Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

Using Causal Graph Model variable selection for BERT models Prediction of Patient Survival in a Clinical Text Discharge Dataset

Omachi Okolo *, B.Y Baha, and M.D Philemon

Department of Information Technology, Modibbo Adama University, Yola, Nigeria;
e-mail : omachi.okolo@gmail.com; bybaha@yahoo.com; dpmanliura@gmail.com
* Corresponding Author : Omachi Okolo

Abstract: Feature selection in most black-box machine learning algorithms, such as BERT, is based on the correlations between features and the target variable rather than causal relationships in the dataset. This makes their predictive power and decisions questionable because of their potential bias. This paper presents novel BERT models that learn from causal variables in a clinical discharge dataset. The causal-directed acyclic Graphs (DAG) identify input variables for patients' survival rate prediction and decisions. The core idea behind our model lies in the ability of the BERT-based model to learn from the causal DAG semi-synthetic dataset, enabling it to model the underlying causal structure accurately instead of the generic spurious correlations devoid of causation. The results from Causal DAG Conditional Independence Test (CIT) validation metrics showed that the conceptual assumptions of the causal DAG were supported, the Pearson correlation coefficient ranges between -1 and 1, the p-value was (>0.05), and the confidence interval of 95% and 25% were satisfied. We further mapped the semi-synthetic dataset that evolved from the Causal DAG to three BERT models. Two metrics, prediction accuracy, and AUC score, were used to compare the performance of the BERT models. The accuracy of the BERT models showed that the regular BERT has a performance of 96%, while Clinical-BERT performance was 90%, and Clinical-BERT-Discharge-summary was 92%. On the other hand, the AUC score for BERT was 79%, ClinicalBERT was 77%, while ClinicalBERT-discharge summary was 84%. Our experiments on the synthetic dataset for the patient's survival rate from the causal DAG datasets demonstrate high predictive performance and explainable input variables for human understanding to justify prediction.

Keywords: BERT prediction; BERT prediction comparison; Causal DAG; Clinical text analysis; Predictor selection.

Received: January, 2nd 2025
Revised: February, 26th 2025
Accepted: March, 3rd 2025
Published: March, 10th 2025
Curr. Ver.: March, 10th 2025

1. Introduction

Feature selection is a critical data preprocessing step in data analytics and machine learning (ML). Most ML algorithms select predictors based on the correlations between features and the target variable rather than the underlying causal relationships. The traditional variable selection methods, such as annotation, filter, embedding, and wrapper methods, are out of sync with the ever-increasing electronic data [1]. Moreover, [2] revealed that the knowledge about the causal relationships between predictors and the target variable has potential benefits for building explainable and accurate predictive models due to the ability of causal discovery to extract the underlying assumptions from the dataset. While causal variable selection has dominated quantitative data analysis, study [3] stressed that its emergence in the nexus of ML and natural language processing (NLP) is new. The delay in coalescing these fields could be attributable to the challenge of combining the high-dimensional textual dataset with causal inference methods, which have been mostly quantitative[4]. Despite the difficulty in combining these two fields, it has become more compelling and advantageous to fix a

common issue with deep learning-inspired black box models like Bidirectional Encoder Representation from Transformers (BERT), which is that they lack transparency and explainable decision [5]. These black-box models are also impaired by the legacy of deep learning (DL) models, which establish correlations between variables and prediction rather than causation, thereby leading to spurious correlations [6]. Therefore, relying on ML models whose variable selection method and decision logic are not transparent is risky.

Studies [2], [3] hinted that introducing causal perspectives to modeling variables in textual data can help mitigate the spurious correlation issues in traditional algorithms and build explainable models. Recent studies in causal ML accorded so much importance to the knowledge of the causal relationship between predictors and labels, improving variable selection, potential confounders, avoidance of bias, and predictive accuracies in NLP tasks[6], [7]. Building causal relations using causal graphs from medical texts can be very important to medical science. It can help identify novel and interesting causal observations from clinical notes, which can help to understand patients' health better. It can also help with clinical diagnosis and determine their prevention and treatment. Research [8] stressed the importance of causal knowledge discovery in the medical diagnostic process, including improving the accuracy of diagnosis, helping to interpret the causal relations in diagnosis, and selecting intervention strategies for a particular disease. Given the critical importance of Causal knowledge discovery (CKD) in clinical decision-making systems, there is traction in research aimed at incorporating medical causal knowledge into clinical decision-support systems by combining related tools such as causal inference, ML, and large language models [2], [9].

Despite the progress made in this research domain, the integration of causal inference with most DL approaches faces notable issues. One of the limitations is the inability to concurrently model high-dimensional relationships with embedded causal knowledge in the datasets. This problem spiraled from the lack of natural mechanisms in ML and DL to explicitly integrate causal knowledge into the learning process [10], [11]. Therefore, in most cases, deep learning-based algorithms such as BERT predict high-dimensional data such as textual clinical datasets with spurious correlations.

Our study proposes a novel approach that leverages the causal power of causal DAG for variable selection while mapping the identified causal variables into the deep learning-based BERT models for prediction to mitigate these limitations. Our method validates the causal DAG assumptions using conditional independence test (CIT) criteria. Here, we present the patients' survival rate as the binary target variable and other relevant causal variables extracted from the semi-structured clinical notes. The causal variables are diseases, treatment, confounders, and the target variable. This method allows for seamless integration of these causal DAG models into the BERT models for prediction. Our research bridges the gap between cutting-edge innovative causal discovery methods and deep learning-based models such as BERT models, offering a more comprehensive approach for causal analysis and prediction in high-dimensional clinical text datasets. The key contributions of this paper are:

1. Perform knowledge discovery from clinical note discharge text
2. Design a causal DAG structure from the discovered knowledge
3. Validate the Causal DAG structure with CIT criteria
4. Predict the semi-synthetic text variables obtained from Causal DAG using regular BERT Models, and its medical variants such as Clinical-BERT and Clinical-BERT discharge summaries.

The rest of the paper is presented as follows: section 2 reviews related works to this study. In section 3, we discuss the material and the methods adopted in this study, while in section 4, we implement our experiment and present the results. Section 5 concludes the work and suggests future work

2. Literature Review

This section reviews the related research at the intersection of causal inference and ML predictions. We then show how our study differs from existing studies. We later establish some causal preliminaries for causal discovery used in this study.

2.1. Causal Inference and Machine Learning Predictions

Causal inference and discovery have continued to impact the general artificial intelligence domain, as [12] stressed that causality is one of the most challenging and open issues in

ML and artificial intelligence. A significant issue on the import of causal structure in prediction modeling was expounded by [9], where a study on DAG and causal thinking in clinical risk prediction submitted that DAG could be used to model a priori causal assumptions to inform variable selection strategies for answering causal questions. Furthermore, they concluded that using DAGs to identify Markov Blanket variables may be a useful, efficient strategy to select predictors in clinical risk prediction models provided strong knowledge of the underlying causal structure can be extracted from the data generating process. Research by [12] focused on how to encode and select appropriate sets of attributes in a clinical text to optimize the results of ML models. They suggested that modeling the Naïve Bayes model with varying sets of attributes showed that extracting the appropriate attributes to be coded (such as diseases, procedures, aggravating factors, etc.) can improve algorithm prediction accuracy.

Moreover, recent research has examined the intersection of NLP and causal modeling, which is called causal NLP [7]. Some of these researches establish the possibility of extracting causal variables from natural language and structured datasets using causal structure and Markov assumptions for onward prediction by ML algorithms.

Table 1. Summaries of related studies of causal learning as a basis for machine learning prediction.

Ref	Research Focus	Method	Evaluation Method
[9]	The use of Directed acyclic graphs in clinical risk prediction modeling	Incorporating causal knowledge into clinical risk prediction model using the Markov principle	Logistic regression model
[12]	Selecting ML features for semi-automatic ICD-9-CM encoding	Extracting clinical discharge notes with graph-assisted matching	Recall rate with Naïve Bayes Model
[6]	Using causality with ML to obviate the limitations of explainable model techniques for identifying predictive variables	Measuring ML predictions from causal structure using synthetic data	Linear Regression (LR), Random Forest (RF), and Neural Network.
[14]	Integrating causal model ontologies with LIME for ML explanations in educational admissions	The use of causal structure and LIME to extract admission criteria from an admission database	Gaussian Naïve Bayes, Decision Trees, and Logistic Regression.
[1]	Design and validation of a Causal Model that focused on educational datasets	The study designed and validated a causal graph from an educational dataset on the Strengthening Education in Northeast Nigeria (SENSE - EGRA) project.	The causal graph from the dataset was validated using the Conditional Independence Test (CIT).
[2]	Causality-based feature selection: Methods and evaluations	Causal-based variable selection using a synthetic and real-world dataset	The study used the CausalFS algorithm

The study by [2], [6], [13] reinforced the assertions that causal variable selection for onward prediction by ML helps build robust and explainable models. However, a study [1], [13] implemented a similar study that uses a student admission dataset and letter identification subtask. Their studies validated the causal DAG with a conditional independence test. However, our study uses causal DAG to design and validate the assumptions of the patient survival rate in the semi-structured clinical text dataset for correctness before mapping it to the BERT models for prediction.

