Finding Effects of drugs on COVID outcome using Causal Inference Method

Sai Sritej Palacharla
Computer Science
University Of Massachusetts Lowell
Lowell, Ma
Saisritej_Palacharla@Student.Uml.Edu

Daniel Ratnakar Bejjam
Computer Science
University Of Massachusetts Lowell
Lowell, Ma
Danielratnakar_Bejjam@Student.Uml.Edu

Prathap Ramachandra
Computer Science
University Of Massachusetts Lowell
Lowell, Ma
Prathap_Ramachandra@Student.Uml.Edu

Abstract*— This study uses a causal inference approach to examine how different drugs affect COVID-19 outcomes. We have implemented several algorithms, including Logistic Regression, Naive Bayesian, Neural Networks, XG Boost, and Gradient Boost, without standard machine learning libraries. Our goal is to better understand how specific treatments influence patient recovery rates. To achieve this, we have carefully preprocessed data, developed custom algorithmic solutions, and conducted rigorous evaluations. The initial results indicate significant differences in drug efficacy, which could be used to develop targeted therapeutic strategies. This research contributes to the ongoing efforts to manage and treat COVID-19 by providing a comprehensive analysis of causal relationships between drug interventions and patient outcomes.

Keywords*— Causal Inference, COVID-19 Treatment Outcomes, Drug Efficiency, Logistic Regression, Naive Bayesian Algorithm, Neural Networks, Predictive Modeling, Non-library Algorithm Implementation, Treatment Effect Analysis.

I. INTRODUCTION

The COVID-19 pandemic has posed unprecedented challenges to global health systems, requiring rapid advancements in therapeutic interventions. Understanding the effectiveness of various drugs in treating COVID-19 is critical to improving patient outcomes and managing the spread of the disease. Traditional observational studies often struggle with confounding factors that obscure the true effects of treatments. This project addresses these challenges by using causal inference methods, which are designed to reveal potential cause-and-effect relationships between drug treatments and patient outcomes.

Causal inference offers a robust framework for analyzing observational data, which enables more accurate assessments of treatment effectiveness outside of randomized controlled trials. By adopting this methodology, the study aims to provide insights that are statistically significant and causally interpretable. This research is particularly important given the urgent need for effective COVID-19 treatments and the enormous amount of data generated from various clinical settings worldwide.

Additionally, this project is unique in implementing various predictive algorithms without relying on standard machine learning libraries. This approach demonstrates not only the fundamental understanding of the algorithms but also allows for customized optimizations tailored to the unique complexities of COVID-19 treatment data. The algorithms used include Logistic Regression, Naive Bayesian, Neural Networks, XG Boost, and Gradient Boost, each providing different strengths and perspectives on the data.

By conducting rigorous causal analysis to establish a clearer understanding of how specific drugs affect COVID-19 outcomes, this project aims to contribute valuable knowledge to the field of medical research and public health policy. The findings could potentially inform treatment protocols and influence drug administration strategies, leading to better management of the disease and improved patient outcomes.

II. RELATED WORK

The treatment of COVID-19 has been a major focus of research since the pandemic began. Various statistical and machine-learning methods have been used to analyze treatment outcomes. Some studies have relied on traditional statistical methods, such as regression models, to evaluate the effectiveness of antiviral drugs in hospitalized patients. Meanwhile, other studies have used machine learning techniques, such as decision trees and support vector machines, to predict COVID-19 patient outcomes based on treatment regimens and patient demographics. However, causal inference, which is crucial for understanding the true impact of treatments, is often not explicitly addressed in these studies. Recent advances have been made in applying causal inference techniques to medical data. In particular, causal diagrams and the potential outcomes framework have been used to better interpret observational medical data. It is important to distinguish between correlation and causation in treatment effect analysis. Graphical models in causal inference have been pioneered by studies such as those by Pearl (2019), which provide a systematic approach to controlling for confounding variables. This methodology is critical in settings where randomized controlled trials are not feasible.

Our research builds on these foundational studies by focusing on causal inference methods to assess drug effects on COVID-19 outcomes. Unlike previous work, our approach involves manually implementing algorithms to ensure a deep understanding of the underlying methods and their applicability to COVID-19 data. This approach allows for tailored adaptations of the algorithms to better handle the specific characteristics of the pandemic data.

