Causal Inference in Depression: Understanding Beyond Correlation

by

Syed Zamil Hasan Shoumo
21166008

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
October 2025.

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Syed Zamil Hasan Shoumo

21166008

Approval

The thesis/project titled "Causal Inference in Depression: Understanding Beyond Correlation" submitted by

1. Syed Zamil Hasan Shoumo (21166008)

Of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on October 19, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
BRAC University

Examiner:
(External)

Prof. Mohammad Zahidur Rahman

Professor
Department of Computer Science and Engineering
Jahangirnagar University, Savar, Dhaka

Examiner:
(Internal)

Dr. Muhammad Iqbal Hossain

Associate Professor
Department of Computer Science and Engineering
BRAC University

Thesis Coordinator:
(Member)

Dr. Md Sadek Ferdous

Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson, Associate Professor
Department of Computer Science and Engineering
BRAC University

Abstract

Depression remains one of the most pressing mental health concerns worldwide, intensified further by the socioeconomic and psychological impacts of the COVID-19 pandemic. Understanding the underlying mechanisms that contribute to depressive symptoms has therefore become a major research priority. While traditional statistical and machine learning models have been effective in identifying associations between risk factors and depression, they often fail to distinguish correlation from causation. Explainable Artificial Intelligence (XAI) methods, such as SHAP and LIME, have improved transparency by revealing which features influence model predictions; however, they remain fundamentally correlational and do not provide insight into the true causal pathways that drive depressive outcomes. To address this limitation, this research integrates machine learning, explainable AI, and causal inference to explore the causal factors behind depression among the Bangladeshi population during the COVID-19 pandemic. Using XGBoost for predictive modeling, the study first evaluates the relative importance of features through gain-based measures and SHAP value interpretation. Subsequently, a causal inference framework is constructed following Judea Pearl's principles to identify and estimate direct causal effects using the backdoor adjustment method with a generalized linear model estimator. Finally, a combined feature-selection pipeline is developed that retains causally significant variables and iteratively removes weakly correlated ones to test their joint predictive strength. The results reveal that while several factors exhibit high correlation and feature importance in black-box models, only a subset demonstrates genuine causal influence on depressive outcomes. This distinction underscores the importance of causal reasoning in mental health analytics. Overall, the study establishes that integrating causal inference within predictive frameworks not only enhances interpretability and trustworthiness but also provides a clearer understanding of which factors can truly influence and potentially mitigate depression.

Keywords: Mental Health, COVID-19, Depression, Causal Inference, Causal AI, Machine Learning, Explainable AI

Dedication

To my mother, whose strength remained unshaken. Even in her lowest moments, she held us high. Her sacrifice shaped me into who I am today, and for that, I am forever grateful.

Acknowledgement

Firstly, all praise to the Almighty, without whose mercy none of this would have been possible. Secondly, to my respected supervisor Prof. Dr. Md. Golam Rabiul Alam, for his unwaivering patience and support. And finally, to our parents for their unconditional love and guidance, because of who I am the man I am today.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Motivation

The COVID-19 pandemic has had a profound and lasting impact worldwide, influencing nearly every aspect of social, economic, and political life [22], [23], [29]. Lockdowns and restrictions led to unemployment, trade disruptions, and interruptions in education, among other consequences [32]. Communities across nations faced rising levels of social isolation, psychological stress, anxiety, and depression. Studies such as [21], [24] have reported both short- and long-term adverse psychological effects among children, youth, and the elderly. Emotional distress, uncertainty, and fear of infection have all contributed to widespread anxiety, stress, and—most prominently—depressive symptoms. This study focuses on the latter.

Depressive Disorder, or depression, is one of the most prevalent mental health conditions globally. According to the World Health Organization (WHO), nearly 332 million people are affected by it. It can emerge from an interplay of psychological, social, and biological factors. Individuals diagnosed with clinical depression often experience persistent sadness, loss of interest, and, in severe cases, suicidal ideation. Prior studies have shown that individuals exposed to challenging, stressful, or adverse life events are more likely to develop depressive disorder [57]. Conditions such as low income, unemployment, and early-life adversity have been strongly linked to depression, as have significant life challenges like marital separation or chronic illness. Although these factors are widely recognized, the mechanisms by which they interact remain uncertain. For example, while low socioeconomic status is associated with a higher prevalence of depression, it is unclear whether financial hardship directly causes depression or whether depressive symptoms increase the likelihood of economic decline.

Much of the existing research has relied on identifying correlations between depressive disorder and potential risk factors using observational data. These studies have offered valuable insights, yet correlation cannot always imply causation. Without a causal framework, it remains difficult to discern which variables genuinely increase the likelihood of depression and which are secondary or spurious. For instance, both [37], [44] reported associations between increased screen time and depressive symptoms among adolescents—particularly during the COVID-19 pandemic. However, it is uncertain whether increased screen time leads to depression or whether adolescents

with depression tend to spend more time online. Similarly, prolonged social isolation has been linked to depressive symptoms, yet individuals already suffering from depression may self-isolate more frequently. A purely correlation-based approach cannot distinguish between these competing explanations. To design effective policy, prevention, or intervention strategies, it is essential to establish cause–and–effect relationships rather than surface-level associations.

Machine learning (ML) models have emerged as powerful tools for prediction and pattern recognition across diverse domains, including mental health. Previous research [33], [35], [36], [48] has demonstrated that multiple factors can be used to predict the likelihood of depressive symptoms, particularly in the context of Bangladeshi students. With the advent of Explainable Artificial Intelligence (XAI), it has become possible to interpret such predictive models and visualize how features contribute to their decisions. Frameworks such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) allow researchers to identify which variables most strongly influence a model's predictions [40], [49], [51]. These tools are valuable for understanding feature associations, yet they fall short in addressing a more fundamental question: which features truly drive the outcome? Most XAI methods are correlation-based—they reveal what the model attends to, but not what it should attend to. For example, a model may highlight "screen time" as a significant predictor, yet this importance may result from hidden confounding factors such as social isolation. Acting on such correlational insights could therefore lead to misguided or ineffective interventions.

Causal inference and Causal AI directly confront these limitations by shifting focus from mere associations to the underlying mechanisms that generate them. Rather than asking which factors correlate with depression, causal methods investigate which factors, if altered, would meaningfully change the outcome. Understanding these relationships enables effective targeted interventions and a more accurate representation of mental health dynamics. Through approaches such as do-calculus, instrumental variables, propensity score matching, and causal discovery algorithms, researchers can estimate and validate causal effects even in observational settings. Graphical causal models, which represent relationships among variables, provide a formal structure for testing hypotheses while accounting for confounders, mediators, and moderators.

In light of these challenges and opportunities, this study aims to explore and analyze the causal effects of various factors contributing to depression among the Bangladeshi population during the COVID-19 pandemic. Rather than focusing solely on associations, this work seeks to identify which factors genuinely contribute to mental health deterioration when confounding influences are controlled. Ultimately, this thesis argues that understanding causal mechanisms behind depression is as crucial as identifying associations. At a policy level, evidence of causal effects can inform resource allocation and guide interventions addressing root causes—such as social isolation, anxiety, or economic vulnerability—thereby enabling more effective responses during future public health crises.

1.2 Research Objectives

Therefore, this research aims to achieve the following objectives:

- Identify the key socio-demographic, psychological, and environmental factors that exert a causal influence on depressive disorders among the population of Bangladesh during the COVID-19 pandemic.
- Differentiate correlation-based associations from genuine causal relationships, ensuring that the findings reflect mechanisms rather than surface-level patterns.
- Employ machine-learning models to predict depressive tendencies from survey data and examine their internal decision patterns through established XAI frameworks.
- Assess whether these XAI explanations align with causally validated features, thereby evaluating the reliability and interpretability of the predictive models.
- Address the limitations of conventional XAI methods, which emphasize feature importance without accounting for intervention-driven insights.
- Demonstrate how integrating causal inference with XAI can move beyond spurious association to identify factors that—if altered—could mitigate depressive symptoms.
- Contribute a framework for causal explainability that can guide future mental health research and data-driven policy formulation in low- and middle-income contexts such as Bangladesh.

1.3 Paper Outline

The rest of this paper is divided into a background study in chapter two, proposed methodology in chapter three, result analysis in chapter four, and a conclusion in chapter five, respectively.

Chapter 2

Literature Review

2.1 Relevant Works

This study aims to combine machine learning approaches with causal inference techniques to understand how different factors influence depression symptoms and how these factors relate to one another. Over recent years, many studies have shown that machine learning can effectively predict the likelihood of depression across different groups. To evaluate these models, researchers have used performance metrics such as precision, recall, F1-score, and AUC, often reporting strong predictive accuracy. In some cases, studies have also examined which factors the models consider important by using Explainable AI (XAI) tools such as LIME, SHAP, and Grad-CAM. These frameworks help to explain how black-box models reach their decisions and make their reasoning process more transparent. However, these XAI approaches mostly depend on correlation-based relationships. They reveal which features a model pays attention to, but they do not explain why those features matter or whether changing them would actually alter the outcome. Because of this limitation, many researchers are now turning to causal inference to uncover the deeper mechanisms that connect risk factors to depressive symptoms. Using Causal AI, it becomes possible to identify which variables have a genuine causal influence on depression and to examine how intervening on those factors might change the outcome. This approach shifts the focus from simply finding associations to understanding the actual causes behind them. Recent research has increasingly focused on applying these approaches to mental health analysis, particularly during and after the COVID-19 pandemic. Several studies have explored how social, behavioral, and environmental factors contribute to depressive symptoms using both traditional statistical methods and machine learning-based frameworks. A smaller but growing number of works have begun to integrate causal reasoning into these analyses, aiming to distinguish between mere associations and true causal influences. The following section reviews key contributions in these areas, highlighting the evolution from applying general correlation-based prediction models to identify the presence of depressive disorders to the application of XAI, and then finally leading up to causal pipelines in mental health research.

2.1.1 Existing Works based on Machine Learning

Health care and behavioral sciences are very popular domains where the inferential powers of machine learning models have been utilized to predict the presence of concerning psychological conditions. These models are often capable of capturing complex relationships among features that traditional statistical models usually overlook. For example, Vandana et al. (2023) had proposed a deep learning–based hybrid model for automatic depression detection in a multimodal setting [43]. A combination of both textual and audio data from the DAIC-WOZ dataset was used to train deep learning models such as convolutional neural networks (CNNs) and long short-term memory (LSTMs). Initially, the models were trained separately on the multi-modal data, followed by a hybrid structure integrating both. The audio CNN achieved the highest predictive score of 98%, outperforming the textual CNN (92%) and the hybrid models. The work demonstrated strong predictive performance. However, since the authors primarily focused on classification accuracy and model architecture, any explainability methods to interpret which features drove predictions were not applied. Haque et al. (2021) explored dimension reduction by applying a wrapper-based feature selection method using the Boruta algorithm and Random Forest on the Young Minds Matter dataset [31]. Their work continued with the implementation of several classifiers, including XGBoost, Decision Trees, Gaussian Naive Bayes, and Random Forest, to predict depression. Further analysis had shown that the Random Forest model performed best, reaching 95% accuracy and 99% precision in identifying depressed individuals. Their work identified key symptoms such as persistent unhappiness, loss of interest, sleep changes, and suicidal thoughts as significant predictors. However, the focus was mainly on optimizing predictive performance and comparing algorithms, rather than addressing feature importance beyond general association. These approaches reflect a stage of using deep learning in depression analysis, where the emphasis was on improving detection performance rather than understanding underlying factors or feature contributions.

While powerful machine learning models can predict linear and nonlinear patterns with high accuracy, their black box nature limits our understanding of their decision-making process. In sensitive fields such as mental health, this lack of interpretability raises various concerns regarding fairness and decision transparency. This necessity has given rise to the development of the field of Explainable Artificial Intelligence (XAI). Now, with the emergence of XAI, we can delve deeper into the models to gain a greater understanding of their decision-making, especially in critical domains like mental health research.

2.1.2 Machine Learning and Explainable AI (XAI)

The study done by Tian et al. (2023) focuses on assessing the likelihood of anxiety and depression among the Chinese population during COVID-19 [45]. The paper first employs machine learning models to predict the likelihood of these disorders and then seeks to identify the optimal combination of variables required to predict the outcome. They analysed feature relevance based on the feature importance provided by their applied tree-based models. Their results showcase that factors such as resilience, social support, being infected with COVID-19, or regular contact with

COVID-19 patients were key factors in predicting depression. The study, however, did not focus on the factors that are responsible for causing depression, but rather on those that can be used as good predictors for depression. Assessments were not conducted to determine whether the selected factors are causally associated or merely correlated with the outcome.

Magboo et al. explored the use of XAI to identify key features associated with depressive disorder. Using machine learning algorithms, the authors developed models capable of classifying depressive states based on behavioral and psychological indicators, while employing XAI techniques to interpret the contribution of each feature to the predictions [34]. The approach allowed the researchers to highlight significant factors correlated with depressive symptoms, providing better transparency for both clinicians and patients. The overall conclusion was that explainability tools could successfully reveal which input variables had the most decisive influence on the model's decision-making process. However, their analysis remained at a correlational level, relying on feature importance rather than causal reasoning. The study did not examine whether the highlighted factors actually lead to depression or merely appear alongside it, which marks a limitation and points to the need for causal inference methods to uncover actual causal mechanisms behind depression risk.