2.2. The Combination of Causal Directed Acyclic Graph and BERT Models

In this study, we propose the possibility of deep learning-inspired models, such as the BERT model, and its medical variants, such as Clinical-BERT and Clinical-BERT discharge summary, which are inherently black-box models to build predictions from extracted causal relations from a clinical dataset to compare their performance on synthetic clinical causal dataset.

Table 2. Regular BERT model and its clinical variants.

Model	Training dataset	Specialty
BERT	Dataset from Book Corpus and Wikipedia	General natural language tasks
ClinicalBERT	Dataset from clinical text and electronic health records	Clinical text analysis
ClinicalBERT-discharge summary	Discharge summary from MIMIC II	Discharge summary analysis

Each BERT model performs differently on different tasks as shown in Table 2. While BERT is versatile for general natural language tasks, ClinicalBERT performs better on electronic clinical text analysis. In contrast, ClinicalBERT-discharge summaries are tuned for clinical applications in discharge summaries [14], offering excellent performance on discharge text-related tasks [15]. Therefore, we generate causal variables from the Causal Directed Acyclic (DAGs) graphs for the eventual prediction of patient survival in a clinical text to showcase the concept of statistical independence and probability and how they can help to extract relevant variables for prediction [6]. The causal concepts used are explained in the next subsection.

2.3. Causal Preliminaries: The Concept of Independence

We adopt the causal formalism of independence, and we show that two variables, X and Y, are independent when X does not change Y, and the reverse is the case. In terms of statistical probability distributions, this is represented as follows:

$$P\,(Y) \;=\; P\,(Y \mid X) \tag{1}$$

$$P\,(X) \;=\; P\,(X \mid Y) \tag{2}$$

This is expressed as the probability of Y happening as the conditional probability of Y given X as in Equation (1). Conversely, the probability of X happening is expressed as the conditional probability of X given Y as in Equation (2). Therefore, this explains that the probability of X happening will not alter the existence of Y, and vice versa. This occurrence is referred to as statistical independence.

The general notation for independence uses the symbol ⫫. Using this symbol, we can state that X and Y are independent in the following way: $X \perp\!\!\!\perp Y$. The concept of conditional independence is critical in establishing causality variables. We can express that X and Y are conditionally independent given Z. This is causally represented as: $X \perp\!\!\!\perp Y \mid Z$.

Similarly, in terms of probabilities, we can express that: $P\,(X, Y \mid Z) = P\,(X \mid Z)\,P\,(Y \mid Z)$.

The above (X, Y, Z) is jointly factorized to give a product of two simple conditions of $(X \mid Z$ and $Y \mid Z)$ using the property introduced earlier $(P\,(X, Y) = P\,(X)\,P\,(Y))$.

2.4. Causal DAG for Causal Inference

Causal DAG formalism can map the conditional independencies in statistical expressions in a directed graph. A directed graph is denoted as $G = (X, E)$, comprises of joint distribution PX as a factorization of the variables $X = \{X1; \dots; Xc\}$ using c corresponding nodes or vertices $v \in V$ and connecting with the directed edges $(i;\, j) \in E$, where $(i;\, j)$ represent a directed edge between vi and vj. The two or more nodes and random variable (V), where $V = X1, X2 \dots Xn$, and the connecting edges are called (E). If all edges are directed without cycles, we refer to them as a class of graphs called DAGs.

We can illustrate that a parent paj as a vertex vi with child vj joined by a directed edge $Xi \rightarrow Xj$ such that $(i; j) \in E$ but $(j;\, i) \neq E$. The first parents are ancestors of the later descendants if there exists a directed path constituting $ik \rightarrow jk + 1$ for all k in a series of vertices in a DAG. Therefore, there exist three relations that can exist in a causal DAG. The first is parents who have a common child represented as $(ik \longrightarrow C \longleftarrow jk)$ is also called a collider or immorality. The second relation exists called a mediator or chain $(ik \longrightarrow C \longrightarrow$

$jk)$ where a parent node ik produces a child node C, which in turn produces another child jk where jk now becomes a grand descendant of ik. Lastly, the third relation exists where a node C, which is a parent, has two descendants ik and jk (expressed as $ik \longleftarrow C \longrightarrow jk$) is referred to as a fork or common cause con-founder. Therefore, these three relations (collider, chain, and fork) are ways that an observational dataset can be represented in a causal DAG to establish the basis for a causal graph and for determining relationships in the data-generating process [1], [6]. DAGs are expected to fulfill the Markov property so that the assumed joint distribution factorizes according to the repetitive decomposition amenable to Bayesian networks[6], as this:

$$P(X) = \prod_{i}^{d} P(Xi|Pai) \tag{3}$$

The Markov assumptions based on Bayesian networks in Equation (3) can be used as a condition or d-separation of the causal DAG structure. We used the backdoor operation for a common cause structure (Fork) in our causal DAG diagram.

3. Proposed Method

This study adopted a causal and ML research method that bears relevance to quantitative methods in similar studies such as [1], [6], [16]. This method discovers causal knowledge, builds the causal DAG structure, validates the causal DAG, and predicts the Patient survival rate from the Semi-synthetic dataset from our Causal DAG using the BERT and its medical variant models such as Clinical-BERT and Clinical-BERT discharge summary.

3.2. Dataset Preparations and Partitioning

This step involves loading the dataset, cleaning the text, selecting the variables of interest, encoding labels, and splitting the data into training, validation, and testing sets. The preprocessing steps transform the original dataset of 757805 records and 33 columns into the required data for the causal modeling and BERT predictions as follows:

1. Data cleaning and removing duplicate values and outliers reduced the dataset to 757123 records and 33 columns.
2. Data cleaning also involves removing missing values, which reduces the dataset further to 755051 records and 33 columns.
3. Feature engineering: A new variable called Survival rate was created from the existing "Patient _Disposition" columns to represent the Patients' Survival rate for causal discovery and modeling from the clinical discharge dataset. The dataset was reduced to 755051 records and 35 columns.
4. The causal discovery and modeling of the patient survival rate reduced the columns to 15000 records and seven columns as sample size. The choice of sample size was predicated by [6], [7], that a small sample size is not a uniquely causal problem since it may lead to statistical and algorithmic bias. The dataset change during the preprocessing is shown in Table 3.

Table 3. Changes in dataset size during preprocessing.

Processing steps	Number of records	Variables
Raw dataset	757805	33
Data Cleaning (Removing duplicates)	757123	33
Data Cleaning (Removing Missing values)	755051	33
Feature engineering	755051	35
Final Causal variables	15000	7

The preprocessing aimed to obtain the causal variables (15000, 7) that can be used to predict the survival rate in the textual clinical dataset. From the causal knowledge variables obtained, we divided the causal clinical text dataset into training, validation, and test sets in the ratio of 70%, 15%, and 15%, respectively [17], for BERT and its clinical variant prediction as in Table 4.

Table 4. Clinical dataset partitions for BERT models.

Models	Train sets	Validation sets	Test sets
BERT	10500	2250	2250
Clinical-BERT	10500	2250	2250
Clinical-BERT discharge summary	10500	2250	2250

3.3. Proposed Research Process Flow

The proposed process flow for the study entails the following steps and are depicted graphically in Figure 1 below:

1. Preprocessing the clinical discharge note datasets through data cleaning and grouping related terms.
2. Knowledge encoding is used to extract relevant variables from the clinical dataset that are relevant for the task of patient survival mining and for designing the causal DAG using the feature engineering and ablation technique to dispose of clinically irrelevant variables [18].
3. Design of Causal DAG from the Clinical discharge dataset.
4. Validate the Assumptions encoded in the causal DAG from the Clinical discharge test dataset using the CIT criteria.
5. Test if the causal assumption is established.
6. Predict the semi-synthetic from the Causal DAG using the BERT Model.

