2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) | 979-8-3503-4501-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/CHASE60773.2024.00011

Biomarker Trajectory Prediction and Causal Analysis of the Impact of the Covid-19 Pandemic on CVD Patients using Machine Learning

Trusting Inekwe
Computer Science Dept
Worcester Polytechnic Institute
Worcester, MA 01609
toinekwe@wpi.edu

Winnie Mkandawire
Genomics and Computational Biology
University of Massachusetts Chan Medical School
Worcester, MA 01655
Winnie.Mkandawire@umassmed.edu

Brian Wee
Chemical and Physical Biology
Harvard College
Cambridge, Massachusetts 02138
brianwee@college.harvard.edu

Emmanuel Agu
Computer Science Dept.
Worcester Polytechnic Institute
Worcester, MA 01609
emmanuel@wpi.edu

Andrés Colubri
Genomics and Computational Biology
University of Massachusetts Chan Medical School
Worcester, MA 01655
Andres.Colubri@umassmed.edu

***Abstract*—Background: The COVID-19 pandemic disrupted healthcare services, increasing the susceptibility of high-risk patients including those with cardiovascular Diseases (CVDs), to adverse outcomes. Biomarkers provide insights into patients' underlying health status. However, few studies have investigated the effects of the COVID-19 pandemic on CVD biomarker trajectories using predictive modeling and causal analyses frameworks. Prior research explored the impacts of the COVID-19 pandemic on CVD severity and prognosis but did not investigate biomarker trajectories using Machine Learning (ML), which can discover complex multivariate relationships in multi-modal data.**

Objective: This study aimed to compare six ML regression models to select the best performing models for predicting biomarker trajectories in CVD patients using retrospective data. Subsequently, these models were used to assess the COVID-19 pandemic's impact on CVD patients and for causal analyses

Approach: Using ML regression and causal inference, this study investigated the pandemic's impact on biomarker values of 80,917 CVD patients and 77,332 non-CVD controls, treated at two hospitals in Central Massachusetts between May 2018 and December 2021. ML regression algorithms, including Neural Networks (NN), Decision Trees (DT), Random Forests (RF), XGBoost, CATBoost and ADABoost, were trained and compared. Important CVD biomarkers (HbA1c, LDL cholesterol, BMI, and BP) were predicted as outcome variables with patients' risk factors (age, race, gender, socioeconomic status) as input variables. Shapley feature importance analyses identified the most predictive features, which were then utilized in Causal Analysis. A Difference-in-Differences (DID) approach within a Double/Debiased Machine Learning (DML) method isolated the pandemic's impact on biomarkers, while minimizing the effects of confounding factors.

Results: CATBoost and XGBoost were the most predictive ML models for LDL cholesterol and HbA1c, yielding R^2 values of 0.13 and 0.10, respectively. RF outperformed other models for BMI and BP, achieving R^2 values of 0.192 and 0.071. The small R^2 values were due to the prevalence of categorical features in the data with substantial variation in biomarker values. Feature importance analysis determined age, socioeconomic status, and race/ethnicity to be important drivers of biomarker changes, highlighting the role of social determinants of health. DML with DID analysis revealed a statistically significant increase (p-value $<$0.05) in BMI and systolic BP values for CVD patients during the COVID-19 pandemic compared to the control group, their HbA1c and LDL cholesterol values actually improved during the pandemic, suggesting differential effects of the pandemic on key CVD biomarkers.

Conclusion: Our proposed ML biomarker prediction models can facilitate personalized interventions and advance risk assessment for CVD patients. The predictive importance of factors such as age, socioeconomic status, and race highlights the need to address health disparities.

***Index Terms*—Keywords: COVID-19, Cardiovascular Diseases, biomarkers, machine learning, regression, predictive modeling, causal inference, Difference-in-Differences, SHAP value, Double/Debiased machine learning.**

I. Introduction

Motivation: The COVID-19 pandemic led a global healthcare crisis, disproportionately affecting individuals with chronic conditions such as cardiovascular diseases (CVDs), respiratory conditions, diabetes, immunocompromised states, obesity, and neurological disorders [1]. Such high-risk patients with these chronic diseases faced an increased risk of contracting the virus, and of succumbing to the infection [1]. Surges in COVID-19 infection rates strained hospitals, leading to delays, cancellations and disruptions of healthcare services and routine medical care of patients [2]. Patients with CVD, a leading cause of mortality [3], were especially vulnerable due to their need for regular medical appointments, care and treatment. Consequently, disruptions caused by COVID-19 raised concerns about potential deterioration of their health.