Jung et al. (2023) conducted a systematic review of XAI in healthcare, focusing on how the generated explanations have been applied and evaluated across different medical domains [41]. They selected six studies that used techniques such as permutation feature importance, LIME, SHAP, class activation maps, and pseudo-coloring to make the behaviour of black box models more interpretable. These approaches mainly focused on identifying which input variables had the most decisive influence on predictions, helping us better understand the behaviour of complex black-box models. This work highlights how XAI methods offer meaningful insight into feature importance compared to traditional machine learning, but also points out the current lack of structure and evaluation standards necessary for reliable clinical deployment.

Similarly, Band et al. had showcased a systematic review analyzing the role of XAI methods across medical domains, classifying primary interpretability techniques such as SHAP, LIME, Grad-CAM, and Layer-wise Relevance Propagation (LRP) and their applications in disease detection [38]. Drawing from 53 studies, the paper summarized how these XAI methods have been used in tasks like brain tumor segmentation, COVID-19 detection, and chronic kidney disease diagnosis. The authors found that XAI techniques improve model transparency and user trust by allowing physicians to visualize or quantify the contribution of clinical features to AI predictions. However, they noted that current healthcare XAI models still operate as descriptive tools—explaining how models make predictions rather than why specific patterns emerge. The study concluded that despite substantial progress in interpretability, XAI remains primarily correlation-based and cannot establish cause-and-effect relationships in clinical outcomes. Integrating causal inference frameworks was identified as a crucial next step for developing AI systems that move beyond interpretability toward causal understanding in medical decision-

making.

The study by Bodria et al. investigates the consistency and reliability of XAI methods across various modalities of data [39]. The paper compares and evaluates the stability and robustness of XAI frameworks such as SHAP, LIME, Counterfactual Explanations, and Rule-based Models. The findings reveal that current XAI methods, while informative, often produce inconsistent or even contradictory explanations for the same predictive model. This inconsistency highlights the trade-off between interpretability and accuracy, where more interpretable models tend to lose predictive strength. The study concludes by emphasizing the importance of developing explanation systems that go beyond correlation-based reasoning. It draws attention to the potential of causal reasoning frameworks as a pathway toward more trustworthy and actionable AI explanations.

From the studies described above, we can attest that although XAI frameworks have improved our ability to interpret complex machine learning models, their explanations remain primarily descriptive rather than causal. They reveal which features a model considers essential, but not whether those features truly influence the outcome. This limitation is a hindrance, especially for critical domains like mental health, where interventions depend on understanding not just naive associations but the underlying causes of psychological changes. To address this gap, recent research has turned toward Causal Inference and Causal AI, which provide a structured framework to understand cause-and-effect relationships from data. Unlike conventional XAI, causal methods focus on identifying the variables that can directly alter outcomes when manipulated, offering a more reliable structure for interpretation and informed decision-making.

The following section discusses recent studies that demonstrate how causal inference enhances interpretability in artificial intelligence, making model insights not only explainable but also actionable.

2.1.3 Stepping into Causality

One such paper by Rawal et al. (2025) explores how integrating causal inference methods can make AI systems more reliable and effective, particularly in areas where correlation-based models fall short [56]. Their study provides a survey of causal inference frameworks such as structural causal models (SCMs) and potential outcomes, showing how these enable AI models to answer counterfactual and interventional questions rather than just relying on surface-level associations. They argue that standard machine learning models often misinterpret statistical dependencies as causal effects, leading to biased feature importance and unreliable decision making, especially in sensitive applications like healthcare. Causal inference techniques, including propensity score methods, covariate balancing, instrumental variables, and causal discovery algorithms, help uncover hidden cause–and–effect relations and can adjust for confounders. This shift from correlation to causation enhances model explainability and lays essential groundwork for applying similar methods to mental health and depression analysis.

Furthermore, Naser et al. in [50] emphasize the limitations of relying purely on statistical regression or AI models that only capture correlations, especially in domains where understanding the data-generating process is essential. The paper continues to explain how causal inference techniques, using directed acyclic graphs (DAGs), can isolate mediators, confounders, and colliders to better understand the complex relationships that simple models might misinterpret. Naser highlights how methods such as score-based and constraint-based causal structure discovery help identify genuine causal links, while causal inference methods quantify the effects of interventions. This provides a deeper insight into feature importance and prevents misinterpretations that often arise from correlated but non-causal variables. These approaches, as a result, become critical for analyzing mental health data, where focusing on simply correlated factors can obscure fundamental causal drivers of depression.

Moreover, the recent study by Tao et al. similarly explains how AI approaches excel at finding associations but lack the tools to identify which features actually drive outcomes, especially in high-dimensional or confounded datasets [54]. The paper further discusses recent developments in combining causal discovery with deep learning, enabling models to separate spurious correlations from real causal effects. It outlines how causal graphs, do-calculus, and counterfactual reasoning provide a more robust way to interpret feature importance, allowing for more reliable insights in applied domains. This is particularly valuable for areas like depression analysis, where social, biological, and behavioral factors interact. By introducing causal principles into the model, researchers can target intervention strategies and produce better explanations that go beyond statistical association, leading to more trustworthy results.

While causal inference methods can be used to identify authentic cause-and-effect relationships from data, combining them with intelligent machines can create a much more advanced and complete framework. Unlike traditional AI models that rely purely on statistical correlations, Causal AI systems are designed to model the underlying data-generating processes and infer how changes in one variable may lead to changes in another. Now models can simulate interventions, counterfactuals, and hypothetical scenarios. In the context of depression analysis, causal inference combined with machine learning allows researchers to not only infer probable outcomes but also identify which factors, if altered, could actually change the outcome. These systems show much promise for psychological and public health research, as they provide interpretable and actionable insights that can guide policy development, clinical interventions, and preventive strategies.

2.2 Causality

The foundation of modern causal reasoning was introduced and popularized by the Turing Award-winning computer scientist Judea Pearl. In his revolutionary works, he introduced numerous mathematical frameworks for identifying, representing, and

quantifying cause-and-effect relationships [3], [6], [7], [9], [10], [19]. His development of Structural Causal Models (SCMs) and the do-calculus provided the backbone that allows distinguishing between mere correlations and genuine causal effects [16]. These tools enable us to model interventions, compute counterfactuals, and quantify how manipulating one variable might change another. This paradigm shift from association-based to intervention-based reasoning laid the groundwork for the emerging field now known as Causal AI. Causal AI combines the predictive power of machine learning with the interpretive capabilities of causal inference. Following his pioneering work, researchers have now extended his theoretical foundations into practical frameworks that can combine quantifying causality with artificial intelligence. These developments have led to the rise of Causal AI, where causal inference and causal reasoning can be embedded into the learning and evaluation steps of AI models. Frameworks such as DoWhy [27], [47], EconML [20], Causal learn [52], and CausalNex [30] have made it possible to estimate causal effects, simulate interventions, and represent causal structures directly from observational data. Causal discovery algorithms have enabled the inference of causal graphs in the form of DAGs from observational data. These extensions bring Pearl's vision of causality into the modern AI landscape, allowing machine learning systems not only to recognise patterns but also to reason about why a specific outcome occurs and how it can be changed. This is particularly important in domains like mental health, where understanding causal mechanisms is essential. For instance, while a traditional machine learning model may regard screen time as a good predictor for depressive symptoms, as increased screen time is associated with depression, a causal model would try to determine whether reducing or increasing screen time would actually cause a change in mental health. Researchers can now develop models that are not only accurate but also capable of determining whether they capture the correct associations present in the data. In this sense, Causal AI represents the next step in the evolution of intelligent systems—moving from predictive accuracy to causal understanding, a transition that directly aligns with the objectives of this research.

2.2.1 Intervention and the Do Operator

A prime concept of causal inference is intervention. In causal inference, an intervention refers to changing or manipulating a variable in a system to observe its effects on other variables. It allows us to infer causal relationships by forcing a variable into a particular state, rather than simply observing how variables behave under natural conditions. We have to iterate the fact that intervening on a variable is not the same as conditioning. Conditioning refers to observing the relationships between variables while we restrict our attention to specific scenarios, that is, to a subset of the entire sample space. It does not necessarily imply any active manipulation or control over the system. For example, we are studying the relationship between income and health, where income has two possible values, 0 (low) and 1 (high). Conditioning on income = 0 reduces the subset to the cases where the people have low income. It divides the already observed data into smaller sets.

On the other hand, intervention involves setting a variable to a specific value and imposing it on the entire population regardless of the factors that cause the variable

to change. Using the same example, if we intervene on income and set it to zero, we are practically modifying the data-generating process. Irrespective of the factors that cause income to change, we are setting it to a specific value in the system and then observing how the rest of the system behaves under this change. This is a key distinction because interventions can provide a clearer understanding of causal effects compared to simply conditioning.

In observational data, it's often challenging to distinguish between correlation and causation. Conditioning does not give us the power to directly manipulate the system or show what would happen under different conditions. In contrast, interventions help us to answer "what if" questions, for example, what would happen if we were to apply specific treatment. Additionally, questions about how it would exactly affect the other variables in the system can also be answered.

Figure 2.1: Pre and Post Interventional Graphs

The do operator (denoted as do(X=x) is a formal tool introduced by Judea Pearl to simulate the effect of an intervention in a causal graph. It allows us to model the causal effect of setting a variable X to a specific value x, while holding the rest of the system fixed. This changes the original graph to a modified structure by removing any incoming arrows to the variable being intervened upon. This severes any dependency between that variable and any of its causes. This modified graph is what we call a manipulated graph. This showcases that X is no longer influenced by its parents in the graph. When intervening, we eliminate the natural causal mechanisms that would otherwise dictate the value of the variable being intervened on. Here, the Figure 2.1 represents the causal graphs before and after the intervention. When we intervene on X, we remove the causal edge from Z to X. This graph is what is known as the manipulated graph.

2.2.2 Confounders and Colliders

A confounder is a variable that influences both the treatment and the outcome. If they are not accounted for, confounders can create spurious associations that can

mislead us about the true causal effect of the treatment. In the Figure 2.2, X is the treatment and Y is the outcome. We have a variable C that is a common cause for both X and Y. We observe that the association from X flows directly to Y and also via a backdoor path, X-C-Y. In this case, if we want to quantify the causal association from X to Y, we must account for the confounder C; otherwise, it will confuse the true causal association with additional spurious associations.

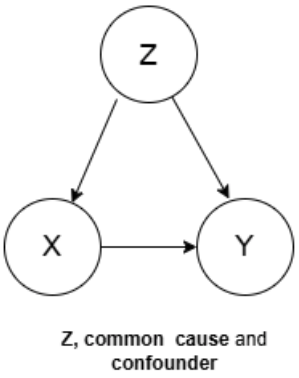

Figure 2.2: Confounders in a DAG

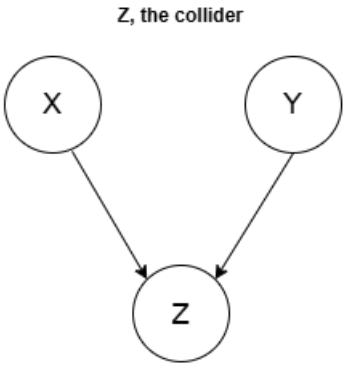

Figure 2.3: Colliders in a DAG

A collider is a variable that is caused by two or more variables. In the Figure 2.3, the variable Z is a collider on the path between X and Y. If we condition on a collider, which in this case is Z, the common effect of X and Y, we may introduce a spurious association between X and Y, even when no causal relationship exists. Initially, X and Y are independent of each other, but conditioning on Z will make them conditionally dependent and may introduce unnecessary association. This phenomenon is called collider bias, which we want to avoid.

Therefore, when we try to estimate the causal effect of one variable on another, we must carefully decide which variables to condition on. The goal is to isolate

the causal relationship from spurious association. Hence, we will condition on confounders to block backdoor paths and exclude colliders to prevent the introduction of new non-causal associations.

These aspects are fundamental when creating a causal inference pipeline. They lay the groundwork when we want to estimate causal effects and quantify causal mechanisms in the system. In the context of depression research, causal inference aligned with machine learning provides a better understanding of how different factors interact to determine mental health outcomes. Accordingly, the present study adopts machine learning and causal AI to identify causal drivers behind depression, addressing the limitations of purely correlation-based models.

The following section begins with a description of the dataset for the study, followed by the proposed model and the methodologies for deploying the experimental pipeline.

Chapter 3

Methodology

This chapter outlines the strategies adopted in this study to examine the causal relationships underlying depressive symptoms among individuals during the COVID-19 pandemic in Bangladesh. The approach combines machine learning, explainable AI, and causal inference techniques to move beyond predictive modeling and toward causal understanding. The methodology adopted is designed to ensure both predictive accuracy and interpretive ability, enabling the identification of features that not only correlate with depression but also causally influence it. The adopted methodologies are divided into two parts: (1) Machine Learning and XAI pipeline, and then (2) Causal AI pipeline. Utilizing the outputs from both pipelines, the paper then initiates an analysis regarding feature relevance and the prevailing causal mechanisms.

Figure 3.1: Machine Learning pipeline.