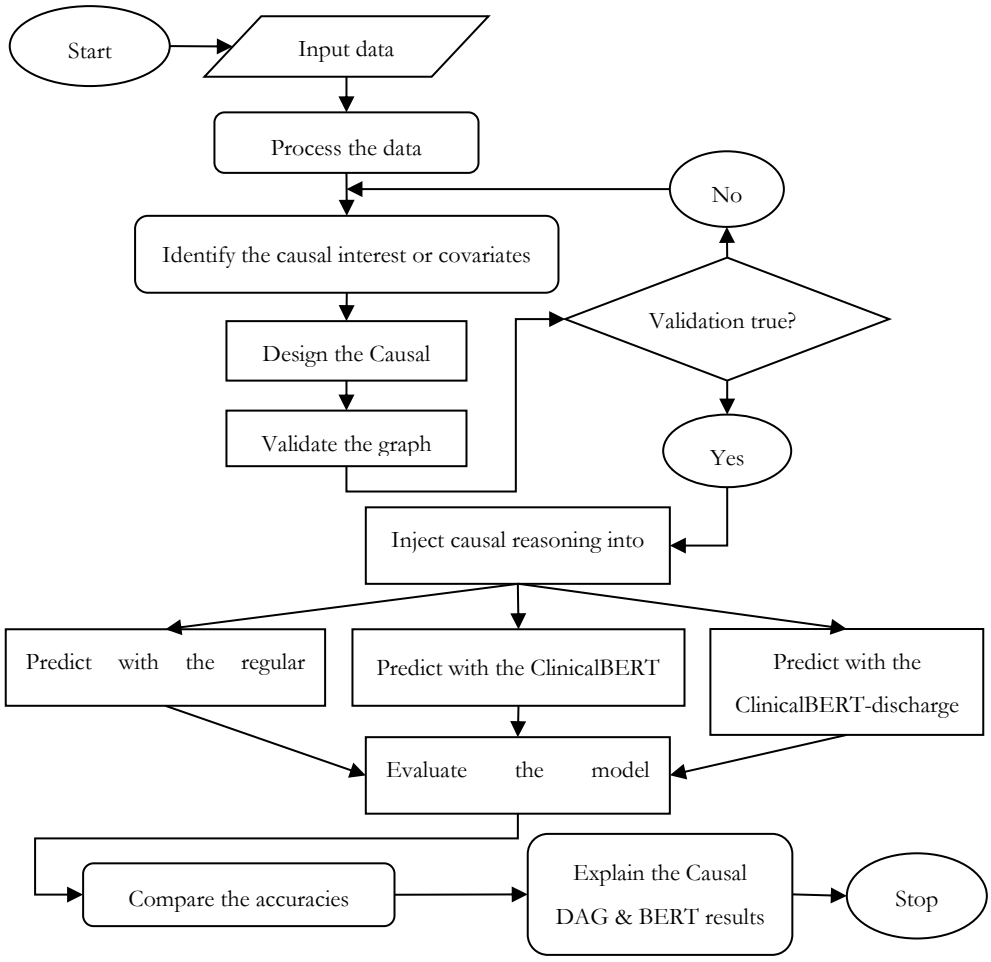


Figure 1. The Process workflow

3.4. The adopted BERT Models

We selected the three BERT models for predicting the causal DAG variable selected from the clinical discharge text: regular BERT and the two other medical variants of the

BERT models, such as the Clinical-BERT and the Clinical-BERT discharge summary. This study trains the semi-synthetic text from the causal DAG on the regular BERT, which uses a large corpus developed for the analysis of general domain text, and on Clinical BERT, which uses text from all clinical note types, and Clinical-BERT discharge summary, which uses only discharge summary to compare the performances of these three BERT models.

3.5. Model training architecture and hyper-parameter selection

The careful selection of the pre-trained BERT architecture and its Clinical variants, such as ClinicalBERT and ClinicalBERT-discharge summary, serves as the foundation for better fine-tuning and greater adaptation of the model to specific tasks of clinical text. While the size of the fine-tuning text dataset plays a secondary role[19]. The architecture of the three pre-trained BERT models and their tokenizer adopted for this study are listed below to eradicate ambiguity concerning the architecture and version of the BERT model adopted[7]. The models adopted in this study is presented in Table 5.

Table 5. Model architecture and hyper-parameter selection.

Model Architecture	Hyperparameters
BERT: BERT-Base (Un-Cased): 12-layer, 768-hidden-nodes, 12-attention-heads, 110M Bert = AutoModel.from_pretrained('bert-base-uncased') Tokenizer = BertTokenizerFast.from_pretrained('bert-base-uncased')	Optimizer: AdamW Learning rate: 1e-5 Epoch: 10
ClinicalBERT: Model = AutoModel.from_pretrained("emilyalsentzer/Bio_ClinicalBERT") Tokenizer = AutoTokenizer.from_pretrained("emilyalsentzer/Bio_ClinicalBERT")	
ClinicalBERT-discharge: Model = AutoModel.from_pretrained("emilyalsentzer/Bio_Discharge_Summary_BERT") Tokenizer = AutoTokenizer.from_pretrained("emilyalsentzer/Bio_Discharge_Summary_BERT")	

3.6. Oversampling to avoid overfitting and underfitting

We used a batch size of 32 and a maximum sequence length of 25 to pre-train the models. Similarly, gradient clipping was used to reduce exploding gradients associated with text tokens classification and gradient accumulation to enable splitting sample batches into smaller mini-batches (25 sequences of 512 tokens each) to optimize the utilization of GPU memory used in the Google Colab platform[20]. During training, the lower binary class (0) was oversampled to match the overrepresented (1) class with random state and SMOTE function to handle imbalance labels and achieve the balance of all classes under analysis. Moreover, we used a random state to implement cross-validation controls to shuffle the data before splitting to ensure that the different data values will result in different splits of the data, ensuring a robust model across different subsets.

3.7. Evaluation of the BERT Models

The BERT models were evaluated using metrics such as precision, recall, and accuracy of the test dataset were calculated. The formula for the evaluation can be seen as below:

- Precision is the ratio of actual survived patients classified by the model and all the clinical text classified by the model as survived. In terms of the true positives (TP) and false positives (FP), precision (p) can be formulated as Equation (4).
- The recall is the ratio of actual survived patients classified by the model and all survived patients in the dataset. In terms of the true positives (TP) and false negatives (FN), recall (r) can be formulated as Equation (5).
- F1-Score: A harmonic average of the precision (p) and recall scores (r), defined as Equation (6).
- Accuracy (acc) is the overall prediction of the target classes; see Equation (7).

$$p = \frac{TP}{TP + FP} \tag{4}$$

$$r = \frac{TP}{TP + FN} \tag{5}$$

$$f1 = 2\frac{p.r}{p + r} \tag{6}$$

$$acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Where TP = True positives, TN = True negatives, FP= False positives, FN = False negatives.

In addition, there is also the AUC-ROC metric, which is a probability evaluation metric that graphically describes the performance of a binary classifier in two forms:

- True Positive Rate (TPR), a sensitivity or recall, measures the proportion of actual positives correctly identified by the model. TPR can be calculated using Equation (5)
- False Positive Rate (FPR): This evaluates the proportion of actual negatives incorrectly identified as positives by the model. FPR is calculated using Equation (8).

$$FPR = \frac{FP}{FP + TN} \tag{8}$$

4. Implementation

This section designs and validates the causal DAG and predicts the patient survival rate using the BERT models. The implementation process involves several development tools. Google Colab was used for data preprocessing, feature engineering, and experiments with the BERT model. The Dagitty package was utilized for designing the Causal DAG ontological framework, as shown in Figure 4, and obtaining model coordinates and the CIT criteria assumptions in the dataset. Additionally, R programming was employed to implement the CIT criteria validation obtained from the Dagitty package.

The implementation follows a structured approach to ensure accurate modeling and prediction of patient survival rates. The following subsections detail each step of the implementation process, from knowledge discovery to model evaluation.

4.1 Causal Knowledge Discovery from Domain Knowledge and Feature Engineering

The target classes and the variables needed for the causal graph modeling were first identified with the aid of the domain knowledge of the clinical dataset through data exploration and feature engineering. The variables labeled Patient_Disposition in the notes/report category were found to be valuable in generating another important variable called Survival_rate. Regarding explaining clinical text column names, Patien_Disposition defines where a patient retires after being discharged from the hospital, as shown in Table 6.