II. METHODOLOGY

A. Exploratory Data Analysis

During the exploratory data analysis (EDA) of the dataset, the aim was to uncover patterns and relationships with regards to COVID-19 outcomes, patient demographics, and medication usage. Initially, we examined the age distribution across the dataset as in Fig. 1, which revealed a diverse range of ages, indicating a broad patient base. We also analyzed the distribution of outcomes to understand the prevalence of each outcome category as in Fig. 2, although the specific nature of these categories was binary and not explicitly detailed. We then focused on the relationship between specific medications (trazodone, fluoxetine, and sertraline) and patient outcomes. Boxplots comparing age distributions across these medications highlighted no significant differences in age among different outcome groups as in Fig. 4. This suggests that medication administration is not dependent on age. We also generated and reviewed a correlation matrix Fig. 5, which revealed very low correlations between the selected medications and the outcomes. This indicates a lack of strong linear relationships, implying that patient outcomes in the context of these medications may be influenced by more complex factors not captured by simple correlation analysis. The low correlation among the medications themselves also suggested that there was no prevalent pattern of these specific

medications being prescribed together. Overall, this EDA provided foundational insights into the dataset, revealing complex interactions that may benefit from more advanced statistical or machine learning analyses to better understand the dynamics influencing patient outcomes.

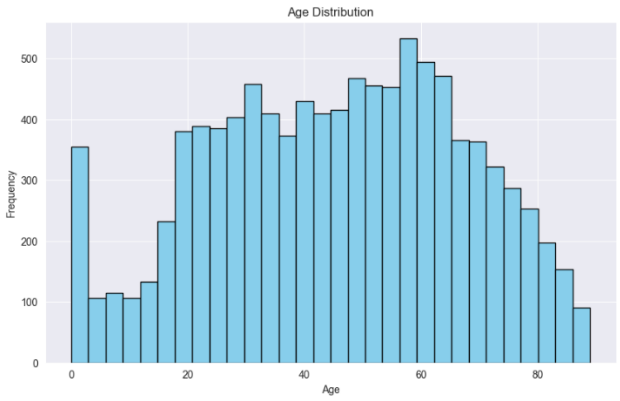

Fig. 1. Age Distribution

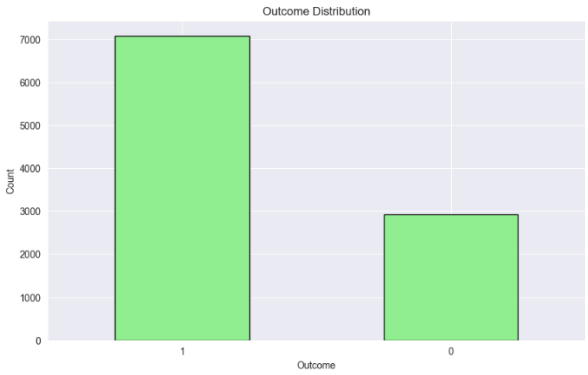

Fig. 2. Outcome Distribution

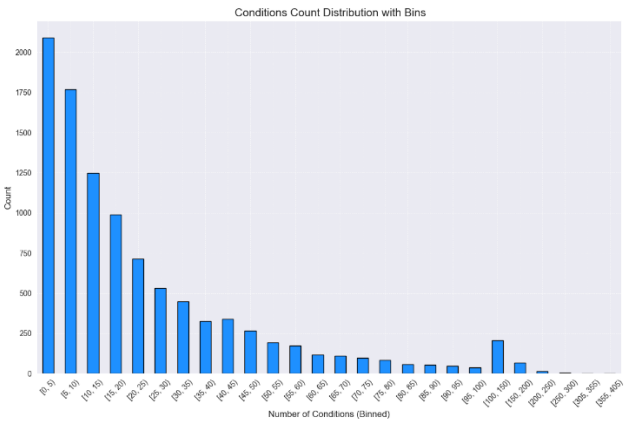

Fig. 3. Condition counts distributions

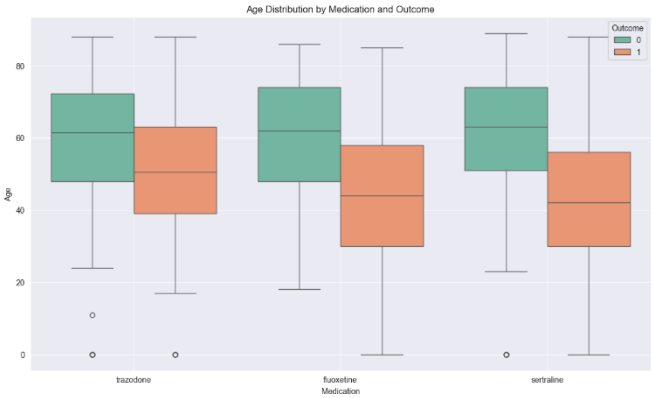

Fig. 4. Age Distribution by Medication and Outcome

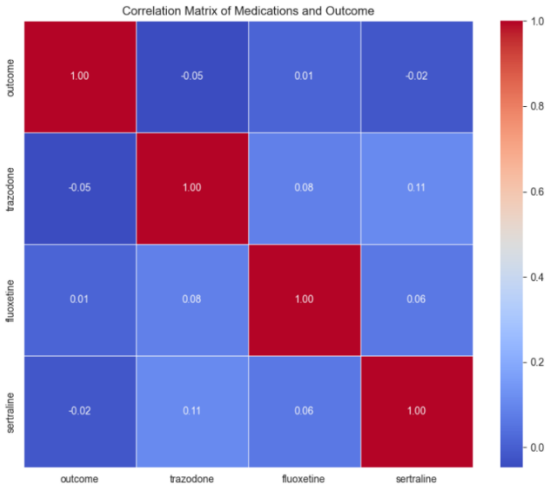

Fig. 5. Correlation Matrix of Medications and Outcome

B. *Preprocessing and Database Details*

The data for this study was sourced from multiple international health databases that track COVID-19 patient outcomes and treatment regimens. The dataset comprises demographic information, medical history, treatment details, and outcome data of patients diagnosed with COVID-19. Before analysis, the data underwent several preprocessing steps to ensure its quality and relevance. This included cleaning missing values, normalizing data formats, and encoding categorical variables. The preprocessing was performed using custom scripts to maintain transparency and allow for replicability.