The ML pipeline (3.1) begins with data preprocessing to ensure it is acceptable to the machine learning and causal AI models. Additionally, on the refined data, some exploratory data analysis (EDA) was performed to gain some insights about the ongoing trends in the collected data. After this, we proceed to the application of machine learning models for depression prediction, followed by the integration of XAI tools like SHAP to interpret model outputs. This is where the machine learning pipeline ends. Following this, the paper introduces the Causal AI pipeline (3.2), which, based on the theory of causal inference, applies different Causal AI frameworks to discover the causal graph, create a causal model, identify the causal and statistical estimations, and finally estimate the causal effects.

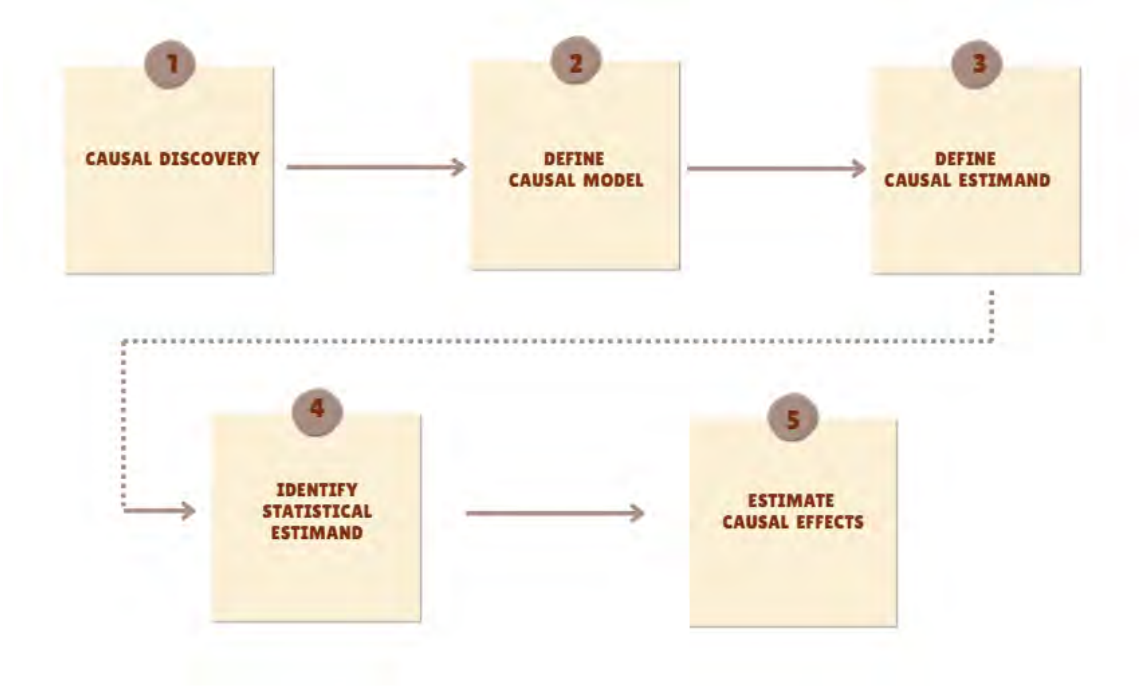

Figure 3.2: Causal AI pipeline

Together, these components form a comprehensive pipeline for understanding the causal mechanisms behind depression by using data-driven and interpretable AI methods, and light on the rthe relevance of characteristicscteristics for assessing depression.

3.1 Dataset Description

This study utilizes a secondary dataset originally collected by Pakpour et al. (2020) during the early phase of the COVID-19 pandemic in Bangladesh [26]. Their paper titled "A population-based nationwide dataset concerning the COVID-19 pandemic and serious psychological consequences in Bangladesh," contains responses from 10,067 participants across all 64 districts of the country. The data were gathered between April 1 and April 10, 2020, through an online cross-sectional survey conducted via social media platforms such as Facebook, WhatsApp, Twitter, and email. Participants aged above 10 years and residing in Bangladesh were eligible to participate. The dataset contains a total number of 73 features covering a wide range of factors, including:

- Socio-demographic characteristics such as age, gender, educational and occupational status, marital status, and area of residence.
- COVID-19 knowledge (awareness of symptoms, transmission, and preventive measures).
- Behavioral responses toward COVID-19 (preventive actions and lockdown practices).

- Psychological measures, including the Bangla Fear of COVID-19 Scale (FCV-19S), Bangla Insomnia Severity Index (ISI), and Bangla Patient Health Questionnaire (PHQ-9) for depression.
- Suicidal ideation related to COVID-19 assessed via a single binary (yes/no) question.

The dataset provides a multidimensional view of the population's mental health, COVID-19 knowledge, and behavioral patterns during the pandemic. The PHQ-9 scale, consisting of nine items rated on a four-point Likert scale, was used to measure depression symptoms. In contrast, the Fear of COVID-19 and Insomnia Severity scales captured the participants' anxiety and sleep difficulties, respectively [4], [53], [55]. The large sample size and nationwide coverage make this dataset particularly valuable for exploring mental health patterns during a national crisis. It also allows for examining how socio-demographic, behavioral, and psychological factors interact to influence depressive symptoms. As this research focuses on understanding the causal relationships behind depression, the dataset provides sufficient variables for building causal models and exploring interdependencies among factors such as fear, insomnia, social isolation, and economic hardship.

The authors obtained ethical approval for the original data. They ensured compliance with the 1975 Helsinki Declaration, as approved by the Biosafety, Biosecurity, and Ethical Committee of Jahangirnagar University and the Institute of Allergy and Clinical Immunology of Bangladesh. The participants were asked to submit written versions of their informed consents before taking part in the survey.

The authors obtained ethical approval for the original data. They ensured compliance with the 1975 Helsinki Declaration, as approved by the Biosafety, Biosecurity, and Ethical Committee of Jahangirnagar University and the Institute of Allergy and Clinical Immunology of Bangladesh. The participants submitted written informed consent prior to the survey. The dataset is publicly available on Harvard Dataverse (DOI: 10.7910/DVN/YKH9C1).

3.2 Machine Learning Pipeline

The following section describes the machine leanring pipeline that will be followed for our analysis. The pipeline starts with exploratory data analysis and preprocessing the data such that it as acceptable to our model. Then we apply an XGBoost classifier as our machine learning model. We will train our model on the data and then perform inference to predict the likelihood of depression for our test set. After this the pipeline concludes with applying SHAP to produce global predictions to understand which features are significant for our model. We will also use compute gain values for the features to understand which features are more suitable for creating splits or branches.

3.2.1 Data Preprocessing and EDA

Before applying any machine learning or causal inference techniques, it was necessary to prepare the dataset for analysis and ensure the quality and consistency of the collected information. The original dataset has been divided into two parts, with "Depression" regarded as the binary target variable and the remaining 72 features as the independent variables. The data, though comprehensive, contained several variables that required cleaning and encoding to make them suitable for statistical and computational analysis. This stage of preprocessing focused on handling missing values, removing irrelevant features, and removing duplicate rows where appropriate. Initially, as most of the features were present in Likert scales, we can treat them as ordinal variables. After the preprocessing steps, we now have a dataset with 73 variables and no missing or duplicate values. Following data cleaning, an exploratory data analysis (EDA) was conducted to gain an initial understanding of the dataset's structure, distributions, and underlying relationships. Because our methods revolve around correlation and causation, we computed Spearman's rank correlation coefficient (ρ) to measure monotonic associations among all variables. To reduce complexity, we will not be including the features that have low (positive or negative) correlation with the target. Therefore, based on this, we have selected a total of 16 ordinal variables for our experimentation. These 16 variables in a final sample size of 8950 showcase the highest amount of correlation with the target. They will be used as input for both the machine learning and causal inference frameworks.

3.2.2 Machine Learning Model

We will initally treat this as a machine learning problem and infer the likelihood of depressive symptoms based on our data. Machine learning provides the ability to detect complex, linear, and nonlinear relationships among features that traditional statistical models may overlook. However, since these models often function as black boxes, interpreting how they arrive at their predictions becomes challenging. To get a look under the hood, we will also evaluate the Gain values for the feature and rank them based on their values to understand which feature acts as a better criterion for creating splits.

Extreme Gradient Boosting (XGBoost):

For prediction, the machine learning model we will use is XGBoost. It is a tree-based model that utilizes boosting to create the final ensemble. As referred in the equation 3.2, the model implements a minimizing objective function where $\ell(y, \hat{y})$ is the differentiable loss measuring the difference between prediction and target, and $\Omega(f)$ is a regularization term on each tree. The loss term of the objective function can be the log-loss for classification problems. Furthermore, as given in equation 3.1, when deciding on the best criterion for splitting, the model uses the following quantities: G_L and G_R are the sums of the first-order gradients for the left and right child, respectively, and H_L and H_R are the sums of the second-order Hessians for the left and right child. Lastly, λ is the L2 regularization, and γ is the penalty

parameter. The split criterion is only accepted if the calculated gain is positive [14], [15].

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right) - \gamma \tag{3.1}$$

$$\mathcal{L}(\boldsymbol{\phi}) = \sum_i \ell\big(\hat{y}_i,\, y_i\big) + \sum_k \Omega(f_k) \tag{3.2}$$

The studies conducted in [28], [42], [45], [46] demonstrate that XGBoost can outperform other models when handling large-scale data for predicting mental health. The model has shown much promise in assessing mental health because of its predictive ability. Still, due to its complexity, it has low interpretability compared to other models such as logistic regression, Naive Bayes, or random forests. For our experiment, we have split our data in the ratio of 60:20:20 for training, testing, and validation. No feature scaling or label encoding was required, as all of the chosen features are in Likert scales.

3.2.3 Explainable AI

As the last step of the machine learning pipeline, to understand the relative importance of different factors, this study employs a combination of machine learning models and explainable AI (XAI) techniques. The higher the predicitve ability, the lower is the interpretibility. In order to better understand our model's internal reasoning, XAI frameworks such as SHAP will be integrated into the analysis to analyze the contribution of each feature to the model's output. This combination allows for accurate prediction of depression and a transparent understanding of the factors that most influence the results, forming an essential bridge between pure prediction and the reasoning of the model.

SHapley Additive exPlanations(SHAP):

Lundberg et al. first introduced the concept of SHAP as an XAI framework in their paper [17]. It is a model-agnostic algorithm that can be used to obtain model explanations at both the local and global levels. The equation 3.3 is the function SHAP aims to approximate. Assuming f(x) is the model's prediction for any instance x, SHAP will try to approximate this function as a linear combination of feature attributions.

$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i \tag{3.3}$$

Here, $f(x)$ represents the model's predicted probability for the sample x, ϕ_0 represents the baseline model prediction (average), and ϕ_i, the actual SHAP value, is the attribution or influence of feature i. M is the total number of features. This means the prediction for each sample equals what the model would predict on average, plus the net effect of each feature's influence. The baseline prediction is what the model

would predict with no information, and we add the sum of the contributions of each feature to it. The equation 3.4 showcases how the shap values are calculated.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(M - |S| - 1)!}{M!} \Big[f_{S \cup \{i\}}\big(x_{S \cup \{i\}}\big) - f_S(x_S) \Big] \tag{3.4}$$

Here, M represents the total number of features, S represents a subset of features not including the feature i whose influence we are calculating, $f_S(x)$ is the model's prediction only using the features in the subset S, and $f_{S \cup \{i\}}(x)$ is the model's prediction using the features in subset S and the feature i.

The term $\big[f_{S \cup \{i\}}(x) - f_S(x)\big]$ is the marginal influence of the feature i. This quantifies how much adding the feature i to the subset S changes the model's prediction for the sample x. The rest of the term before this is simply a weight factor to be multiplied when averaging the marginal contributions of feature i. The final SHAP value for the feature i shows how much the feature pushes the predicted probability away from the baseline or average. The SHAP values tell us why the predicted value is $f(x)$ instead of the baseline.

SHAP will be applied to our model predictions to gauge how much the features contribute to the final prediction. We will focus on global explanations rather than on local-level explanations. For a specific feature, we will first calculate the sum of the absolute values of the SHAP values generated for every sample of our test set, then average it. That will give us a mean contribution of the feature for the entire test set. This will be done for all the features in question to get an overall view of the feature importance assigned by our specified model. With respect to our target variable, *Depression*, for instance, a mean SHAP value of +0.25 for feature i indicates that, on average, knowing this feature changed the model's predicted probability by +0.25 compared to the baseline.

Therefore, with the help of SHAP, we can get a better understanding of which feature contributes to what extent in the model's prediction. But what we would still be missing out on is why a feature has such an influence on the prediction. Furthermore, whether changing the feature in the real world would cause the prediction to change is another question we should focus on. SHAP tells us why the model predicts the presence of depressive disorder, while we are more interested in what truly causes depression. SHAP may give high importance to a variable that is correlated with depression, but it may not be a direct cause of depression in the real world.

It is necessary to distinguish statistical association from genuine causation. We propose to achieve this by utilizing the scientific methods of causal inference in the form of a Causal AI framework.