From Table 6, it was discovered that the Expired label shows the number of patients who died after hospital admission. This column was important for identifying and modeling the number of patients who survived or died after being admitted to a particular hospital and administered some treatment. The label was encoded and separated into binary target labels as 'Survived' or 'Died', as Figure 2 illustrates the feature engineering process applied to the Patient Disposition column.

The knowledge discovery process produced the target class called the Survival_rate variable from the Patient Disposition variable. Thus, after identifying the target variable, the study identified the causal variables that could help model the causal relationship. At this stage, we discovered unnecessary categories of information in the dataset since they will not provide significant insight into the model target. Those data classes, such as Hospital Information, Billing Method, and Cost, were ablated or constrained. This was done in consultations with clinicians and medical domain specialists. The ablation technique and expert knowledge were needed to remove some columns that could introduce bias into the model [1], [9].

Furthermore, as shown in Table 7, we had left seven variables from the dataset needed for further modeling. The knowledge discovered from the clinical text was encoded to make an informed decision on the number of people that died or survived in the dataset. Therefore, the final variables selected for the model are Age_Group, Gender, APR_MDC_Description, Severity_of_Illness_Description, CCSR_Procedure_Description, and Survival_rate.

Table 6. Patient Disposition Categories.

Patient Disposition	Frequency
Home or self-care	482999
Home w/ Home Health Services	116579
Skilled Nursing Home	64578
Expired	25703
Left Against Medical Advice	20946
Short–term Hospital	12638
Inpatient Rehabilitation Facility	12413
Hospice - Home	4791
Psychiatric Hospital or Unit of Hosp	3921
Hospice – Medical Facility	3432
Another Typed Not Listed	3305
Facility W/ Custodial / Supportive Care	1947
Court / Law Enforcement	1410
Hosp Based Medicare Approved Swing Bed	1044
Medicare Cert Long Term Care Hospital	898
Cancer Center or Children's Hospital	632
Medicaid Cert Nursing Facility	327
Federal Healthcare Facility	179
Critical Access Hospital	58

```
df.loc[df['Patient_Disposition'] != 'Expired', 'Survival_rate'] = 'Survived'
df.loc[df['Patient_Disposition'] == 'Expired', 'Survival_rate'] = 'Died'

print(df[['Patient_Disposition', 'Survival_rate']].head())
```

Figure 2. Feature engineering on the Pateint_Disposition column

Table 7. The final variables selected and their descriptions.

Variable	Description
Age_Group (AG)	The age distribution of the patients
Gender	Sexual identities of patients
APR_MDC_Description	All patients refined of Major Diagnostic Categories (MDC) description
APR_Severity_of_Illness_Description	All patients refined (APR) the severity of the illness. It groups the severity of illness into four.
CCSR_Procedure_Description	Clinical Classifications Software Refined (CCSR) procedure description based on international classification of Diseases (ICD)
APR_Risk_of_Mortality	This groups the disease mortality level into four groups.
Patient_Disposition or Survival_rate	A place or setting where a patient was discharged to stay on the day of discharge.

The knowledge discovered from the clinical text dataset consists of [Diseases, Gender, Age, severity of the disease, mortality rate, treatment, and the target variable] as shown in Table 7. The causal DAG was depicted graphically, as shown in Figure 3. The causal DAG

structure can help to understand the specific variables extracted from the dataset that can be used to predict the survival rate in the dataset. This DAG structure harnessed the rich ontological framework of causal discovery to uncover the nature of relevant variables, relationships, and insights that can help in clinical decision-making in the dataset used in this study [21].

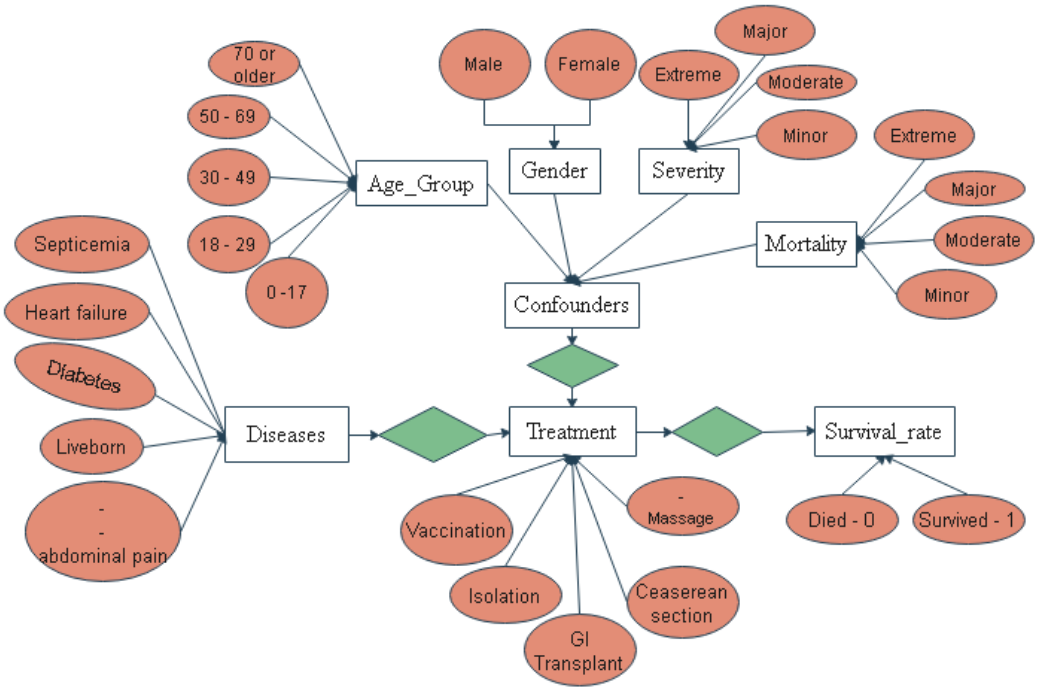


Figure 3. Patient Survival Rate Prediction Data flow using the Causal DAG Framework

4.2 Causal DAG Framework Knowledge Encoding and Formulation

Applying a Causal graph ontological framework can help test the statistical implications of the conceptual assumptions encoded in a given Causal diagram, and this can help researchers discover errors in the model, avoid erroneous conclusions based on spurious correlations, and build better models. Therefore, the causal graph model goes through systematic scrutiny and validation to ensure the correctness of the conclusions of a causal graph-based analysis and its underlying assumptions.

The causal knowledge of survival rate from the clinical text was encoded into the causal graph. We establish a causal assumption for the survival rate in the clinical text as part of the causal model to imply the following conditional independences and assumptions as Equation (9).

$$\text{Diseases} \perp \text{Survival\_rate} \mid \text{Confounders}, \text{Treatment} \tag{9}$$

The symbol "⊥" or "_||_"stands for independent of, and "|" stands for, given or conditioned on.

Therefore, Equation 10 can be interpreted as follows: [Diseases] is independent of [Survival_rate] given or conditioned on Confounders and Treatment. Simply put, that a person is sick does not equal surviving or dying unless you consider other factors such as confounding variables and the treatment administered. Where: $Confounders = \{Age\_Group, Gender, APR\_Risk\_of\_Mortality,$ and $APR\_Severity\_of\_illness\_Description\}$

However, from the Causal DAG in Figure 4, there was a biasing path that was opened by the [Confounders] that requires minimal adjustment sets for controlling or conditioning the information flow in the graph and for estimating the total effect of Diseases on Survival_rate given other variables. Therefore, conditioning or d-separation is necessary for blocking paths between the sets of nodes in a causal graph produced by confounders. Therefore, we adjusted or intervened on confounders using a backdoor adjustment operation to apply independence.