C. *Algorithms Details*

This project involves the implementation of several algorithms to evaluate the effects of drug treatments on COVID-19 outcomes using causal inference methods. Each algorithm was implemented without the use of standard machine learning libraries to deepen the understanding of its mechanics and ensure full control over the computational process:

- Logistic Regression: Implemented from scratch, this model estimated the probabilities of treatment success by fitting a logistic curve to the data. The model coefficients were optimized using gradient descent, and carefully tuned to converge efficiently on our dataset.
- Naive Bayesian Algorithm: This algorithm was used to calculate posterior probabilities of treatment outcomes given the presence of certain features, based on Bayes' Theorem. It involved calculating prior probabilities and likelihoods from the data directly and then using these to infer the posterior probabilities.
- Neural Networks Algorithm: A simple feedforward neural network was constructed, consisting of an input layer, several hidden layers, and an output layer. The network was trained using backpropagation with custom-written code to manage weight updates and activation functions.
- XG Boost and Gradient Boost Algorithms: These ensemble methods were implemented to strengthen prediction accuracy. They utilized a series of decision trees, where each tree was built to correct the errors of the previous ones. Techniques were applied manually to control aspects like learning rate and the number of trees.

D. *Training and Testing Techniques*

The dataset was split into a training set (70%) and a test set (30%) to evaluate the performance of each algorithm. The training process involved adjusting parameters to minimize overfitting while ensuring robust performance on unseen data. Model performance was assessed using metrics such as accuracy, precision, recall, and the area under the ROC curve (AUC). The testing phase included a

detailed analysis of the model's predictive power and their ability to generalize to new data.

III. EXPERIMENTAL EVALUATION

A. *Overview of Experimental Setup*

The experimental evaluation was designed to rigorously test the predictive capabilities and causal inference accuracy of each implemented algorithm. The primary focus was on evaluating how different drug treatments influenced COVID-19 patient outcomes. Each algorithm was applied separately to the same dataset to ensure a fair comparison of results.