3.3 Causal AI Pipeline

Causal inference builds the foundation of the second part of this research, guiding the process of identifying and estimating the genuine cause-and-effect relationships that drive depressive symptoms. Building upon the teachings of Judea Pearl, causal

inference can be used to derive causal relations, interventional distributions, and causal effects solely based on observation data [16], [19]. Pearl's framework, expressed through Structural Causal Models (SCMs) and Directed Acyclic Graphs (DAGs), provides a formal and structured way to represent causal relations present in the data. His framework of do-calculus offers a set of rules for estimating the effects of interventions, allowing researchers to investigate not just what is associated with depression, but also what causes it. Moving over, the field of Causal AI extends these principles into the world of machine learning. While traditional AI systems excel at prediction, they often rely on purely statistical correlations, limiting their ability to generalize across contexts or reason about interventions. Causal AI can be used to bridge this gap by integrating causal mechanisms into machine learning pipelines, allowing algorithms to learn causal structures, estimate treatment effects in high-dimensional data, and simulate counterfactual scenarios. This marriage of causal inference and AI enables models to capture causal mechanisms from high-dimensional observational data and establish a stronger and more interpretable foundation for inferential tasks.

3.3.1 Causal Discovery

The first step of the causal estimation pipeline starts with learning the causal structure directly from either domain knowledge or from observational data. This step is known as causal discovery. In traditional causal inference, the causal structure is visually represented in the form of a Directed Acyclic Graph (DAG). The DAG consists of features as nodes, with a directed edge between the nodes showcasing the cause-and-effect relationship. A directed edge from node A to node B means that A is the cause of B. Therefore, a change in the probability distribution of A will directly alter the probability distribution of B. This DAG can be drawn from domain knowledge, or in our case, from observational data. The causal discovery algorithms aim to learn the structure by taking advantage of the dependence and conditional independence among the variables. Leveraging these relations along with initial assumptions, they help to signify the potential cause-and-effect relationships in the data. In this way, a causal graph can be obtained for later use in causal effect estimation and counterfactual reasoning.

In addition, several methods have been developed for causal discovery, with each technique making different statistical assumptions and employing various strategies to approximate the causal structure. For example, two very popular choices are the PC (Peter-Clark) algorithm [1] and the Fast Causal Inference (FCI) algorithm [13]. They fall under the class of Constraint-based causal discovery methods that use conditional independence tests to infer causal relations. Another class of algorithms is the score-based causal discovery methods. The Greedy Equivalence Search (GES) falls under this umbrella. GES finds the causal graph that best fits the data according to a scoring function, which can be the BIC score [2] or the generalized score [18]. Another widely used method, LiNGAM (Linear Non-Gaussian Acyclic Model), assumes linear relationships and non-Gaussian noise to infer causal directions [5], [8], [11].

When determining which causal discovery method to select, it is necessary to focus

on the class of variables present in the data and the research objective. For our work, we propose to use the FCI algorithm to obtain the causal graph. Now, our dataset consists of multiple categorical variables in Likert scales, which represent psychological and behavioral attributes, along with a binary outcome variable indicating the presence or absence of depressive symptoms. As the variables are ordinal and non-continuous, the relationships among them may be nonlinear. Furthermore, another highly plausible scenario is the presence of confounding variables such as personality traits, life experiences, or environmental stressors that may affect both treatment and outcomes. The PC algorithm assumes that all relevant variables are observed and present in the data, and that no hidden confounders can influence the system. But In psychological or social datasets, this assumption may often break as many influential factors are unmeasured or unobservable. Thus, choosing the PC algorithm under the risk of unobserved causes may lead to a biased or incomplete causal graph, where spurious correlation links may appear. Moreover, LiNGAM assumes that the relationships between variables are linear and that the residual errors are non-Gaussian and independent. Since Likert-scale data are discrete, ordinal, and not normally distributed, the assumptions required by LiNGAM are not satisfied in this case. On the other hand, the FCI algorithm was specifically designed to handle scenarios where unobserved confounders may exist, and it does not impose any strict assumptions on linearity either. It uses conditional independence testing that can accommodate ordinal data, making it suitable if the experimental setup is similar to ours.

Fast Causal Inference (FCI) for Causal Discovery:

We choose to obtain our causal graph using the FCI algorithm. In order to find the causal relationships and orient the edges, FCI applied the following steps:

- The FCI algorithm starts with a complete undirected graph among all the variables. In this version of the graph, every variable is connected with every other via undirected edges.
- Then, for each pair of variables X and Y for the given data, FCI checks if they are conditionally independent given a subset of other variables Z. If they are found to be conditionally independent, the undirected edge between X and Y is removed. The conditioning sets Z are recorded as separation sets. After this step, a final undirected graph (also called a *skeleton* graph) remains that represents the conditional independencies among the observed variables.
- The next step focuses on detecting colliders. The algorithm looks for triplet connections like $A - B - C$, where A and C are not adjacent. If A and C are not adjacent, and if B is not in the separation set of A and C, then the triplets are oriented as $A \rightarrow B \leftarrow C$; otherwise, they remain unoriented. This is the first step where directionality is introduced in the graph.
- Once the initial colliders are oriented, the next stage focuses on orienting additional edges based on the propagation rules proposed by Christopher Meek in [12]. These rules focus on propagating the effects of the oriented edges to orient the undirected ones. The four rules are described as follows:

- **R1:** If $A \to B$, B and C share an undirected edge ($B - C$), and A and C are *not* adjacent, then we can orient $B - C$ as $B \to C$.
- **R2:** If there exists a directed path like $A \to \cdots \to B$, and $A - B$ is undirected, then to avoid cycles we orient the edge as $A \to B$.
- **R3:** If $A - B$, and $B - A$ is undirected, $B \to C$ exists, and B and C are not adjacent, then the undirected edge $A - B$ is oriented as $A \to B$.
- **R4:** $A - B$, $A - C$, and $B \to C$ exist, and A and C are adjacent. If $A - C$ is not oriented yet, then we can orient $A - B$ as $A \to B$.

- The final causal graph is what we call a *Partial Ancestral Graph (PAG)*. Each endpoint in a PAG edge can have one of three possible markings. A tail (—) indicates the node is a possible ancestor of the other. An arrowhead (>) indicates the node is not an ancestor of the other, and a circle (∘) represents an undetermined direction. In simpler terms, when the conditional independence structure suggests a relationship that cannot be explained by the observed data circle endpoints (∘). Lastly, bidirectional edges (↔) are introduced to denote possible unobserved confounding.

In addition, like all causal discovery algorithms, FCI also relies on a few key assumptions. For example, it assumes the Causal Markov Condition, meaning that each variable is independent of its non-effects given its parents (direct causes). It also assumes faithfulness, implying that all observed conditional independencies in the data are due to the underlying causal structure rather than the influence of spurious effects. Additionally, FCI assumes the causal graph to be acyclic, such that the causal structure can be represented as a DAG or its corresponding PAG. Under these assumptions, FCI can efficiently identify not only direct and indirect causal links but also highlight possible confounding relationships.

3.3.2 Define Causal Model: The Structural Causal Model

Once we obtain the causal graph (DAG) through causal discovery, the next step is to formally define a causal model that can be used to represent the causal mechanisms present in the data. Therefore, to formally reason about causality, we utilize the framework of Structural Causal Models (SCMs). SCMs provide a mathematical framework that represents assumptions about how variables in a system causally influence one another. An structural causal equation consists of three components: a set of *exogenous variables* (U) representing external or unobserved factors whose parents are not present in the causal graph, a set of *endogenous variables* (V) whose causal mechanisms will be modeled from the graph, and a set of *structural functions* (f) that assign values to each endogenous variable as a function of other observed variables [7], [16], [19]. An SCM is a collection of structural equations that describe the relationships of all variables in the model. The exogenous ones determine the values of the endogenous variables. That is, every endogenous variable is a child of one or more exogenous variables. On the other hand, exogenous variables do not have any parents; therefore, they cannot be children of any exogenous variable. Once the values of all exogenous variables are specified, the values of all endogenous variables can be computed deterministically through the structural equations.

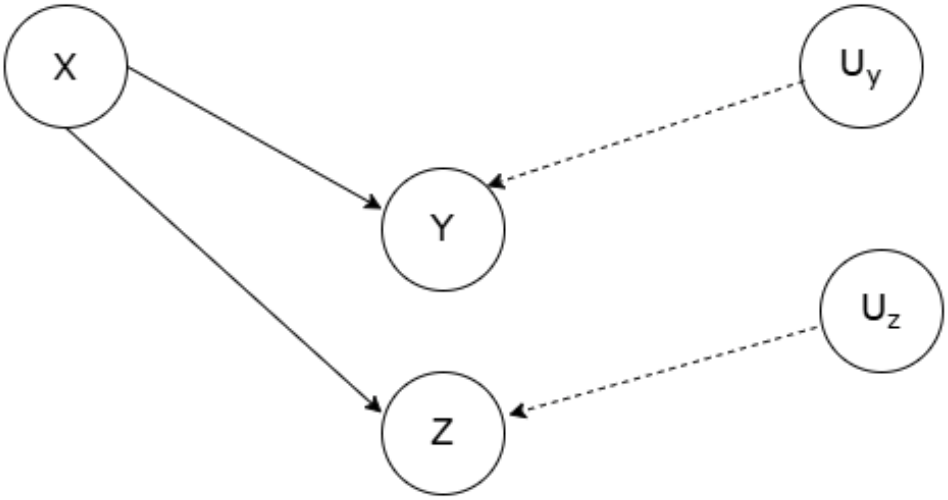

Figure 3.3: Structural Causal Model

As mentioned earlier, every SCM corresponds to a graphical causal model, typically represented in the form of a DAG, where nodes denote variables and directed edges represent causal dependencies implied by the structural functions. In figure 3.3, the corresponding graph refers to a system where X is a direct cause of Y and Z. Again, U_z and U_y are noise terms that are independent of each other. We can represent the structural functions for this model as,

$$Y := f(X, U_y), \qquad Z := f(X, U_z),$$

Here, Y and Z are called endogenous variables - meaning variables whose parents need to be modeled. Whereas, X, U_y, and U_z are the exogenous variables. They are terms whose parents are not a part of the system and hence, do not need to be modeled. Here, as X is a direct cause of Y and Z, it appears in the structural function that is used to model Y and Z. Each structural function showcases how each endogenous variable can be mapped to its causes and noise terms. The SCM is the entire system - a collection of structural functions that models the whole system. This represents a mathematical, structured formulation of the DAG and its inherent causal mechanisms. The SCM will be used to construct the causal questions and estimate the causal effects.

3.3.3 Define Causal Estimand

After identifying the causal graph and model, the next step is to pinpoint the causal estimand we want to estimate. We need to specify the causal question that we want to answer. The expression whose causal effects we aim to obtain. The causal estimand represents the quantity of interest in causal terms. It is an expression that contains the do operator as explained in section 2.2.1. This step allows us to represent the research question as a formal, mathematical causal quantity.
Now, we are focused on finding the average treatment effect (ATE) on the outcome.

That means we want to see the likelihood of depressive symptoms when the independent variables are intervened on. The equation 3.5 represents the structured formulation of the causal estimand. This asks about the difference in the expected outcome of Y when we are intervening on X.

$$E[\,Y \mid do(X = x_1)\,] \;-\; E[\,Y \mid do(X = x_0)\,] \tag{3.5}$$

This basically represents the average causal effect of changing the treatment X from x_0 to x_1 on the outcome Y. If Y is binary, the expectation, $E[Y]$, is just the average likelihood of Y. In contrast, if Y is continuous, then $E[Y]$ is just the mean of the quantitative values. The total difference here represents how much the expected likelihood of the outcome changes when you intervene on X if you change its value from x_0 to x_1 [16], [25]. This is not correlation; rather, this is a causal quantity. This showcases the actual change in the outcome that is caused by changing the treatment. As our target variable is categorical, we will estimate the change in the average likelihood of the target when intervening on any of the other factors.

3.3.4 Identify Statistical Estimand

Before moving on to estimation, we have to remember that the causal estimand deals with the *do* operator, meaning it utilizes potential outcomes. Therefore, the causal estimand revolved around both unobserved and observed outcomes. But our data does not account for unobserved outcomes or counterfactuals. To compute the interventional distribution, we have to repeat the experiment under the intervention. But using the rules of do calculus, we do not have to repeat the experiment. The interventional distribution can be computed from the observational data. As we want to quantify an effect, we need to convert the causal estimand into a mathematical formulation that will be calculated from the observed data, which will be a statistical quantity. We will not have to rely on any unobserved effects. Hence, the process of converting the causal estimand into a statistical estimand that can be computed from the observed data is called **Identification**. Different methods can be adopted to identify the statistical estimand. For our research purpose, we choose to utilize the Backdoor Criterion.