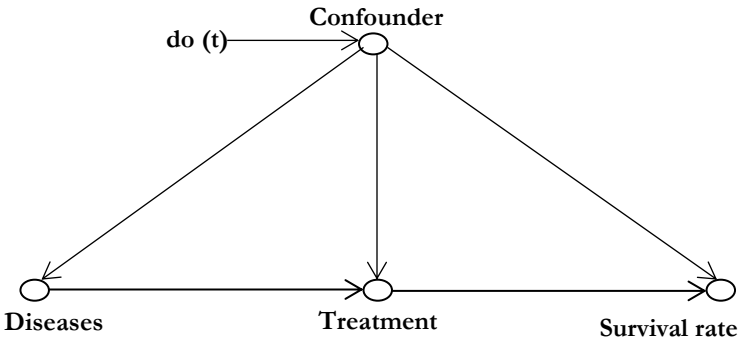


Figure 4. The Causal DAG structure and the adjustment on confounders

The adjustment concept or d-separation method applied to the causal graph is sufficient to identify the mathematical formula for adjusting covariates and estimating the causal impact of the intervention by using the do-action formula, i.e. $(y|(t))$. We perform an intervention on the $[Confounders - (y|(confounders))]$ as in Figure 4 to block the backdoor paths. Therefore, the d-separation on the confounders eliminates the confounding bias produced by the following equations, assumptions, or mathematical formulas for testing the encoded model using the CIT. This implies:

$$\text{Dsss } \_||\_ \text{ Srv\_ } | \text{ Gender, Trtm} \tag{10}$$

$$\text{Dsss } \_||\_ \text{ Srv\_ } | \text{ Ag\_G, Trtm} \tag{11}$$

$$\text{Dsss } \_||\_ \text{ Srv\_ } | \text{ S\_\_I, Trtm} \tag{12}$$

$$\text{Dsss } \_||\_ \text{ Srv\_ } | \text{ R\_\_M, Trtm} \tag{13}$$

Furthermore, the identified CIT assumptions above were used alongside the dataset to perform the CIT statistical test. The overarching objective of testing the causal diagram or knowledge discovered is to confirm or reject the CIT assumptions encoded and identified in the Causal graph ontological framework.

4.3. Causal DAG Model Validation and Results Explanation

The causal graph model encoded in Section 4.2 undergoes systematic validation to ensure the correctness of the Causal DAG assumptions. Figure 4 illustrates the encoded knowledge in the Causal DAG for the clinical discharge dataset's survival rate. The validation process verifies whether these assumptions hold using the Conditional Independence Test (CIT). The validation requires two key components: (i) the coordinates and CIT criteria derived from the design process in Digitty and (ii) the dataset. The CIT criteria obtained from the causal graph structure ontology, as presented in Equation (10), have been validated. Table 8 presents the results of the CIT test based on the assumptions from Equations (9)–(13), which were derived from the Causal Directed Acyclic Graph (DAG) framework in Figure 4. The study conditions on confounder variables {Age_Group, Gender, APR_Risk_of_Mortality, and APR_Severity_of_Illness_Description} to generate the causal assumptions for the CIT test.

Table 8. Results of the CIT criteria for each instance of confounders using Equations (9) – (13).

Confounders	CIT Criteria	LocalTest		95% Confidence Interval	
		p-coefficient	p.value	2.5%	97.5%
Gender	Dsss _| |_ Srv_ | Gndr, Trtm	-0.005141721	0.5289365	-0.02114432	0.01086351
Age_Group	Dsss _| |_ Srv_ | Ag_G, Trtm	-0.08660599	0.08177213	-0.1024691	-0.07069977
Severity	Dsss _| |_ Srv_ | S__I, Trtm	-0.01353112	0.09751101	-0.02952907	0.002473756
Mortality	Dsss _| |_ Srv_ | R__M, Trtm	-0.002604563	0.7497719	-0.01860813	0.01340034

The validation results in Table 8 indicate that the Pearson correlation coefficient estimates for all variables range between -1 and 1, and are close to zero, with a narrow confidence interval (CI) at 25% and 95%. Additionally, the p-values exceed the 0.05 threshold, confirming statistical independence [22], as indicated in Table 9.

Table 9. CIT criteria Metrics.

Metrics	Lower bound	Upper bound
Pearson correlation coefficient estimates	-1	1
p-value	>0.05	1
Confidence interval (CI)	25%	97%

The LocalTest results in Table 9 employ four key metrics to assess whether the conceptual assumptions of the causal graph hold: (1) Pearson correlation coefficient estimates, (2) p-value, and (3) confidence interval (CI) for the assumed conditional independence in the causal graph. The Pearson correlation coefficients for all variables range between -1 and 1, suggesting no strong correlation. The p-values (>0.05) indicate statistical significance, supporting the assumed independence in the causal structure.

The conditional independence assumptions are confirmed if the correlation coefficient is close to zero and the p-value is high (>0.05). Conversely, the causal structure may not hold in the dataset if the correlation coefficient is high and the p-value is low. The confidence intervals for the correlation coefficient should ideally be close to zero, as wider intervals indicate weaker validation of conditional independence assumptions. The Causal DAG ontology is validated if the CIT assumptions are confirmed. Otherwise, causal relationships or dataset adjustments may be required [1], [22].

The results confirm the assumptions proposed in Equations (9)–(13), derived from the causal diagram in Figure 4. This validation supports the conceptual Causal DAG model for patient survival prediction, ensuring its robustness in the clinical discharge dataset used in this study.

4.4. Using the Causal DAG Knowledge from the Clinical Text Dataset for BERT Predictions

The knowledge discovered from the clinical discharge text using the Causal DAG variables for the survival rate of patients at the hospitals was mapped into the BERT model. The Causal DAG variable was then used as a basis for BERT model predictors' variable selection in formulating a semi-synthetic dataset for clinical text classification [20], [23], [24]. Text classification is a core task in this section, which is amenable to the deep learning-inspired BERT model. We categorized text into predefined categories or classes. After re-arranging the clinical dataset from the discovered causal knowledge based on the predicted class of patient survival rate, the dataset consists of two columns: the 'text' and the 'label'. The column "text" contains the clinical text obtained from the causal graph ontological framework, and the "label" is a binary variable where 1 means the patient survived, while 0 means the patient died.

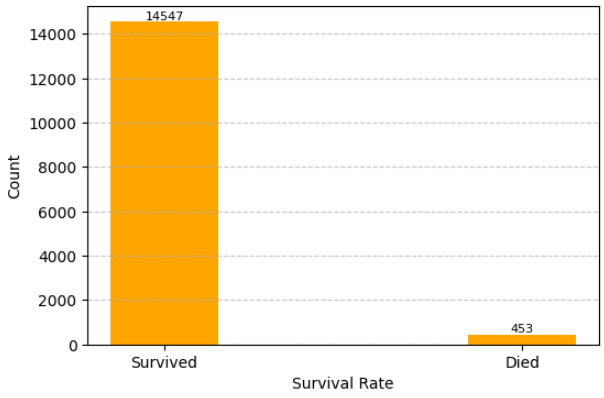


Figure 5. Patients' survival rate label distribution.

Table 10. Random samples of dataset text and labels from the clinical text.

ID	Text	Label
8871	DISEASES AND DISORDERS OF THE MUSCULOSKELETAL SYSTEM AND CONNECTIVE TISSUE-F-70 or Older-Moderate-Moderate…	1
3048	DISEASES AND DISORDERS OF THE SKIN, SUBCUTANEOUS TISSUE AND BREAST-M-30 to 49-Moderate-Minor-SPONTANEOUS V…	1
9033	PREGNANCY, CHILDBIRTH AND THE PUERPERIUM-F-30 to 49-Moderate-Minor-SPONTANEOUS VAGINAL DELIVERY…	1
4126	DISEASES AND DISORDERS OF THE MUSCULOSKELETAL SYSTEM AND CONNECTIVE TISSUE-M-50 to 69-Minor-Minor-KNEE ART…	1
9600	INFECTIOUS AND PARASITIC DISEASES (SYSTEMIC OR UNSPECIFIED SITES)-M-70 or Older-Extreme-Extreme-TOE AND MI…	1
8393	DISEASES AND DISORDERS OF THE RESPIRATORY SYSTEM-M-70 or Older-Extreme-Extreme-ADMINISTRATION OF THERAPEUT…	1
9622	DISEASES AND DISORDERS OF THE CIRCULATORY SYSTEM-M-70 or Older-Moderate-Major-ADMINISTRATION OF THERAPEUTI…	1
1232	DISEASES AND DISORDERS OF THE NERVOUS SYSTEM-F-30 to 49-Extreme-Extreme-COMPUTERIZED TOMOGRAPHY (CT) WITHO…	0
10823	DISEASES AND DISORDERS OF THE KIDNEY AND URINARY TRACT-M-70 or Older-Extreme-Major-ADMINISTRATION OF ANTIB…	0
7273	DISEASES AND DISORDERS OF THE CIRCULATORY SYSTEM-M-70 or Older-Major-Extreme-NON-INVASIVE VENTILATION…	0
13887	INFECTIOUS AND PARASITIC DISEASES (SYSTEMIC OR UNSPECIFIED SITES)-F-70 or Older-Extreme-Extreme-CHEST TUBE…	0
7268	INFECTIOUS AND PARASITIC DISEASES (SYSTEMIC OR UNSPECIFIED SITES)-F-50 to 69-Extreme-Extreme-MECHANICAL VE…	0

The dataset structure shown in Table 10 above comprised the text and their corresponding labels. The clinical text dataset fine-tuned on the BERT models was 15000 instances drawn from the entire dataset. The patients' survival rate prediction dataset contains 14547 patients who survived, while 453 patients died after undergoing a series of prescribed treatments, see Figure 5.

