative benchmarking of causal discovery algorithms. Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, 46–56. https://doi.org/10.1145/3152494.3152499
[17] Cao, L., Su, J., Cao, Y., Siang, L. C., Li, J., Saddler, J., & Gopaluni, R. B. (2022). Causal Discovery Based on Observational Data and Process Knowledge in Industrial Processes. In Industrial & Engineering Chemistry Research. https://doi.org/10.1021/acs.iecr.2c01326
[18] Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. In Frontiers in Genetics. https://doi.org/10.3389/fgene.2019.00524
[19] Lu, N. Y., Zhang, K., & Yuan, C. (2021). Improving Causal Discovery By Optimal Bayesian Network Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(10), 8741–8748. https://doi.org/10.1609/aaai.v35i10.17059
[20] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science Advances, 5(11). https://doi.org/10.1126/sciadv.aau4996
[21] Tu, R., Zhang, K., Ackermann, P. W., Bertilson, B. C., Glymour, C., & Zhang, C. (2018). Causal Discovery in the Presence of Missing Data. https://doi.org/10.48550/arxiv.1807.04010
[22] Qiao, J., Chen, Z., Yu, J., Cai, R., & Hao, Z. (2024). Identification of Causal Structure in the Presence of Missing Data With Additive Noise Model. In Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v38i18.30036
[23] Niu, Y., & Fei, J. (2021). A Sparsity-Assisted Fault Diagnosis Method Based on Nonconvex Sparse Regularization. In Ieee Access. https://doi.org/10.1109/access.2021.3073072
[24] Dong, Z., Lin, G., & Nian-dong, C. (2021). An Inexact Penalty Decomposition Method for Sparse Optimization. In Computational Intelligence and Neuroscience. https://doi.org/10.1155/2021/9943519
[25] Kalainathan, D., & Goudet, O. (2019). Causal Discovery Toolbox: Uncover causal relationships in Python.
[26] Kawaguchi, H. (2023). Application of Quantum Computing to a Linear Non-Gaussian Acyclic Model for Novel Medical Knowledge Discovery. In Plos One. https://doi.org/10.1371/journal.pone.0283933
[27] Kawaguchi, H. (2022). Application of quantum computing to a linear non-Gaussian acyclic model for novel medical knowledge discovery. https://doi.org/10.21203/rs.3.rs-1264829/v1
[28] Wu, J., & Drton, M. (2023). Partial Homoscedasticity in Causal Discovery With Linear Models. IEEE Journal on Selected Areas in Information Theory, 4, 639–650. https://doi.org/10.1109/JSAIT.2023.3328476
[29] Yu, Y., Chen, J., Gao, T., & Yu, M. (2019). DAG-GNN: DAG Structure Learning With Graph Neural Networks. https://doi.org/10.48550/arxiv.1904.10098
[30] Huang, Y., Kleindeßner, M., Munishkin, A. A., Varshney, D., Guo, P., & Wang, J. (2021). Benchmarking of Data-Driven Causality Discovery Approaches in the Interactions of Arctic Sea Ice and Atmosphere. In Frontiers in Big Data. https://doi.org/10.3389/fdata.2021.642182
[31] Lee, S., & Honavar, V. (2016). On Learning Causal Models from Relational Data. In Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v30i1.10417
[32] Ahsan, R., Arbour, D., & Zheleva, E. (2023). Learning Relational Causal Models With Cycles Through Relational Acyclification. In Proceedings of the AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v37i10.26434
[33] Yuan, Y., Ding, X., & Bar-Joseph, Z. (2020). Causal Inference Using Deep Neural Networks. https://doi.org/10.48550/arxiv.2011.12508
[34] Akil, Y. S., & Lateko, A. A. H. (2023). Analisis Kausalitas Antara Produksi Listrik RE Dengan CPI Dan GDP Di Indonesia. In Jurnal Teknik Elektro Uniba (Jte Uniba). https://doi.org/10.36277/jteuniba.v8i1.234
[35] Naibaho, R. (2020). Analisis Tingkat Pengungkapan Transaksi Pihak Berelasi Dan Pengaruhnya Terhadap Nilai Perusahaan (Studi Pada Industri Manufaktur). In Abis Accounting and Business Information Systems Journal. https://doi.org/10.22146/abis.v7i4.58861
[36] Guyon, I., Aliferis, C. F., Cooper, G. S., Elisseeff, A., Pellet, J., Spirtes, P., & Statnikov, A. (2011). Causality Workbench. https://doi.org/10.1093/acprof:oso/9780199574131.003.0026
[37] Lauritzen, S., & Spiegelhalter, D. (1988). Local Computations With Probabilities on Graphical Structures and Their Application to Expert Systems. In Journal of the Royal Statistical Society Series B (Statistical Methodology). https://doi.org/10.1111/j.2517-6161.1988.tb01721.x
[38] Bryon Aragam. (2024). Greedy equivalence search for nonparametric graphical models. Arxiv.
[39] Chickering, M. (2020). Statistically Efficient Greedy Equivalence Search. In J. Peters & D. Sontag (Eds.), Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI) (Vol. 124, pp. 241–249). PMLR.