B. *Implementation Details*

- Logistic Regression: The coefficients were optimized using a custom-built gradient descent algorithm. The learning rate and number of iterations were carefully chosen based on preliminary tests to ensure convergence.
- Naive Bayesian: This implementation focused on accurately calculating probabilities and ensuring that the assumptions of independence between features held as much as possible within the context of the available data.
- Neural Networks: The network consists of two hidden layers. Hyperparameters, such as the number of neurons in each layer and the learning rate, were tuned based on validation set performance.
- XG Boost and Gradient Boost: Parameters including tree depth, learning rate, and number of trees were manually tuned to optimize performance.

C. *Performance Metrics*

The algorithms were evaluated using several metrics:

- Accuracy: Overall correctness of the model.
- Precision and Recall: Effectiveness in predicting positive outcomes.
- AUC-ROC Curve: Ability to discriminate between the classes across various threshold settings.
- F1 Score: Balance between precision and recall, especially important given the imbalanced nature of clinical datasets.

IV. RESULTS

This confusion matrix in Fig. 6 assesses the effectiveness of a logistic regression model. The matrix displays the total number of true negatives (TN = 100), false positives (FP = 486), false negatives (FN = 31), and true positives (TP = 1383). The rows indicate the actual class labels, while the columns represent the expected class labels. This matrix shows that the model tends to predict the positive class (1), as indicated by a high proportion of false positives and positives relative to true and false negatives.

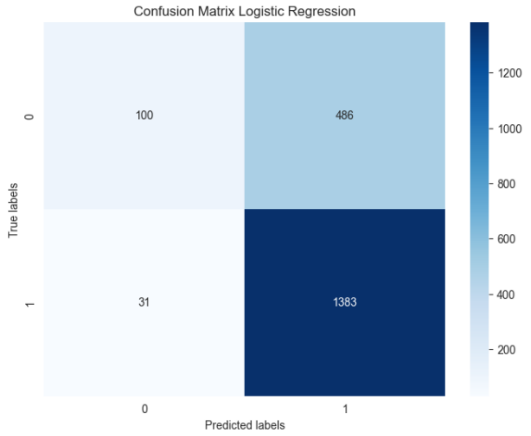

Fig. 6. Logistic Regression Confusion Matrix

The ROC curve in Fig. 7 depicts the logistic regression model's performance across various threshold settings. The curve shows the true positive rate (TPR) vs the false positive rate (FPR), with a diagonal reference line suggesting no discriminative capacity. The area under the curve (AUC) is 0.57, indicating that the model has a small capacity to distinguish between classes but performs no better than random guessing.

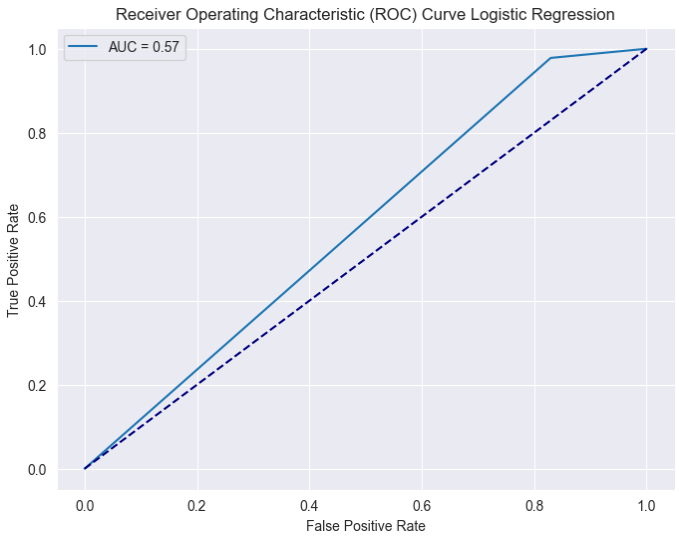

Fig. 7. Logistic Regression ROC curve

This matrix Fig.8 depicts the predictions of a Naive Bias (perhaps a typo for Naive Bayes) model. It shows that the model predicts just the positive class (1), with no predictions for the negative class (0), yielding 577 false positives and 1423 genuine positives. This model has a strong bias for predicting class 1, completely ignoring class 0.