4.5. Tokenization of the semi-synthetic Clinical Discharge Dataset

Before feeding the clinical text data into the BERT models, we first converted it into a format that the model can process through a step known as tokenization. We used the Hugging Face AutoTokenizer from three BERT model variants to tokenize the text data. Example of Tokenization Process:

- **Original Text:** DISEASES AND DISORDERS OF THE RESPIRATORY SYSTEM-M-70 or Older-Major-Extreme-ISOLATION PROCEDURES
- **Tokenized Output:** ['diseases', 'and', 'disorders', 'of', 'the', 'respiratory', 'system', '-', 'm', '-', '70', 'or', 'older', '-', 'major', '-', 'extreme', '-', 'isolation', 'procedures']
- **Token IDs:** [7870, 1998, 10840, 1997, 1996, 16464, 2291, 1011, 1049, 1011, 3963, 2030, 3080, 1011, 2350, 1011, 6034, 1011, 12477, 8853]

To adapt the text for BERT processing, we included special classification tokens [CLS] and [SEP], which help the model differentiate the beginning and end of a sequence.

- **Original text with special tokens:** diseases and disorders of the respiratory system-m-70 or older-major-extreme-isolation procedures
- **processed input with special tokens:** tensor([101, 7870, 1998, 10840, 1997, 1996, 6091, 2291, 1011, 1042, 1011, 2753, 2000, 6353, 1011, 3576, 1011, 3576, 1011, 12477, 8853, 102, 0, 0, 0])

This tokenized representation ensures that the text is properly structured for input into the BERT model, enabling efficient processing and learning from clinical text data. The presence of padding tokens (zeros) accounts for fixed-length input requirements in BERT-based architectures.

4.6. Model Training Performance

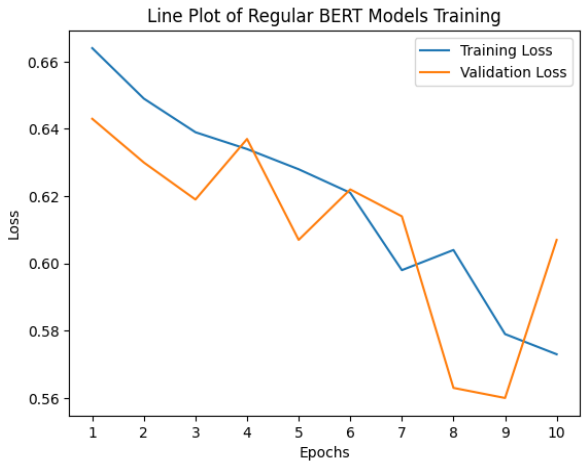
The variants of the BERT models adopted for this study, such as regular BERT, Clinical-BERT, and the Clinical-BERT discharge summary, were trained on the synthetic clinical dataset. The results of the training and visualization are shown below in Figure 6(a) and (b) for Regular BERT, Figure 7 (a) and (b) for Clinical-BERT, and Figure 8(a) and (b) for Clinical-BERT discharge summary.

4.7. Prediction Accuracy and Performance of different BERT models on Clinical Text Dataset

After the models had been trained on the clinical text classification task against the training and the validation datasets, the model was tested on the held-out test datasets. The accuracy of the three variants of the BERT models using the classification summary function is shown below in Table 11.

Epochs	Training Loss	Validation Loss
1	0.664	0.643
2	0.649	0.630
3	0.639	0.619
4	0.634	0.637
5	0.628	0.607
6	0.621	0.622
7	0.598	0.614
8	0.604	0.563
9	0.579	0.560
10	0.573	0.607

(a)

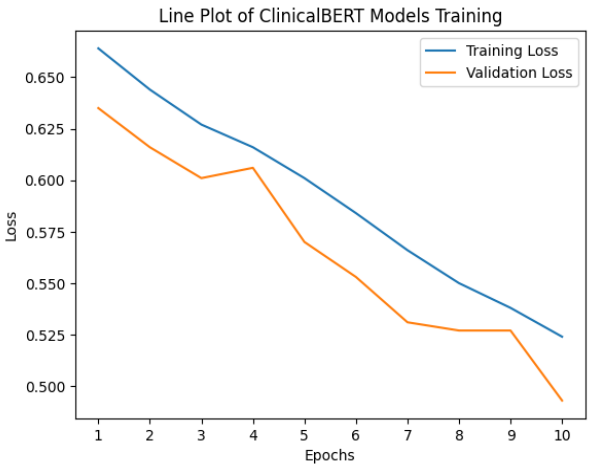


(b)

Figure 6. (**a**) Training and validation loss values across 10 epochs for the Regular BERT model; (**b**) Line plot showing the trend of training and validation loss during the model training process.

Epochs	Training Loss	Validation Loss
1	0.664	0.635
2	0.644	0.616
3	0.627	0.601
4	0.616	0.606
5	0.601	0.570
6	0.584	0.553
7	0.566	0.531
8	0.550	0.527
9	0.538	0.527
10	0.524	0.493

(a)

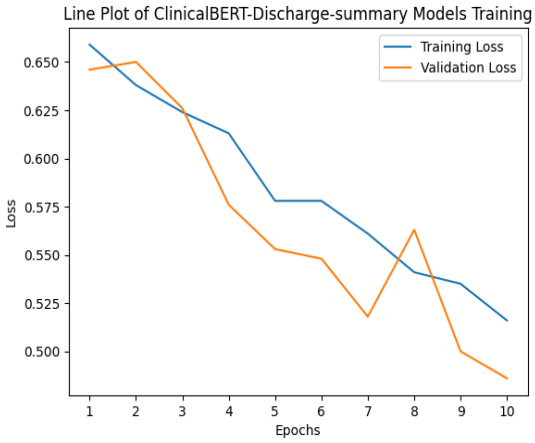


(b)

Figure 7. (**a**) Training and validation loss values across 10 epochs for the Clinical BERT model; (**b**) Line plot showing the trend of training and validation loss during the model training process.

Epochs	Training Loss	Validation Loss
1	0.659	0.646
2	0.638	0.650
3	0.624	0.626
4	0.613	0.576
5	0.578	0.553
6	0.578	0.548
7	0.561	0.518
8	0.541	0.563
9	0.535	0.500
10	0.516	0.486

(**a**)



(b)

Figure 8. (a) Training and validation loss values across 10 epochs for the Clinical BERT-discharge-summary model; (**b**) Line plot showing the trend of training and validation loss during the model training process.

Table 11. Performances of the three BERT models

Label/ Metrics	Regular BERT model				ClincalBERT Model				ClinicalBERT-Discharge-Summary			
	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Precision
0-label	0.39	0.54	0.45	68	0.17	0.60	0.26	68	0.21	0.60	0.31	68
1-label	0.99	0.97	0.98	2182	0.99	0.91	0.95	2182	0.99	0.93	0.96	2182
Acc			**0.96**	2250			**0.90**	2250			**0.92**	2250
Macro avg	0.69	0.76	0.72	2250	0.58	0.76	0.60	2250	0.60	0.77	0.63	2250
Weighted avg	0.97	0.96	0.96	2250	0.96	0.90	0.92	2250	0.96	0.92	0.94	2250

We used the classification report summary to evaluate the performance of the models as shown in Table 11 above. The regular BERT model achieved an accuracy of 96%, the clinical-BERT model achieved a performance of 90%, and the clinical-BERT discharge summary achieved an accuracy of 92% accuracy. The precision, recall, and f1-score performances between the survival [1] and death [0] labels were comparatively according to the number of classes in the test partition. Since the classes were imbalanced in favor of survived class, the performances were in that order. The regular BERT had the highest performance of 96%, followed by clinical-BERT discharge summary at 92%, while the clinical-BERT achieved the lowest accuracy of 90%. In addition, this study also uses the receiver operating characteristic curve to determine the probability of true positives and false negatives in predictions. The results of the ROC-AUC measurements are presented in Figure 9.