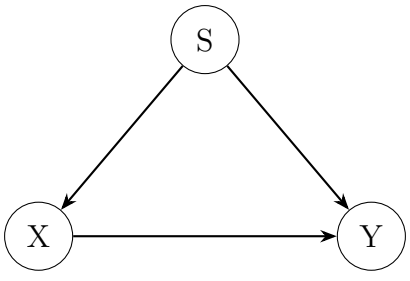

Figure 3.4: Backdoor example with confounder

In the figure 3.4, X is the treatment variable that we intervene on, S is a common cause for both X and Y. Additionally, Y happens to be the outcome and a confounding variable (as shown in section 2.2.2). Then, a *backdoor path* for X is any

non-causal path that enters X through a confounding path. Therefore, the path from $X - S - Y$ is the backdoor path.
A set of nodes Z will satisfy the backdoor criterion for an ordered pair of variables (X, Y) in a DAG, G, if the set of nodes Z satisfies the following constraints:

- Block all backdoor paths between X and Y
- Leave all directed paths from X to Y open
- No node in Z is a descendant of X

If the set of nodes Z satisfies the backdoor criterion, then the likelihood of Y intervening on X can be computed as the expression showcased in equation 3.6.

$$P\big(Y \mid do(X{=}x)\big) \;=\; \sum_{z} P\big(Y \mid X{=}x,\, Z{=}z\big)\, P\big(Z{=}z\big) \tag{3.6}$$

This is what is known as the *backdoor adjustment*, where the set of nodes Z is a sufficient adjustment set [16], [25]. In this case, the adjustment set Z will consist only of node S, as it is part of the only backdoor path, and conditioning on S will block this backdoor path. The above expression is a statistical estimand and can be computed from the observational data.

3.3.5 Estimate Causal Effects

After we identify the statistical estimand, all that is left is using the estimands to compute the causal effects. That is the final stage of the pipeline. We will use machine learning models as estimators to quantify the causal effect. For this, we propose a Generalized Linear Model (GLM), which includes logistic regression, as an estimator. The model will adjust for confounders using the backdoor criterion. The final estimate will show us how much the probability of having depression changes when the treatment variables change from 0 to a specific value, with confounders adjusted for. We will be using treatment = 0 as the control value (baseline).

$$P(Y \mid \mathrm{do}(X = x_0)) - P(Y \mid \mathrm{do}(X = x_1)) = \sum_{z} \Big[P(Y \mid X = x_0, Z = z) - P(Y \mid X = x_1, Z = z)\Big]\, P(Z = z) \tag{3.7}$$

The equation 3.7 is the final expression that will be used to estimate the average treatment effect as our desired causal effect. Here, Z is a sufficient adjustment set that satisfies the backdoor criterion.

All preprocessing steps, tasks related to training and inference for machine learning models, causal discovery, identification of estimands, and causal estimations will be implemented by the scikit-learn, DoWhy, and Causal Learn libraries.

Chapter 4

Result Analysis

This section presents the study's results and analysis. The analysis was carried out in three primary stages: exploratory data analysis (EDA), predictive modeling, and causal inference.

The exploratory phase focused on understanding the relationships among the key variables through summary statistics and correlation. This was followed by the implementation of a pipeline designed to predict depressive symptoms based on multiple psychosocial and social factors. As part of the machine learning pipeline, SHAP has been applied to generate global explanations for the model. The purpose of this stage is to get a better idea about the model's internal reasoning. This stage aims at figuring out which factor is the model considering necessary in predicting depressive disorder. Finally, a causal AI framework based on the principles of causal inference was developed to move beyond correlational and model-dependent associations, allowing for the estimation of the actual causal effects of various factors behind depressive symptoms. Together, these three stages provide a comprehensive understanding of the data, from detecting statistical associations to uncovering potential cause-and-effect relationships.

4.1 Results from EDA

We first performed a correlation test to gauge the association between all the variables. Since our input variables were mostly on Likert scales and our target was binary, we used Spearman's correlation coefficient (ρ) to assess correlation. Among the 73 variables, we chose the 16 variables that had the highest correlation with our target variable, "Depression". The figure 4.1 showcases the 16 variables and their degree of correlation with the binary target. We can see that features P1–P9 exhibit high correlation with the target, whereas features F1 to F7 show a medium to low correlation.

The table 4.1 showcases the 16 features that were selected and the corresponding questions that they refer to.

Table 4.1: Description of Features and Corresponding Questions

Features	**Questions**
F1	I am most afraid of Coronavirus-19.
F2	It makes me uncomfortable to think about Coronavirus-19.
F3	My hands become clammy when I think about Coronavirus-19.
F4	I am afraid of losing my life because of Coronavirus-19.
F5	When watching news and stories about Coronavirus-19 on social media, I become nervous or anxious.
F6	I cannot sleep because I'm worrying about getting Coronavirus-19.
F7	My heart races or palpitates when I think about getting Coronavirus-19.
P1	Little interest or pleasure in doing things.
P2	Feeling down, depressed or hopeless.
P3	Trouble falling or staying asleep, or sleeping too much.
P4	Feeling tired or having little energy.
P5	Poor appetite or overeating.
P6	Feeling bad about yourself—or that you are a failure or have let yourself or your family down.
P7	Trouble concentrating on things, such as reading the newspaper or watching television.
P8	Moving or speaking so slowly that other people could have noticed. Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual.
P9	Thoughts that you would be better off dead, or of hurting yourself.

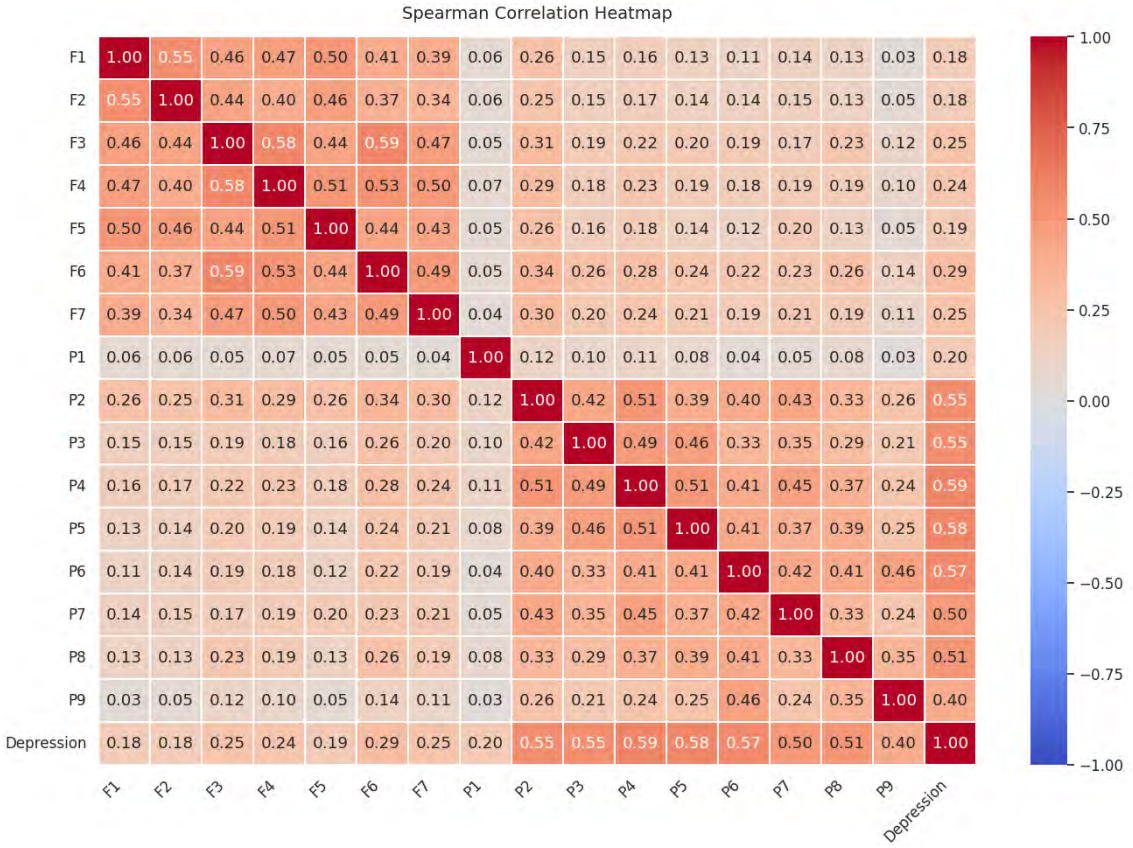

Figure 4.1: Spearman's correlation heatmap

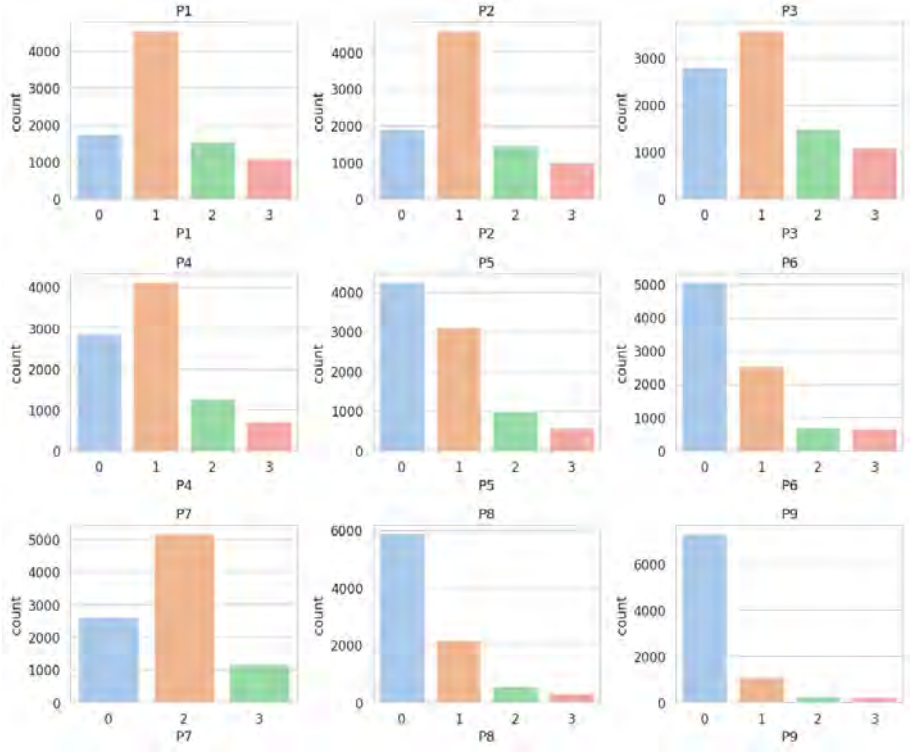

Figure 4.2: Count plot of features P1–P9.

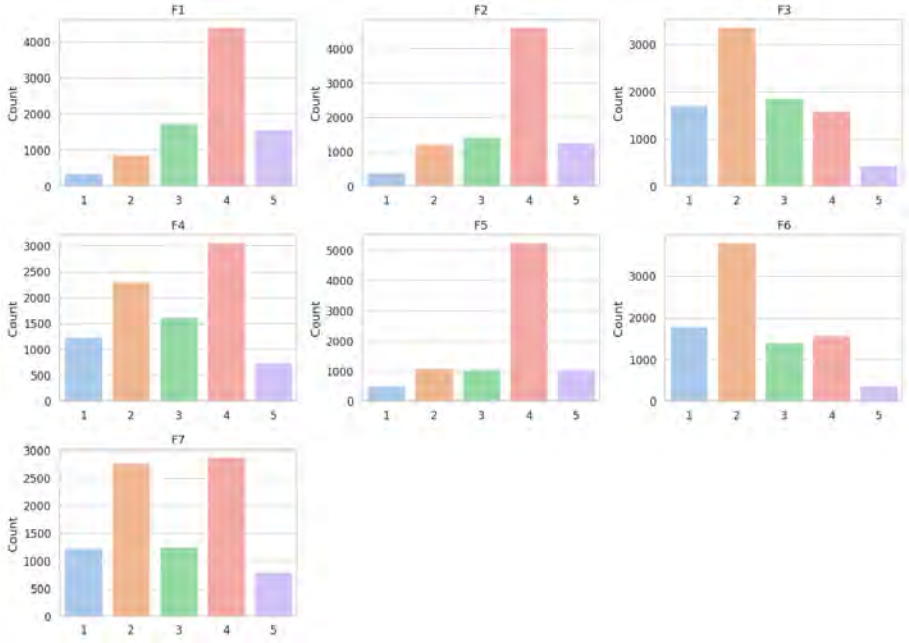

Figure 4.3: Count plot of features F1–F7.

After selecting the features that will be used for all future analysis, we now showcase their distributions using a count plot. The count plots are represented in Figures 4.2 and 4.3 respectively. Each of the features P1 to P9, and F1 to F7 are measured on Likert Scales. For features P1 to P9, most responses are clustered around 0, 1, and 2, indicating that participants reported experiencing depressive symptoms either "not at al" or "several day" or "more than half day" during the assessment period. Features such as P6, P8, and P9 show particularly high frequencies in the "0" category, suggesting that severe symptoms like restlessness, suicidal ideation, or feeling bad about oneself were relatively uncommon among respondents. In contrast, items such as P1 and P2 display a more balanced distribution between the "0","1", and "2" categories, indicating a higher prevalence of mild to moderate symptoms such as lack of interest and feelings of sadness. Overall, the distributions reveal a skew toward lower symptom intensity, implying that while depressive traits were present in the population, severe forms of depression were less frequently reported.

Moreover, for features F1 to F7, most responses are clustered around 2, 3, and 4, exhibiting moderate disagreement to moderate agreement. Features such as F1, F2, and F5 show a right-skewed distribution, where the majority of responses fall in category 4, indicating a strong inclination toward the higher end of the scale.