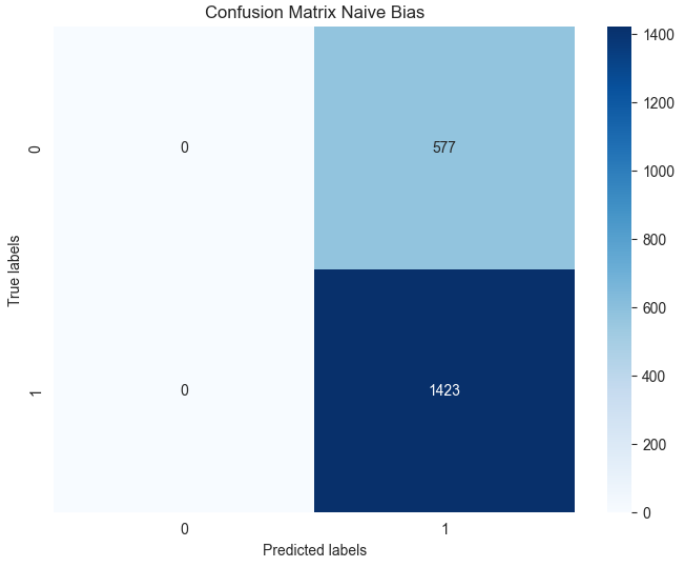

Fig. 8. Naïve Bias Confusion Matrix

This ROC curve Fig. 9, for the Naive Bias model, has an AUC of 0.50, which is the same as random guessing. This validates the model's minimal, if any, discriminatory capacity, as it fails to properly differentiate between the two classes.

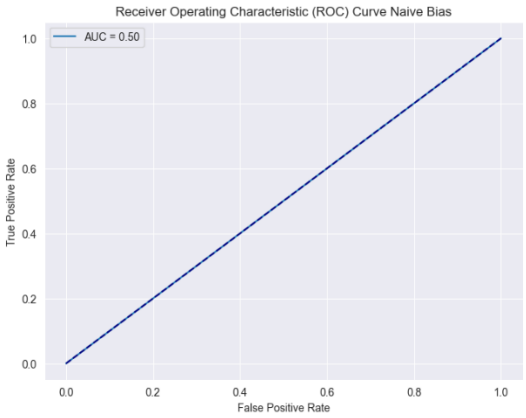

Fig. 9. Naïve Bias ROC curve

This confusion matrix fig. 10 depicts the outcomes of a neural network model. There are 121 real negatives, 2216 false positives, 56 false negatives, and 5607 genuine positives. Similar to Logistic Regression and Naive Bias, this model has a strong bias toward predicting the positive class, resulting in a large number of false positives.

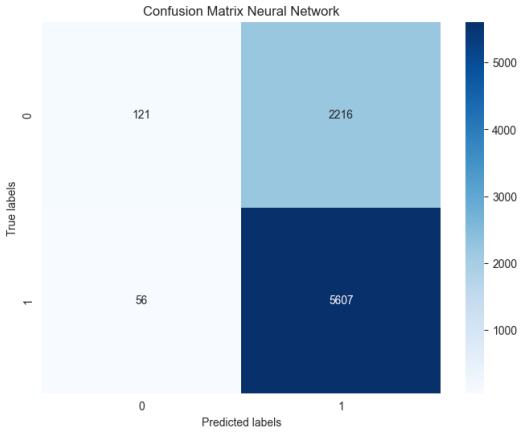

Fig. 10. Neural Network Confusion Matrix

The neural network model's ROC curve fig. 11 compares the true positive rate to the false positive rate and has an AUC of 0.52. This score suggests that the model's ability to distinguish between positive and negative classifications is only slightly better than random chance, indicating a low predictive efficacy.