4.8. Results Discussion of Causal DAG and BERT Predictions

The causal DAG provides intuitive knowledge that appeals to the BERT models' human understanding and decision-making process [6]. Our Causal DAG structure in Figure 10 shows that a patient's [treatment] mediates between the diseases and survival rate. However, some unobserved variables called [Confounders – Age, Gender, disease severity, and mortality index of the disease] could impact the diseases, the treatment, and the survival rate of the patients as in Figure 10.

This causal relation assumption was tested and validated to provide the basis for variable selection for the BERT model in this study [6], [13]. This provides a novel approach [causal variables] for variable selection for black-box models such as BERT.

Our BERT model built on a causally tuned synthetic text dataset presents the promise of constraining the unnecessary text columns that add bias in the training dataset through ablation [25]. This was expressed by [26] in their study on causal structure discovery from electronic health records, which revealed that the proposed method achieved a higher recall and precision than the general-purpose methods. They asserted that the clinical text's causal re-

lationships help the model adapt to the dataset. The proposed method is more suitable for use in clinical decision support than the general-purpose method.

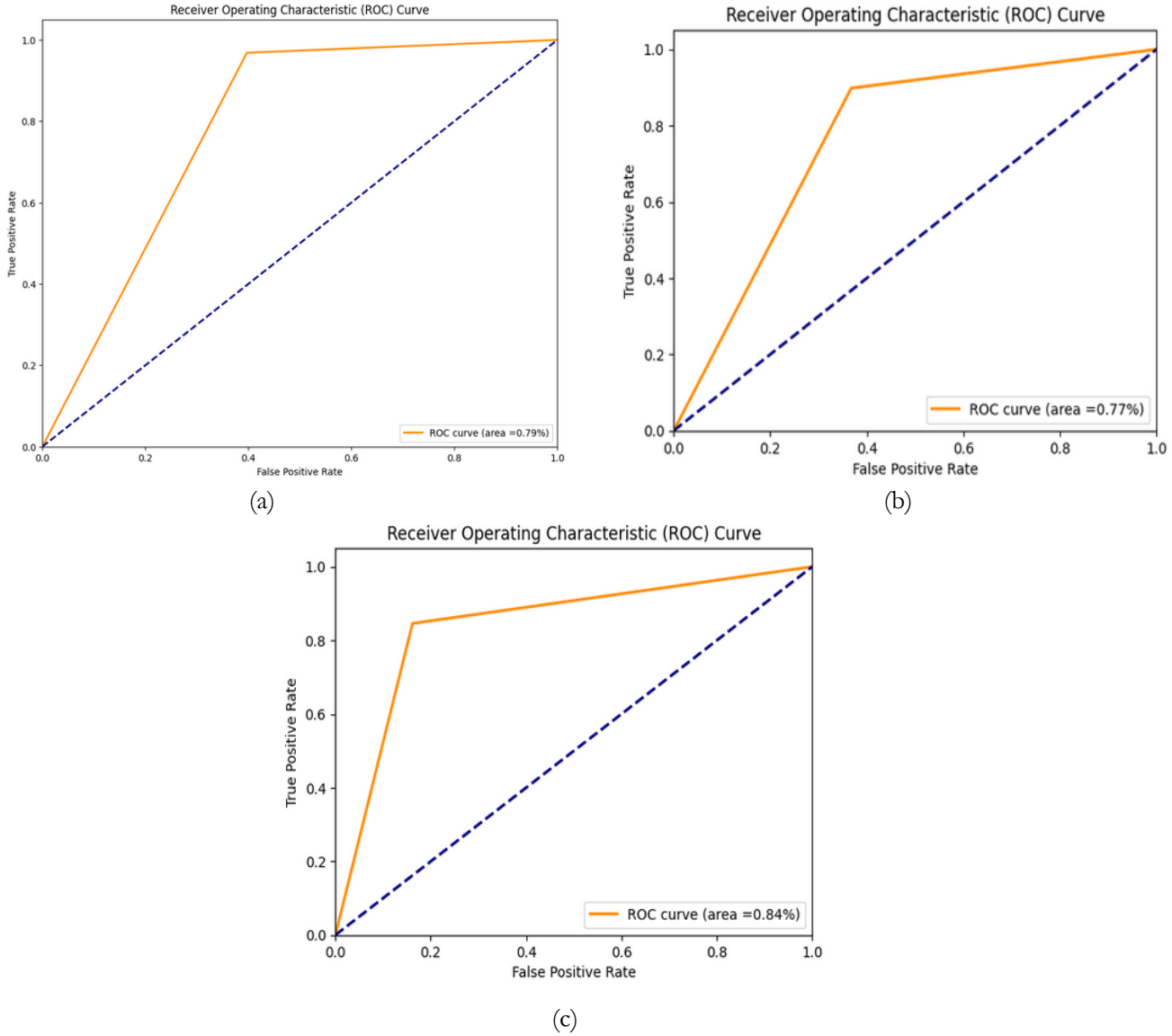


Figure 9. AUC-ROC Curve (**a**) BERT - AUC Score: 0.7856594058338275; (**b**) ClinicalBERT - AUC Score: 0.7653689006308299; (**c**) ClinicalBERT-discharge summary - AUC Score: 0.8421240631908127.

The results of the three variants of BERT Models with such as regular BERT, Clinical-BERT, and the Clinical-BERT-discharge-summary showed that the regular BERT had a performance accuracy of 96%, while Clinical-BERT performance was 90%, and Clinical-BERT-Discharge-summary was 92%. The regular BERT model pre-trained on a large and varied text domain from the Wikipedia corpus performed better than the two other domain-specific models such as Clinical-BERT and the Clinical-BERT-discharge-summary. Similar studies such as [15], [27] supported this result and offered that BERT in general, is superior to contextual embedding on a variety of tasks, including those in the clinical domains. A study by [15] added that BERT's superior performance is closely tied to its deeper and much more training parameters, thus possessing greater predictive power. More importantly, [study by [20] provided another reason for relative performance on the BERT model and added that the dataset on which a BERT model has been pre-trained could affect performance. Likewise, the Clinical-BERT and Clinical-BERT-discharge-summary used in this research were pre-trained MIMICII clinical notes and discharge text respectively while the clinical text in this study was extracted from the Statewide Planning and Research Co-

operative System (SPARCS) clinical discharge dataset. Hence, the regular BERT learned better than the two domain models. However, the Clinical-BERT-discharge-summary performed better than the Clinical-BERT model. Similarly, the results from the AUC in Figure 9 (a), (b), (c) showed that the ClinicalBERT-discharge summary performed better than the ClinicalBERT and the regular. ClinicalBERT-discharge summary has an AUC score of 84%, while BERT and ClincalBERT have an AUC score of 79% and 77%, respectively. Clincal-BERT-discharge summary AUC score superior performance was based on the fact that the dataset used in this study is a clinical discharge dataset. The Clinical-BERT-discharge-summary maximized the advantage of the domain power inherent in the pre-trained model to perform better than the Clinical-BERT.

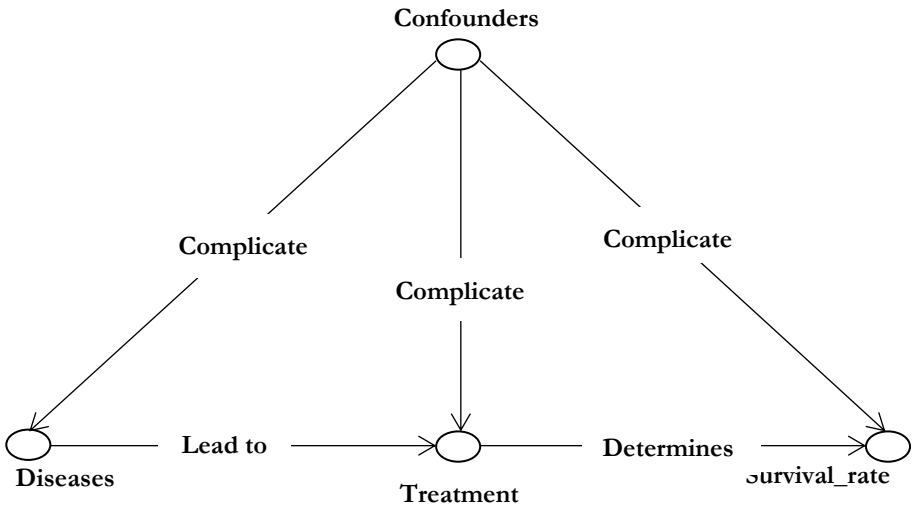


Figure 10. Causal DAG Explanations.