The COVID-19 pandemic may have caused deterioration of the health of CVD patients directly due to changes in

2832-2975/24/$31.00 ©2024 IEEE
DOI 10.1109/CHASE60773.2024.00011

lifestyle, reduced physical activity and unhealthy diet [4], and indirectly by preventive public health measures to curb disease spread. Lockdowns, social distancing, mask-wearing mandates, remote work policies and isolation recommendations disrupted healthcare access. Additionally, despite requiring critical medical attention, many patients were unable or reluctant to seek care due to apprehensions about potential exposure to SARS-CoV-2 [5]. Patients also faced delays in receiving healthcare during in-person routine check-ups, screenings and follow-up treatment appointments [6]. For CVD patients, such disruptions could increase the likelihood of heart attacks, strokes, heart failure and, ultimately, death [7].

Knowledge gap our work addresses: Despite extensive research elucidating the intricate interplay between health deterioration of CVD patients and the COVID-19 pandemic [4], a crucial knowledge gap persists in the lack rigorous computational approaches to quantifying the pandemic's impact on CVD patients' biomarkers—measurable indicators of an individual's health status. Biomarkers can either be measured directly from body fluids or indirectly from physical manifestations in physical examination (vital signs).

CVD biomarkers our study focuses on: The values of specific biomarkers such as (Systolic) Blood Pressure (BP), Low Density Lipoprotein (LDL) cholesterol, glycated hemoglobin (HbA1c), and Body Mass Index (BMI), provide valuable insights into the health status of CVD patients. Regular testing and monitoring of these biomarkers enable healthcare professionals to evaluate CVD progression, identify potential risks, and guide treatment [8]. Values of these biomarkers outside normal ranges have been directly linked to poor COVID-19 outcomes, especially in CVD patients [9]. For instance, elevated LDL cholesterol levels may indicate an increased risk of atherosclerosis and coronary artery disease [8]. BMI, a measure of body weight in relation to height, provides insights into obesity-related CVD risks. Elevated BP readings signal potential hypertension and increased stress on the cardiovascular system [10]. Elevated levels of HbA1c, a biomarker that reflects long-term blood sugar control may increase cardiovascular risk [11]. Prior research focused on investigating the severity and prognosis of COVID-19 among CVD patients using various biomarkers, including LDL cholesterol, BP, HbA1c, and BMI [8], [9]. While studies have demonstrated the heightened susceptibility of CVD patients to morbidity and complications caused by COVID-19, the impact of the disrupted healthcare systems, lockdowns and social isolation that were caused by the pandemic has not been investigated. Specifically, there has been no prior work that utilized computational approaches to predict trajectories and perform causal analyses of CVD patient biomarkers during the Covid-19 pandemic.

Our approach: This study investigates how the COVID-19 impacted four CVD patient biomarkers: BMI, LDL cholesterol, BP, and HbA1c. We combine ML regression with Double/debiased Machine Learning (DML) [12] for Difference-in-Differences (DID) approach [13](DMLDID)– a causal inference method proposed in econometrics [14], on the biomarker values of 80,917 CVD patients (experimental group) treated at hospital sites throughout central Massachusetts between May 2018 and June 20 2021 and 77,332 patients without CVD (control group). Since the World Health Organization (WHO) declared Covid-19 a pandemic on March 10, 2020 [15], the selected timeframe spans pre- and post-COVID-19 onset with statewide lockdowns. By comparing health biomarker values of a CVD group with controls, DMLDID effectively isolates the pandemic's distinct effect from confounding factors to generate robust estimates of the pandemic effect on CVD patients. Additionally, Tree-based Machine Learning (ML) regression algorithms were investigated for biomarker value prediction including the DT, RF, Extreme Gradient Boosting (XGBoost), Adaptive Boosting Regression (AdaBoost) and CatBoost, and Neural network. These ML models were used to predict HbA1c, LDL cholesterol, BMI, and BP CVD biomarker values from patients' risk factors (age, race, gender, socioeconomic status) as input variables. Beyond the capabilities of traditional statistical methods, ML can discover complex, non-linear, multivariate relationships between the input variables and biomarker values.