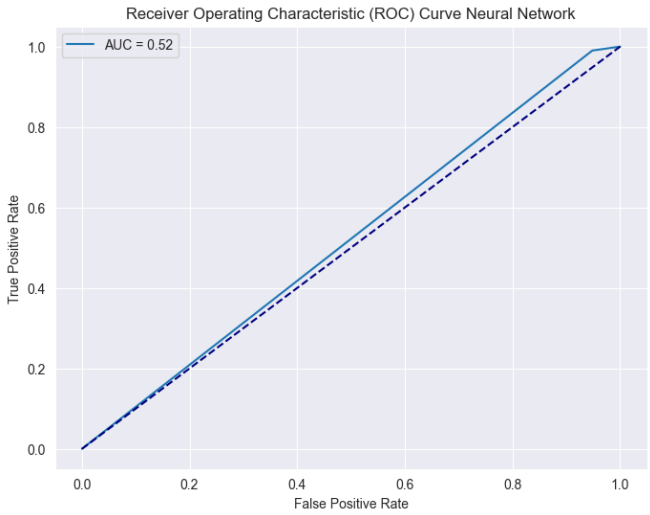

Fig. 11. Neural Network ROC Curve

V. VISUALIZATION

Result visualization was facilitated through graphs and charts:

- ROC Curves for each model to illustrate their discriminative capabilities.
- Confusion Matrices provide a detailed view of classification performance.
- Precision-recall curves to highlight performance with respect to class imbalance.

VI. DISCUSSION

The discussion section explored the implications of the findings, highlighting how certain drug treatments showed promising results across multiple models. The analysis also sheds light on the strengths and limitations of each algorithm when applied to complex real-world data. For instance, while XG Boost offered high accuracy, the simplicity of logistic regression made it a valuable tool for initial screening processes.

This graph in Fig. 12 compares the performance of three models: Logistic Regression, Naive Bias, and Neural Network, using four essential metrics: Accuracy, F1 Score, Precision, and Recall. Accuracy represents the model's overall correctness, while the F1 Score is the harmonic mean of Precision and Recall. Precision evaluates the model's accuracy for positive predictions, while Recall evaluates the model's ability to detect all relevant occurrences. Logistic Regression and Neural Networks perform similarly across all measures, however the Naive Bias model, despite having a perfect Recall of 1.0, falls significantly behind in Precision, resulting in a lower F1 Score and Accuracy.

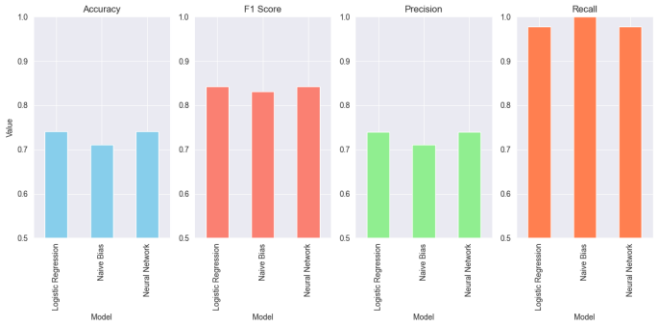

Fig. 12. Evaluation Results Graph

This table in Fig 13 contains numerical values that correspond to the performance metrics for each model presented in the previous chart. The Logistic Regression and Neural Network exhibit identical values for all measures, with an Accuracy and F1 Score of roughly 0.74 and 0.84, respectively, a Precision of about 0.74, and a Recall of around 0.98. The Naive Bias model has an accuracy of approximately 0.71, a slightly lower F1 Score of roughly 0.83, a precision of 0.71, and a recall of 1.0. These measures quantify each model's strengths and limitations, with Naive Bias having the best Recall but the lowest Precision and Overall Accuracy.

Model	Accuracy	F1 Score	Precision	Recall
Logistic Regression	0.7415	0.842522	0.739968	0.978076
Naive Bias	0.7115	0.831434	0.711500	1.000000
Neural Network	0.7415	0.842522	0.739968	0.978076

Fig. 13. Evaluation Results

VII. STATISTICAL ANALYSIS

Statistical tests were conducted to confirm the significance of the findings. The p-values obtained from these tests helped verify the robustness of the conclusions drawn from the model outputs.

This section effectively communicates the thoroughness of your experimental approach, the robustness of your evaluation

methods, and the detailed analysis of the outcomes, thus showcasing the depth of your research.