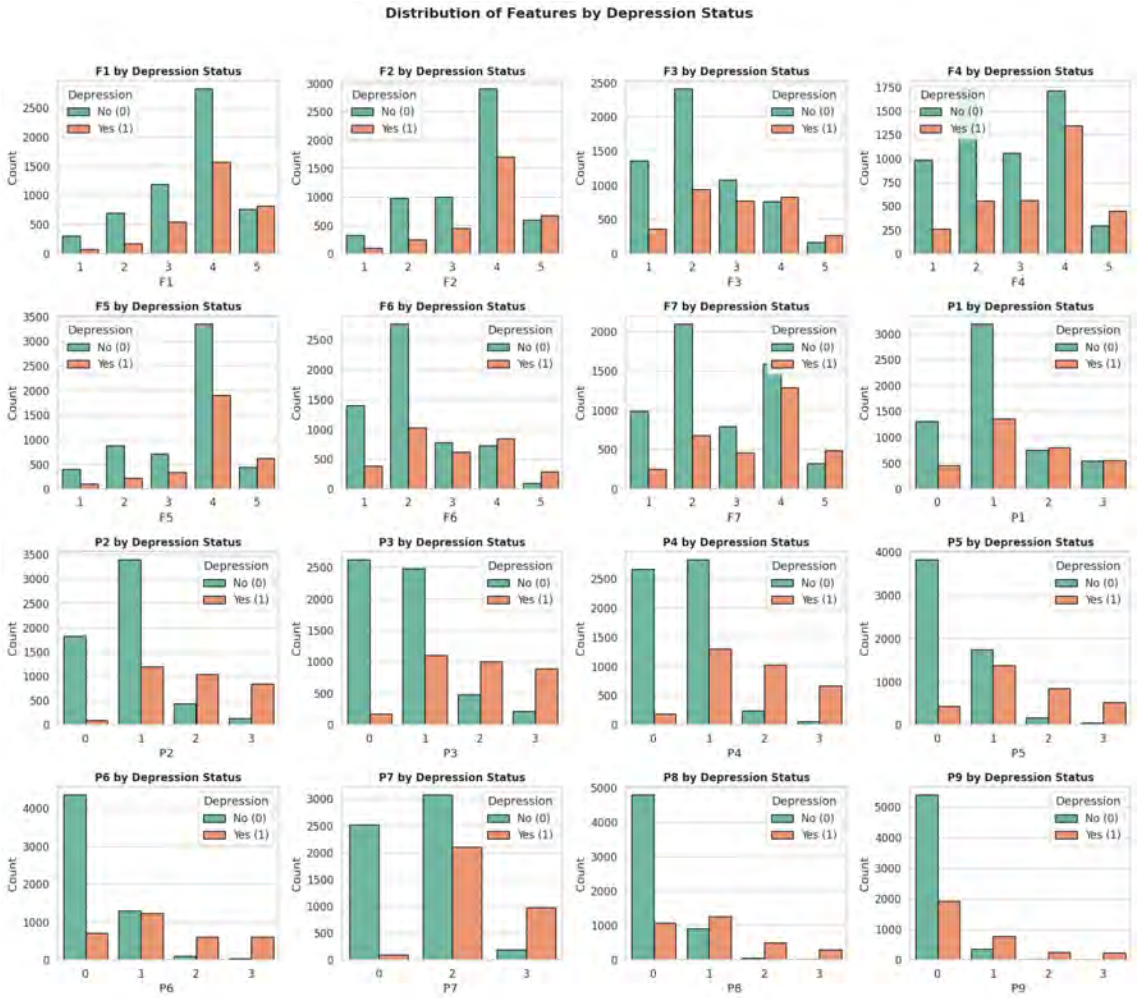

Figure 4.4: Distributions of features P1–P9 and F1–F7 by depression status (No vs. Yes).

Figure 4.4 presents the distribution of responses for both the independent variables F1–F7 and P1–P9, segmented by depression status (No vs. Yes). Across most variables, individuals without depression tend to cluster around lower or mid-range categories, whereas participants with depression show a visible shift toward higher response levels. For example, in features P1–P9, depressed individuals report higher frequencies in categories 2 and 3, indicating a greater intensity and frequency of depressive experiences. We can observe that the features F1 to F7 and P1–P9 are not uniformly distributed across groups. These observed trends provide early evidence of potential associations between the predictors and depression, justifying the application of both machine learning and causal inference analyses to examine these cases.

4.2 Results of the Machine Learning pipeline

In this section, we train an XGBoost classifier using a 60:20:20 train-validation-test ratio. For this, all 16 independent variables (F1-F7, P1-P9) are used. Later, we used SHAP to generate explanations for the model. This would help us better understand the internal workings of the model and assess which features it considers good predictors. The table 4.2 shows the precision, recall, and prediction accuracy of the XGBoost model. So based on the 16 high correlated features - from this score, we see that our model is capable of predicting the likelihood of depression. But we are less focused on model accuracy, rather we want to focus on what helped the model get this score.

Precision	**Recall**	**Accuracy**
95.40%	93.30%	96.20%

Table 4.2: Results of model with all 16 features

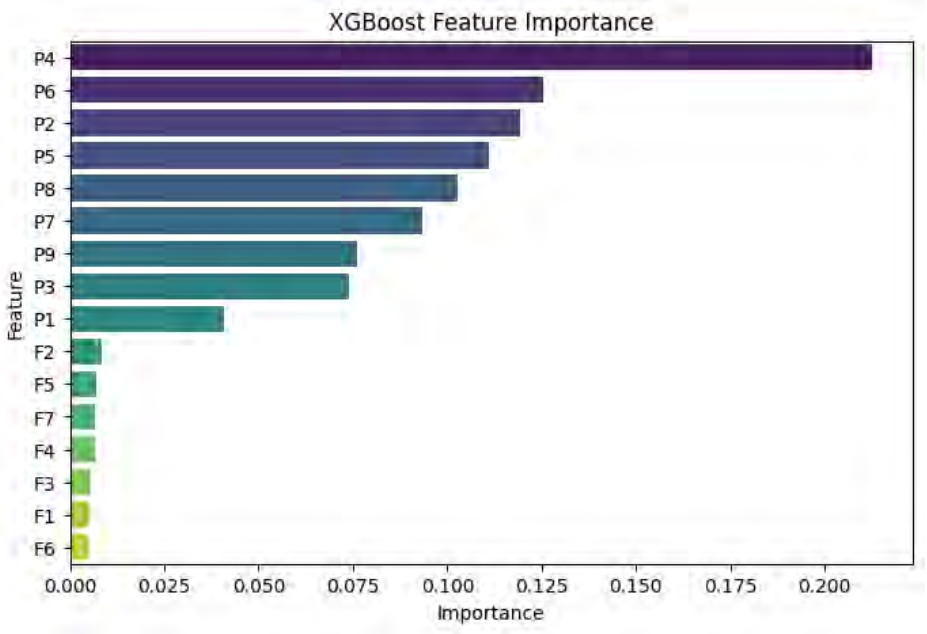

Figure 4.5: XGBoost Feature Importance based on Gain.

The XGBoost model uses *gain* as the criterion for choosing the best feature and threshold for splitting at each node. The figure 4.5 showcases the gain value for each feature. It helps us to understand the influence of each feature based on gain. This shows that P4 is the feature with the highest gain and, therefore, reduces entropy the most. One common trend we observe is that all features from P1 to P9 are considered necessary by the model when creating splits. On the other hand, features F1 to F7 are given comparatively less importance due to their low gain value. This corresponds with the correlation test as well. Just as the features P1-P9 are more correlated than the features F1-F7, the same group of features also work as better criterias for split compared to F1-F7.

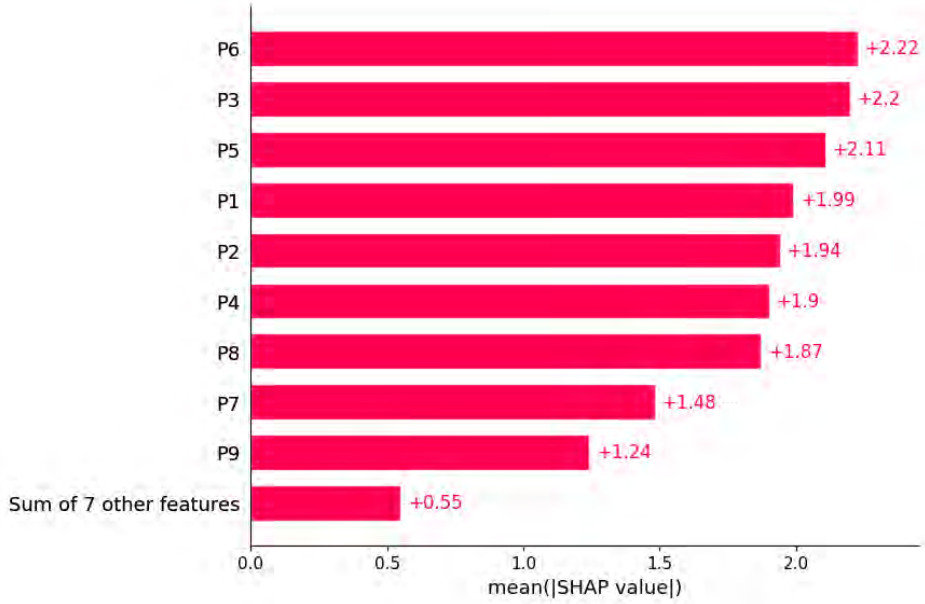

Figure 4.6: Mean SHAP value by feature (global importance).

In the next stage, we applied the *TreeExplainer* module from SHAP to explain the predictions made by the XGBoost model. We will look into the global explanations by calculating the mean of the absolute values of the SHAP values for all instances per feature. The figure 4.6 allows us to visualize this. This plot focuses solely on the average strength of the impact of the features, irrespective of direction. A higher value means the feature has a greater average influence on the model's output. These values are always positive, since they measure magnitude, not direction. Each SHAP value for the features basically shows how much that feature pushes the predicted probability away from the baseline or average.

The features P6, P3, and P5 exhibit the highest mean SHAP values (2.22, 2.20, and 2.11, respectively), indicating that they are the most influential predictors of the model's output. A value of +2.22 for *P6* indicates that, overall, considering all the data points, this feature changed the model's prediction by +0.25 compared to the baseline. In contrast, items such as P7, P9, and the combined contribution of seven features F1 to F7 demonstrate comparatively weaker effects.
We have accumulated the results using both the gain and shap values of the 16 selected features and have ranked them, which is shown in the table 4.3. From the table we see that the highly correlated features P1-P9 - all showcase high gain and shap values compared to the less correlated F1-F7 features. Based on both the gain and the shap values, we can understand that the model pays more importance to the P1-P9 features compared to F1-F7 when assessing depression. The correlation found in the data has been perfectly reflected in the rankings based on shap and gain. Overall we can say that the same features that show high correlations with "Depression", are the same ones that have been assigned both high gain values and high shap values compared to the less correlated F1-F7 features.

Now, while tools such as SHAP and model-based feature importance measures like gain can provide valuable insights into how a model makes predictions, they remain

Table 4.3: Ranking of features based on Gain and SHAP values

Feature	**Question**	**Gain**	**SHAP**
P1	Little interest or pleasure in doing things.	9	4
P2	Feeling down, depressed or hopeless.	3	5
P3	Trouble falling or staying asleep, or sleeping too much.	8	2
P4	Feeling tired or having little energy.	1	6
P5	Poor appetite or overeating.	4	3
P6	Feeling bad about yourself—or that you are a failure or have let yourself or your family down.	2	1
P7	Trouble concentrating on things, such as reading the newspaper or watching television.	6	8
P8	Moving or speaking so slowly that other people could have noticed. Or the opposite—being so fidgety or restless that you have been moving around a lot more than usual.	5	7
P9	Thoughts that you would be better off dead, or of hurting yourself.	7	9
F1	I am most afraid of Coronavirus-19.	15	15
F2	It makes me uncomfortable to think about Coronavirus-19.	10	12
F3	My hands become clammy when I think about Coronavirus.	14	11
F4	I am afraid of losing my life because of Coronavirus-19.	13	14
F5	When watching news and stories about Coronavirus-19 on social media, I become nervous or anxious.	11	16
F6	I cannot sleep because I'm worrying about getting Coronavirus-19.	16	13
F7	My heart races or palpitates when I think about getting Coronavirus-19.	12	10

fundamentally correlation-driven. They help identify which features the model considers good predictors, but not which features can actually change the outcome in the real world. Feature importance metrics, such as gain in XGBoost, measure the contribution of each variable to the model's predictive performance rather than its causal relevance. In addition, SHAP assumes conditional or marginal independence among the input features- a constraint that is often broken in psychological or social data where confounding and interdependence are common. As a result, these methods fail to uncover the underlying causal mechanisms. Hence, while XAI enhances explainability in model behavior, it cannot establish whether manipulating a given feature would actually change the outcome. Therefore, we now move on to the causal AI pipeline and analyze its results to understand which features, when intervened on, can actually change the outcome.

4.3 Results of the Causal AI Pipeline

After the machine learning pipeline, we lean towards the Causal AI pipeline and analyze the obtained results. Using the 16 features as independent predictors and Depression as the target, we first showcase the causal graph obtained using the FCI algorithm for causal discovery. Later, we fit the SCM model to instantiate the causal model that will be used to perform interventions and estimate causal effects.