5. Conclusion and suggestions for future study

This study uses feature engineering and domain knowledge to discover patient survival knowledge in the clinical text. The causal knowledge encoded into the causal DAG was validated with the use of CIT and found to be substantiated by fulfilling the CIT metrics such as P-value, Pearson correlation, and confidence interval. The regular BERT model and its clinical variants, such as ClinicalBERT and ClinicalBERT-discharge summary, were used to predict the extracted causal knowledge. The accuracy of the BERT model performed better than the other two variants. However, the AUC score metric showed that the Clinical-BERT-discharge summary was superior because of the domain adaptation of the clinical discharge text used in this study. This study contributes to knowledge discovery in causal ML, where predictions are made on general datasets rather than causal variables, leading to spurious correlation. Therefore, reinforces the deficiency in ML prediction in the clinical domain and the widespread assertion that correlation is not causation as practiced in most classical ML predictions. Thus, adopting a causal DAG method to select causal variables in ML prediction is important. This method will help to reduce computation costs since the causal variables used are constrained or limited to important ones. This will also help explainable models from the model input perspectives and reduce the black-box scenarios that doubt the predictive power of opaque algorithms such as deep learning-inspired BERT models. However, since causal DAG may be flawed by design, this study suggests that other causal algorithm methods should be used to select causal variables from this dataset for ML predictions and improve causal ML development.

Author Contributions: Mr. Omachi Okolo conceptualized the research idea, analyzed the data, and drafted the first manuscript; Prof. B.Y Baha fine-tuned the methodology and the technicality; while Dr. M.D. Philemon reviewed and edited the manuscript. Prof B.Y Baha and Dr. MD Philemon both supervised the study. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive any external funding.

Data Availability Statement: The dataset used in this study is publicly available. Here is the link:
https://healthdata.gov/State/Hospital-Inpatient-Discharges-SPARCS-De-Identified/szqf-xu7c/about_data.

Acknowledgments: This study acknowledges the role played by healthdata.gov in making the dataset publicly available for research.

Conflicts of Interest: The authors declare no conflict of interest in this study.

References

[1] G. T. Ayem, O. Asilkan, and A. Iorliam, "Design and Validation of Structural Causal Model: A Focus on EGRA Dataset," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 86–103, Nov. 2023, doi: 10.33633/jcta.v1i2.9304.

[2] K. Yu *et al.*, "Causality-based Feature Selection," *ACM Comput. Surv.*, vol. 53, no. 5, pp. 1–36, Sep. 2021, doi: 10.1145/3409382.

[3] A. Feder, N. Oved, U. Shalit, and R. Reichart, "CausaLM: Causal Model Explanation Through Counterfactual Language Models," *Comput. Linguist.*, vol. 47, no. 2, pp. 333–386, 2021, doi: 10.1162/coli_a_00404.

[4] K. A. Keith, D. Jensen, and B. O'Connor, "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates," *ArXiv*. May 01, 2020. [Online]. Available: https://arxiv.org/abs/2005.00649

[5] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, Sep. 2019, doi: 10.1145/3236009.

[6] M. J. Vowels, "Trying to outrun causality with machine learning: Limitations of model explainability techniques for exploratory research.," *Psychol. Methods*, Sep. 2024, doi: 10.1037/met0000699.

[7] A. Molak and A. Jaokar, *Causal Inference and Discovery in Python: Unlock the secrets of modern causal machine learning with DoWhy, EconML, PyTorch and more*. Packt Publishing, 2023. [Online]. Available: http://ieeexplore.ieee.org/document/10251331

[8] K. Lyu and others, "Causal knowledge graph construction and evaluation for clinical decision support of diabetic nephropathy," *J. Biomed. Inform.*, vol. 139, p. 104298, 2023, doi: 10.1016/j.jbi.2023.104298.

[9] M. Piccininni, S. Konigorski, J. L. Rohmann, and T. Kurth, "Directed acyclic graphs and causal thinking in clinical risk prediction modeling," *BMC Med. Res. Methodol.*, vol. 20, no. 1, p. 179, Dec. 2020, doi: 10.1186/s12874-020-01058-z.

[10] M. Liu, D. R. Bellamy, and A. L. Beam, "DAG-aware Transformer for Causal Effect Estimation," *ArXiv*. Oct. 13, 2024. [Online]. Available: https://arxiv.org/abs/2410.10044

[11] J. Zhang, J. Jennings, A. Hilmkil, N. Pawlowski, C. Zhang, and C. Ma, "Towards Causal Foundation Model: on Duality between Causal Inference and Attention," *ArXiv*. Oct. 01, 2023. [Online]. Available: https://arxiv.org/abs/2310.00809

[12] J. Medori and C. Fairon, "Machine learning and features selection for semi-automatic ICD-9-CM encoding," in *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, 2010, pp. 84–89. [Online]. Available: https://aclanthology.org/W10-1113/

[13] B. I. Igoche, O. Matthew, P. Bednar, and A. Gegov, "Integrating Structural Causal Model Ontologies with LIME for Fair Machine Learning Explanations in Educational Admissions," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 65–85, Jun. 2024, doi: 10.62411/jcta.10501.

[14] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission," *ArXiv*. Apr. 10, 2019. [Online]. Available: http://arxiv.org/abs/1904.05342

[15] E. Alsentzer *et al.*, "Publicly available clinical BERT embeddings," in *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 2019, pp. 72–78. [Online]. Available: https://aclanthology.org/W19-1909/

[16] S. Gopalakrishnan, V. Z. Chen, W. Dou, G. Hahn-Powell, S. Nedunuri, and W. Zadrozny, "Text to Causal Knowledge Graph: A Framework to Synthesize Knowledge from Unstructured Business Texts into Causal Graphs," *Information*, vol. 14, no. 7, p. 367, Jun. 2023, doi: 10.3390/info14070367.

[17] S. Khanna, "A Comprehensive Guide to Train-Test-Validation Split in 2024," *Analytics Vidhya*, 2024. https://www.analyticsvidhya.com/back-channel/download-pdf.php?pid=134366&next=

[18] R. Pryzant, D. Card, D. Jurafsky, V. Veitch, and D. Sridhar, "Causal Effects of Linguistic Properties," *ArXiv*. Oct. 24, 2020. [Online]. Available: http://arxiv.org/abs/2010.12919

[19] LLM, *Large Language Models (LLMs) Interview Question*. Medium, 2024. [Online]. Available: https://masteringllm.medium.com/recent-11-large-language-models-llms-interview-questions-

[20] A. Turchin, S. Masharsky, and M. Zitnik, "Comparison of BERT implementations for natural language processing of narrative medical documents," *Informatics Med. Unlocked*, vol. 36, p. 101139, 2023, doi: 10.1016/j.imu.2022.101139.

[21] H. Alkattan, S. K. Towfek, and M. Y. Shams, "Tapping into Knowledge: Ontological Data Mining Approach for Detecting Cardiovascular Disease Risk Causes Among Diabetes Patients," *J. Artif. Intell. Metaheuristics*, vol. 4, no. 1, pp. 08–15, 2023, doi: 10.54216/JAIM.040101.

[22] A. Ankan, I. M. N. Wortel, and J. Textor, "Testing Graphical Causal Models Using the R Package 'dagitty,'" *Curr. Protoc.*, vol. 1, no. 2, Feb. 2021, doi: 10.1002/cpz1.45.

[23] A. S. Maiya, "CausalNLP: A Practical Toolkit for Causal Inference with Text," *ArXiv*. Computer Science - Computation and Language, Jun. 15, 2021. [Online]. Available: http://arxiv.org/abs/2106.08043

[24] V. Veitch, D. Sridhar, and D. M. Blei, "Adapting Text Embeddings for Causal Inference," in *Conference on Uncertainty in Artificial Intelligence, 2020*, May 2019, pp. 919–928. [Online]. Available: http://arxiv.org/abs/1905.12741

[25] S. Sheikholeslami, "Ablation Programming for Machine Learning," KTH-Royal Institute of Technology, 2019. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1349978/FULLTEXT01.pdf

[26] X. Shen, S. Ma, P. Vemuri, M. R. Castro, P. J. Caraballo, and G. J. Simon, "A novel method for causal structure discovery from EHR data and its application to type-2 diabetes mellitus," *Sci. Rep.*, vol. 11, no. 1, p. 21025, Oct. 2021, doi: 10.1038/s41598-021-99990-7.

[27] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An Attention Enhanced Graph Convolutional LSTM Network for Skeleton-Based Action Recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 1227–1236. doi: 10.1109/CVPR.2019.00132.