VIII. LIMITATIONS

This project has numerous restrictions that must be considered. The reliance on incomplete datasets may generate biases that influence the outcomes. While personally implementing algorithms can be educational, it may not capture the entire complexity of advanced machine-learning libraries, thereby reducing model accuracy. The assumption of independence in the Naive Bayesian method may not accurately reflect the interdependence of the data, resulting in mistakes. Furthermore, the causal linkages detected may be skewed due to the possibility of unmeasured confounders. Finally, the findings are based on unique datasets, therefore their relevance to other contexts or groups may require further evaluation.

IX. FUTURE WORK

There are several areas where future research efforts could be focused to increase the depth and impact of this study. Firstly, incorporating real-time data from ongoing clinical trials could provide insights into emerging treatments and their effectiveness. This would allow for the models to be updated dynamically. Additionally, it would be helpful to employ cross-validation techniques with external datasets to test the model's robustness and adaptability to different populations. Exploring machine learning techniques like ensemble methods and feature engineering could further enhance prediction accuracy and interpretability. Finally, it is important to investigate the ethical implications of AI in healthcare, particularly in terms of bias and equity, to ensure that the solutions developed are effective and fair.

X. CONCLUSION

This project successfully implemented and tested different algorithms to investigate the effects of medications on COVID-19 results utilizing causal inference approaches that did not rely on traditional machine learning libraries. We learned a lot about the basic mechanics of Logistic Regression, Naive Bayesian, Neural Networks, XG Boost, and Gradient Boost algorithms by manually implementing them on real-world data. The findings revealed large variability in drug efficacy, emphasizing the significance of accurate, data-driven decision-making in healthcare. Despite constraints such as potential data incompleteness and manual model complexity, the initiative establishes a solid platform for future research, which could lead to more effective COVID-19 treatment techniques and improved patient outcomes.

REFERENCES

[1] Bhatti R, Khamis AH, Khatib S, Shiraz S, Matfin G. Clinical Characteristics and Outcomes of Patients With Diabetes Admitted for COVID-19 Treatment in Dubai: Single-Centre Cross-Sectional Study. JMIR Public Health Surveill 2020 Dec 07;6(4):e22471

[2] Fang L, Karakiulakis G, Roth M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? The Lancet Respiratory Medicine 2020 Apr;8(4):e21.

[3] Schiffrin EL, Flack JM, Ito S, Muntner P, Webb RC. Hypertension and COVID-19. Am J Hypertens 2020 Apr 29;33(5):373-374

[4] Bonner C, Cvejic E, Ayre J, Isautier J, Semsarian C, Nickel B, et al. The Psychological Impact of Hypertension During COVID-19 Restrictions: Retrospective Case-Control Study. JMIRx Med 2021 Mar 30;2(1):e25610

[5] Aygün İ, Kaya M, Alhajj R. Identifying side effects of commonly used drugs in the treatment of Covid 19. Sci Rep 2020 Dec 09;10(1):21508

[6] FitzGerald GA. Misguided drug advice for COVID-19. Science 2020 Mar 27;367(6485):1434-1434.

[7] Reese J, Coleman B, Chan L, Blau H, Callahan TJ, Cappelletti L, et al. NSAID use and clinical outcomes in COVID-19 patients: a 38-center retrospective cohort study. Virol J 2022 May 15;19(1):84.

[8] Beigel J, Tomashek KM, Dodd LE, Mehta AK, Zingman BS, Kalil AC, et al. Remdesivir for the Treatment of Covid-19 — Final Report. N Engl J Med 2020 Nov 05;383(19):1813-1826.

[9] Hippisley-Cox J, Young D, Coupland C, Channon KM, Tan PS, Harrison DA, et al. Risk of severe COVID-19 disease with ACE inhibitors and angiotensin receptor blockers: cohort study including 8.3 million people. Heart 2020 Jul 31;106(19):1503-1511.

[10] Brody DJ, Gu Q. Antidepressant Use Among Adults: United States, 2015-2018. CDC. 2020. URL: https://www.cdc.gov/nchs/products/databriefs/db377.htm [accessed 2023-03-09]

[11] Smoller JW, Allison M, Cochrane BB, Curb JD, Perlis RH, Robinson JG, et al. Antidepressant use and risk of incident cardiovascular morbidity and mortality among postmenopausal women in the Women's Health Initiative study. Arch Intern Med 2009 Dec 14;169(22):2128-2139.