4.3.1 Causal Discovery using FCI

The figure 4.7 represent the causal graph that was obtained using the FCI algorithm. From the obtained DAG, we noticed that only the features P1-P9 had direct causal link with the target variable, "Depression". Therefore, to reduce complexity, we only showcase the part of the graph where we the features that have direct causal effects on the target. We can see that features P1 to P9 have direct causal effects on Depression. Furthermore, we can visualize what causal relationships they showcase among each other. Now that we have the final causal structure, we can move on to quantifying and analyzing the average treatment effects on Depression.

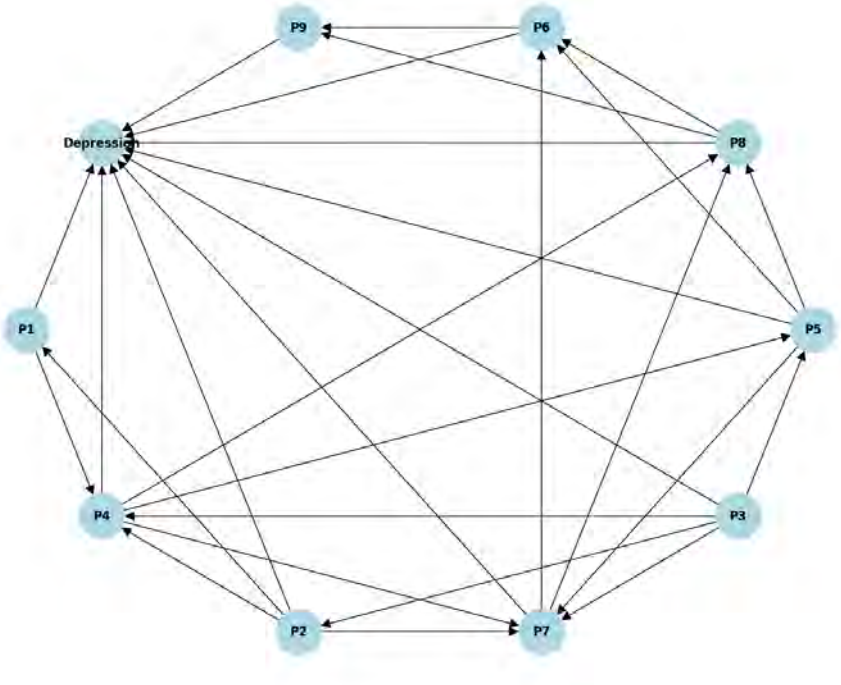

Figure 4.7: Causal graph (DAG) discovered via FCI

4.3.2 Estimation of Causal Effects

At the final stage of the Causal AI pipeline, we will implement the SCM on the basis of our causal graph to serve as our causal model. To quantify the causal effects of each independent variable P1 to P9 on depressive disorder, the causal model was implemented using the `DoWhy` library in Python. Each of the features, P1-P9 were used as a treatment variable, keeping Depression as the outcome.

The causal effects were estimated using the strategies following Judea Pearl's do calculus. Once the statistical estimands were properly identified, sufficient adjustment sets were utilized to compute the causal effects. The estimation of the direct causal effects was carried out using a Generalized Linear Model (GLM). Since our outcome variable, "Depression" is a binary valued feature, the GLM (a member of the binomial family) was defined with a logit function, hence making it equivalent to a logistic regression model. This means the model estimates the likelihood of depression as a linear combination of the treatment and adjustment variables. The logistic regression model here helps to assess how changes in each treatment category affect the probability of depression, while holding the confounders constant. The coefficients from the GLM represent the change in the likelihood of depression for a one-unit increase in the treatment variable.

For each treatment variable, we set the control or baseline level as 0, and the treatment levels as all other possible values of the features. The model compares the expected probability of depression between the baseline level and each of the higher response categories of the treatment level. This results in a set of Average Treatment Effects (ATEs) for every feature. Each value is our desired causal effect of the features from P1 to P9 on "Depression". These values represent how changes in the features from P1 to P9 influence the likelihood of depression when confounding variables are appropriately adjusted.

The table 4.4 exhibits the estimated causal effects and the adjustment sets for each treatment variable when Depression is the outcome and the features P1-P9 are used as the corresponding treatment. Here we can see that the features P2 to P6 are the top contributing variables, with P3 having the highest causal effect on the outcome. For one-unit change in P3 can alter the likelihood of Depression by a value of 0.509.

Table 4.4: Estimation of Causal Effects

Feature	**Average Causal Effects**	**Adjustment sets**
P1	0.154	backdoor: ['P2']
P2	0.439	backdoor: ['P3']
P3	0.509	backdoor: []
P4	0.372	backdoor: ['P3', 'P1', 'P2']
P5	0.365	backdoor: ['P3', 'P4']
P6	0.278	backdoor: ['P7', 'P5', 'P8']
P7	0.272	backdoor: ['P2', 'P3', 'P4', 'P5']
P8	0.299	backdoor: ['P4', 'P7', 'P5']
P9	0.272	backdoor: ['P6', 'P8']

Now, table 4.4 represents a final ranking of the features combining the results from the three applied approaches — gain-based feature importance, SHAP value interpretation, and causal effect estimation. This provides a comprehensive understanding of how different depressive symptoms influence the likelihood of overall depression. The gain scores from the XGBoost model indicate that features such as P4, P6, and P2 contribute most to reducing entropy and uncertainty. Moreover, the SHAP analysis, which explains the internal reasoning of the model, ranks P6, P3, and P5 as the most influential features in determining prediction outcomes. Meanwhile, the causal analysis reveals that P3, P2, and P4 exert the largest direct causal effects on depression probability, with causal magnitudes of 0.509, 0.439, and 0.372, respectively.

Table 4.5: Feature ranking based on Gain, SHAP, and ATE

Feature	**Gain**	**Effects**	**ATE**
P1	9	4	9
P2	3	5	2
P3	8	2	1
P4	1	6	3
P5	4	3	4
P6	2	1	6
P7	6	8	7
P8	5	7	5
P9	7	9	8

We can see that some overlap exists in our results. For example, the features P2, P3, and P4 appear consistently important across all three methods. The correlation that was showcased on the data, is prevalent in the causal mechanisms as well.

On the other hand, noticeable discrepancies are prevalent as well. Features like P6, which have high gain and SHAP values, show a comparatively lower causal effect. This suggests that while the model heavily relies on them for prediction, their influence may be more correlational than causal. Again, P3, which ranks lower in gain but shows a high causal effect, can indicate that specific symptoms may exert strong real-world causal influence even if the model doesn't find them as significant predictors. These variations highlight a critical distinction between what the model finds useful for prediction and what truly drives depressive outcomes.

The results show that while models such as XGBoost can effectively predict depressive symptoms and identify statistically significant associations through gain and SHAP analyses, these techniques alone are insufficient for drawing causal conclusions. Several features, such as P6 and P5, appeared highly important in model-based interpretations but showed weaker causal effects, indicating that their relevance may stem from correlation rather than direct influence. Conversely, features like P3 and P2, which displayed stronger causal effects despite modest predictive rankings, highlight that certain depressive symptoms may play a more fundamental role in driving mental health deterioration. These findings reinforce the crucial distinction between predictive importance and causal importance.

Overall, this comparison emphasizes that while traditional machine learning and XAI methods can effectively identify associations and help to understand model-based importance, they do not necessarily showcase causal mechanisms. The causal inference results extend this understanding by identifying which features are likely to cause changes in depression probability rather than merely accompany them.

4.4 Dimensionality Reduction Based on Correlation and Causation

Building on this, we also propose a pipeline for dimensionality reduction. We have created three versions of our model, M1, M2, and M3. Now, the first version, M1 trains our model on all 16 features (P1-P9, F1-F7) and then evaluates it on the basis of prediction accuracy. The second version, M2 is trained only on the features that have direct causal effects on the target. Hence, features P1 - P9 are used to train this model, as they are the only ones with a direct causal link to the "Depression. Lastly, we built a third version of our model, M3, that will be trained on only those features that have a direct causal link with the target (P1-P9) and some selected features from F1-F7. A reminder that the features F1-F7, even though were correlated with the outcome, they were not direct causes of Depression. Now we will use backward selection based on correlation to select features from F1 - F7. We want to evaluate how predictive performance changes when we include variables that are both correlated and causally related to depression. We propose an iterative feature selection pipeline. We begin the process by dividing the dataset into two feature groups: the fixed features, which include P1 to P9 and were previously identified as causally related to depression. Then we create a set of variable features (F1 to F7), which represent additional factors that were determined not to be causally related but still highly correlated with Depression. We used the absolute values of the correlations to rank the variable features in descending order of strength. The next stage involved modeling a loop, where the XGBoost classifier was trained multiple times, each time using a smaller set of variable features (F1-F7). In each iteration, the feature with the weakest correlation to depression was removed from the input set, while the causally significant P1 - P9 features remained fixed. The model was retrained at every step, and its predictive performance was evaluated using prediction accuracy. We aimed to examine how model accuracy changes as correlated but non-causal features are gradually excluded as model input.

After further evaluation, we record that the causal features, along with two features from the variable set, F6 and F7, record the highest prediction accuracy. That means that even though the features F1 - F5 showcase high correlation with the target, not all of the features are necessary for predicting when the causal links are present. The table 4.6 helps to visualize the results of all three versions of the model. From this, we can see that even though the features F1-F7 are correlated, not all of them are needed for our model to converge. By keeping the direct causes constant, and only using two additional features, our model performed better than the version of the model that was trained on all 16 features. One more thing we can notice is that for optimal performance, simply the direct causes were not sufficient. At the

end of the day, machine learning models take advantage at correlation. Since the direct causes were not enough, the model needed assistance of additional correlated features, which in this case are F6 and F7 respectively.

Therefore, the results from this pipeline highlight which features contribute meaningfully to prediction and which act as redundant or weakly informative variables. By comparing performance across iterations, it was possible to assess the marginal utility of including correlated variables alongside the causally verified ones. This approach bridges the predictive insights from machine learning with the causal findings from earlier analysis, demonstrating how incorporating causally relevant variables yields both interpretable and empirically strong models.

Table 4.6: Comparison Of Models before and after dimensionality reduction

Model	**Prediction Accuracy**	**Feature Set**
M1	96.2	All 16 features
M2	93.8	Feature P1 – P9
M3	97	Features P1 – P9, F6 and F7

Based on the overall results, it is evident that for inferential tasks, if the experimental setup is similar to ours, we can implement a similar feature selection technique to reduce model complexity. We see that even though machine learning models focus on correlation, using causal factors can help the model to converge more efficient and gives more meaning to its internal reasoning.

Chapter 5

Conclusion

This study set out to understand the causal effects of different factors behind depression, using a combination of machine learning, explainable AI, and causal inference techniques. While existing research has largely focused on identifying correlations and associations, this work aimed to go beyond prediction and uncover the underlying mechanisms that truly drive depressive outcomes. By integrating these three analytical layers, the research not only achieved high predictive accuracy but also provided a transparent, causally grounded interpretation of which features genuinely influence depression.

Furthermore, the dimensionality reduction pipeline combining causal and correlated variables demonstrated that incorporating causally verified features ensures robust and interpretable model performance. Even when correlated non-causal features were excluded, predictive accuracy remained high, confirming that models grounded in causal understanding can retain both explanatory clarity and empirical strength.

Overall, this thesis emphasizes the value of combining causal inference with explainable AI to move from "which features matter" to "which features truly make a difference." By identifying symptoms that not only predict but also causally influence depression, this research offers insights that are both scientifically meaningful and practically actionable. At the policy and clinical level, such findings can support targeted interventions aimed at modifying causal drivers rather than merely correlated indicators, contributing toward more effective prevention and treatment strategies for mental health.

In this sense, causal AI extends and strengthens the goals of XAI. It does not discard feature-importance methods but situates them within a broader framework of explanation that accounts for interventions and counterfactual reasoning. By clarifying what should be changed, rather than just what is correlated, causal AI provides actionable insights that move beyond transparency toward genuine understanding and control. They represent the model's dependence on correlated features rather than actual interventional effects. Therefore, a causal inference framework, as employed in this study is necessary to distinguish statistical association from genuine causation.

5.1 Future Works

While this study offers an integrated framework for understanding the causal mechanisms behind depression, several extensions remain open for future exploration. One promising direction would be the application of counterfactual analysis, which can provide individual-level explanations by estimating how outcomes might change under hypothetical interventions. This would allow a deeper understanding of not only which factors cause depression but also how altering these factors could prevent or mitigate it.

Another area for development lies in hybrid causal-learning architectures, where causal inference principles are embedded directly into machine learning pipelines. Such approaches could enhance the interpretability and stability of predictions while addressing confounding bias at the model level. Lastly, expanding the dataset to include diverse demographic or cultural contexts beyond Bangladesh would allow for cross-cultural validation and generalization of causal relationships in mental health.

Through these extensions, future research can continue to refine the balance between predictive performance and causal interpretability, moving closer to the long-term goal of using Causal AI for personalized and effective mental health interventions.

Bibliography

[1] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[2] D. M. Chickering, "Optimal structure identification with greedy search," *Journal of machine learning research*, vol. 3, no. Nov, pp. 507–554, 2002.

[3] J. Pearl, "Statistics and causal inference: A review," *Test*, vol. 12, no. 2, pp. 281–345, 2003.

[4] A. N. Chowdhury, S. G. Sayanti Ghosh, and D. S. Debasish Sanyal, "Bengali adaptation of brief patient health questionnaire for screening depression at primary care.," 2004.

[5] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, and M. Jordan, "A linear non-gaussian acyclic model for causal discovery.," *Journal of Machine Learning Research*, vol. 7, no. 10, 2006.

[6] J. Pearl, "Causal inference in statistics: An overview," 2009.

[7] J. Pearl, *Causality*. Cambridge university press, 2009.

[8] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer, "Estimation of a structural vector autoregression model using non-gaussianity.," *Journal of Machine Learning Research*, vol. 11, no. 5, 2010.

[9] J. Pearl, "An introduction to causal inference," *The international journal of biostatistics*, vol. 6, no. 2, p. 7, 2010.

[10] J. Pearl, "Causal inference," *Causality: objectives and assessment*, pp. 39–58, 2010.

[11] S. Shimizu et al., "Directlingam: A direct method for learning a linear non-gaussian structural equation model," *Journal of Machine Learning Research-JMLR*, vol. 12, no. Apr, pp. 1225–1248, 2011.

[12] C. Meek, "Causal inference and causal explanation with background knowledge," *arXiv preprint arXiv:1302.4972*, 2013.

[13] P. L. Spirtes, C. Meek, and T. S. Richardson, "Causal inference in the presence of latent variables and selection bias," *arXiv preprint arXiv:1302.4983*, 2013.

[14] T. Chen et al., "Xgboost: Extreme gradient boosting," *R package version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[15] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

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

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[18] B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and C. Glymour, "Generalized score functions for causal discovery," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 1551–1560.

[19] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic books, 2018.

[20] M. Oprescu, V. Syrgkanis, K. Battocchi, M. Hei, and G. Lewis, "Econml: A machine learning library for estimating heterogeneous treatment effects," in *33rd Conference on Neural Information Processing Systems*, 2019, p. 6.

[21] W. Cullen, G. Gulati, and B. D. Kelly, "Mental health in the covid-19 pandemic," *QJM: An International Journal of Medicine*, vol. 113, no. 5, pp. 311–312, 2020.

[22] N. Donthu and A. Gustafsson, *Effects of covid-19 on business and research*, 2020.

[23] A. Haleem, M. Javaid, and R. Vaishya, "Effects of covid-19 pandemic in daily life," *Current medicine research and practice*, vol. 10, no. 2, p. 78, 2020.

[24] K. Kontoangelos, M. Economou, and C. Papageorgiou, "Mental health effects of covid-19 pandemia: A review of clinical and psychological traits," *Psychiatry investigation*, vol. 17, no. 6, p. 491, 2020.

[25] B. Neal, "Introduction to causal inference," *Course lecture notes (draft)*, vol. 132, 2020.

[26] A. H. Pakpour, F. Al Mamun, I. Hosen, M. D. Griffiths, and M. A. Mamun, "A population-based nationwide dataset concerning the covid-19 pandemic and serious psychological consequences in bangladesh," *Data in brief*, vol. 33, p. 106 621, 2020.

[27] A. Sharma and E. Kiciman, "Dowhy: An end-to-end library for causal inference," *arXiv preprint arXiv:2011.04216*, 2020.

[28] A. Sharma and W. J. Verbeke, "Improving diagnosis of depression with xgboost machine learning model and a large biomarkers dutch dataset (n= 11,081)," *Frontiers in big Data*, vol. 3, p. 523 466, 2020.

[29] C. A. Tisdell, "Economic, social and political issues raised by the covid-19 pandemic," *Economic analysis and policy*, vol. 68, pp. 17–28, 2020.

[30] P. Beaumont et al., *CausalNex*, Oct. 2021. [Online]. Available: https://github.com/quantumblacklabs/causalnex.

[31] U. M. Haque, E. Kabir, and R. Khanam, "Detection of child depression using machine learning methods," *PLoS one*, vol. 16, no. 12, e0261131, 2021.

[32] J. Hoofman and E. Secord, "The effect of covid-19 on education," *Pediatric Clinics of North America*, vol. 68, no. 5, p. 1071, 2021.

[33] R. Rois, M. Ray, A. Rahman, and S. K. Roy, "Prevalence and predicting factors of perceived stress among bangladeshi university students using machine learning algorithms," *Journal of Health, Population and Nutrition*, vol. 40, no. 1, p. 50, 2021.

[34] V. P. C. Magboo and M. S. A. Magboo, "Important features associated with depression prediction and explainable ai," in *International Conference on Well-Being in the Information Society*, Springer, 2022, pp. 23–36.

[35] U. B. Munir, M. S. Kaiser, U. I. Islam, and F. H. Siddiqui, "Machine learning classification algorithms for predicting depression among university students in bangladesh," in *Proceedings of the Third International Conference on Trends in Computational and Cognitive Engineering: TCCE 2021*, Springer, 2022, pp. 69–80.

[36] M. I. H. Nayan et al., "Comparison of the performance of machine learning-based algorithms for predicting depression and anxiety among university students in bangladesh: A result of the first wave of the covid-19 pandemic," *Asian Journal of Social Health and Behavior*, vol. 5, no. 2, pp. 75–84, 2022.

[37] M. Trott, R. Driscoll, E. Iraldo, and S. Pardhan, "Changes and correlates of screen time in adults and children during the covid-19 pandemic: A systematic review and meta-analysis," *EClinicalMedicine*, vol. 48, 2022.

[38] S. S. Band et al., "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Informatics in Medicine Unlocked*, vol. 40, p. 101 286, 2023.

[39] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *Data Mining and Knowledge Discovery*, vol. 37, no. 5, pp. 1719–1778, 2023.

[40] H. Byeon, "Advances in machine learning and explainable artificial intelligence for depression prediction," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.

[41] J. Jung, H. Lee, H. Jung, and H. Kim, "Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review," *Heliyon*, vol. 9, no. 5, 2023.

[42] R. Kamil and A. R. Abbas, "Predicating depression on twitter using hybrid model bilstm-xgboost," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3620–3627, 2023.

[43] N. Marriwala, D. Chaudhary, et al., "A hybrid model for depression detection using deep learning," *Measurement: Sensors*, vol. 25, p. 100 587, 2023.

[44] N. Priftis and D. Panagiotakos, "Screen time and its health consequences in children and adolescents," *Children*, vol. 10, no. 10, p. 1665, 2023.

[45] Z. Tian et al., "Predicting depression and anxiety of chinese population during covid-19 in psychological evaluation data by xgboost," *Journal of Affective Disorders*, vol. 323, pp. 417–425, 2023.

[46] A. Ayodele, A. Adetunla, and E. Akinlabi, "Prediction of depression severity and personalised risk factors using machine learning on multimodal data.," *International Journal of Online & Biomedical Engineering*, vol. 20, no. 9, 2024.

[47] P. Blöbaum, P. Götz, K. Budhathoki, A. A. Mastakouri, and D. Janzing, "Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models," *Journal of Machine Learning Research*, vol. 25, no. 147, pp. 1–7, 2024.

[48] A. H. Chowdhury, D. Rad, and M. S. Rahman, "Predicting anxiety, depression, and insomnia among bangladeshi university students using tree-based machine learning models," *Health Science Reports*, vol. 7, no. 4, e2037, 2024.

[49] A. A. Jo, E. D. Raj, A. S. Vino, and P. V. Menon, "Exploring explainable ai for enhanced depression prediction in mental health," in *2024 First International Conference on Innovations in Communications, Electrical and Computer Engineering (ICICEC)*, IEEE, 2024, pp. 1–7.

[50] M. Z. Naser, "Causality and causal inference for engineers: Beyond correlation, regression, prediction and artificial intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 4, e1533, 2024.

[51] J. H. Rony, M. M. Syeed, R. H. Khan, K. Fatema, M. S. Hossain, and M. F. Uddin, "Predicting depression among university students: A comparative assessment of ml & dl models using xai," in *2024 23rd International Symposium on Communications and Information Technologies (ISCIT)*, IEEE, 2024, pp. 175–180.

[52] Y. Zheng et al., "Causal-learn: Causal discovery in python," *Journal of Machine Learning Research*, vol. 25, no. 60, pp. 1–8, 2024.

[53] M. S. Azad, S. I. Leeon, R. Khan, N. Mohammed, and S. Momen, "Sad: Self-assessment of depression for bangladeshi university students using machine learning and nlp," *Array*, vol. 25, p. 100 372, 2025.

[54] G. Carloni, A. Berti, and S. Colantonio, "The role of causality in explainable artificial intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 15, no. 2, e70015, 2025.

[55] M. Mia et al., "Adaptation and validation of the phq-9 scale in bangla for bangladesh police personnel," *Journal of Police and Criminal Psychology*, pp. 1–10, 2025.

[56] A. Rawal, A. Raglin, D. B. Rawat, B. M. Sadler, and J. McCoy, "Causality for trustworthy artificial intelligence: Status, challenges and perspectives," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–30, 2025.

[57] World Health Organization, *Depressive disorder (depression)*, https://www.who.int/news-room/fact-sheets/detail/depression, Accessed: 2025-10-02, 2025.

Appendix

Survey Questions from the Nationwide COVID-19 Dataset (Bangladesh)

The following section lists all the survey questions included in the nationwide study titled *"A population-based nationwide dataset concerning the COVID-19 pandemic and serious psychological consequences in Bangladesh"* [26].

Socio-demographic Information

Participants were asked to provide information on:

- Age
- Gender
- Educational status
- Occupational status
- Place of residence (village, town, city)
- Marital status
- Smoking and alcohol-drinking habits
- Presence of comorbidities or chronic illnesses
- Current health condition
- Frequency of social media use
- Sources of information about COVID-19 (e.g., social media, TV, newspaper)

COVID-19 Knowledge (True/False Questions)

1. COVID-19 can spread through coughing or exhalation from infected individuals.
2. COVID-19 can spread by touching infected individuals.
3. COVID-19 can spread from wild animals.
4. COVID-19 can spread from faces of infected individuals.

5. COVID-19 can spread from companion animals such as cats and dogs.
6. COVID-19 can spread through parcels from infected countries.
7. The incubation period ranges from 2 to 14 days.
8. Some infected individuals may not develop symptoms.
9. Common symptoms include fever, tiredness, and dry cough.
10. Individuals may develop respiratory problems.
11. Some individuals may experience aches, sore throat, or diarrhea.
12. Individuals with comorbidities are more likely to develop serious illness.
13. Washing hands regularly for 20 seconds is necessary.
14. Avoid touching eyes, nose, and mouth.
15. Wearing masks is mandatory.
16. Avoid close contact with infected persons.
17. Maintain at least one-meter distance from others.
18. Quarantine at home if feeling unwell or isolate infected individuals.
19. Taking paracetamol can help manage fever.
20. There is no vaccine or specific antiviral treatment for COVID-19.

COVID-19 Preventive Behaviors (Likert Scale)

1. How often do you clean your hands with alcohol-based rub or soap and water?
2. How often do you cover your mouth and nose when coughing or sneezing?
3. How often do you maintain at least one-meter distance from others?
4. How often do you stay at home if you feel unwell?
5. How many days have you self-isolated during the past week?
6. How many days did you go outside for more than 15 minutes in the past week?
7. How many days did you have face-to-face contact with others for more than 15 minutes?

Lockdown-Related Questions

1. Feeling uncomfortable during lockdown.
2. Unable to buy necessary items.
3. Unable to maintain usual daily routine.
4. Unable to engage in physical exercise.
5. Afraid of going outside to sunbathe or walk.
6. Unable to play in the field.
7. Unable to concentrate on household activities.
8. Facing other problems not listed above.
9. Having enough food supply.
10. Experiencing panic due to economic recession.
11. Facing economic hardship.

Fear of COVID-19 Scale (F1 - F7)

1. I am most afraid of Coronavirus-19.
2. It makes me uncomfortable to think about Coronavirus-19.
3. My hands become clammy when I think about Coronavirus-19.
4. I am afraid of losing my life because of Coronavirus-19.
5. When watching news about Coronavirus-19, I become nervous or anxious.
6. I cannot sleep because I am worrying about getting Coronavirus-19.
7. My heart races or palpitates when I think about getting Coronavirus-19.

Insomnia Severity Index (ISI)

1. Difficulty falling asleep.
2. Difficulty staying asleep.
3. Problems waking up too early.
4. Satisfaction with current sleep pattern.
5. How noticeable your sleep problem is to others.
6. How worried or distressed you are about your current sleep problem.
7. To what extent your sleep problem interferes with daily functioning.

Patient Health Questionnaire (P1-P9)

1. Little interest or pleasure in doing things.
2. Feeling down, depressed, or hopeless.
3. Trouble falling or staying asleep, or sleeping too much.
4. Feeling tired or having little energy.
5. Poor appetite or overeating.
6. Feeling bad about yourself or that you are a failure.
7. Trouble concentrating on reading or watching television.
8. Moving or speaking slowly or feeling restless.
9. Thoughts of being better off dead or self-harm.

Suicidal Behavior

- Do you think about committing suicide, and are these thoughts persistent and related to COVID-19 issues?