Novelty of our work: Prior work focused on predicting various COVID-19 health outcomes, including diagnosis [10], severity [16], and mortality risk [17]. Some of this research investigated chronic disease patients as a group that included CVD patients [18] and CVD patients exclusively [19]. While these studies demonstrate the ability of ML to predict the pandemic's impact on CVD patients using biomarkers as feature inputs, there has been no work that performed prediction and causal analyses of how the pandemic affected CVD patient biomarker. Our work addresses this gap, with the hypothesis that the pandemic resulted in significant deterioration in the values of four CVD biomarkers (BMI, BP, LDL cholesterol, and HbA1c) in CVD patients compared to controls. Finally, we innovatively adapt DMLDID, to minimize confounding and generate robust estimates of the disruptive effects of COVID-19 and preventive measures such as lockdowns on CVD patient biomarkers.

Research Questions: addressed are:

1) How accurately can ML regression predict pandemic-induced changes in key CVD biomarker values?
2) What were the most important features for predicting CVD patient biomarker values? How predictive are social determinants of health, age, race, gender and socioeconomic status?
3) How effective is the DMLDID causal analyses method in isolating confounding effects to robustly predict pandemic-induced biomarker changes?

Contributions: of our study are:

1) *Proposing an ML-based regression methodology to predict CVD patient biomarker trajectories:* to discover complex, non-linear, multivariate relationships between input features including, age, race and socioeconomic status, and four key CVD biomarkers (HbA1c, LDL cholesterol, BMI, and BP) as output features.

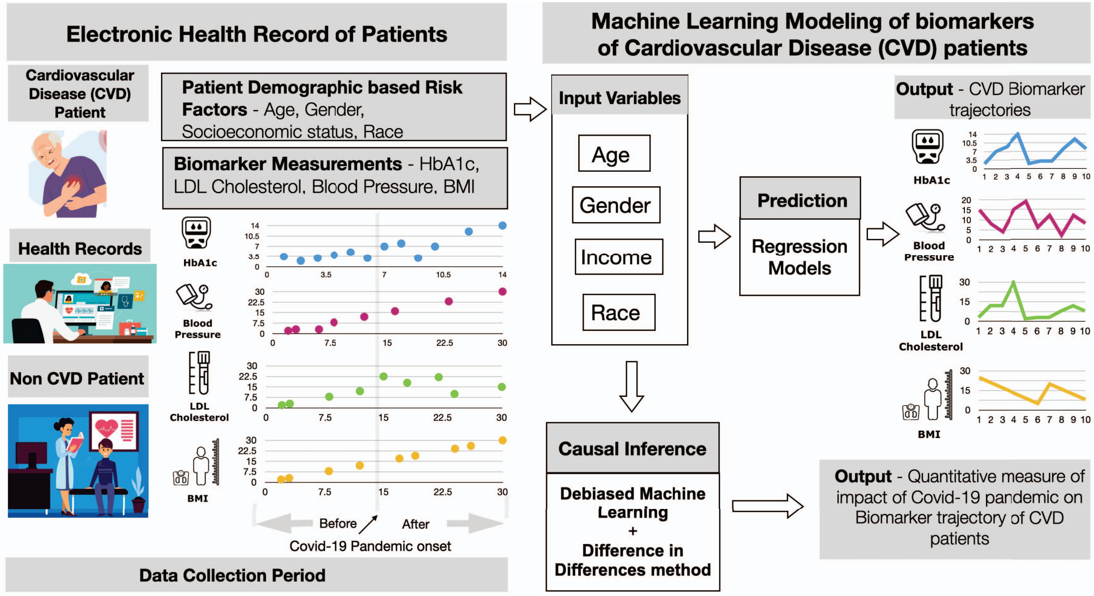

Electronic Health Record of Patients
Cardiovascular Disease (CVD) Patient
Patient Demographic based Risk Factors - Age, Gender, Socioeconomic status, Race
Biomarker Measurements - HbA1c, LDL Cholesterol, Blood Pressure, BMI
Health Records
Non CVD Patient
HbA1c
Blood Pressure
LDL Cholesterol
BMI
Before
After
Covid-19 Pandemic onset
Data Collection Period
Machine Learning Modeling of biomarkers of Cardiovascular Disease (CVD) patients
Input Variables
Age
Gender
Income
Race
Prediction
Regression Models
Output - CVD Biomarker trajectories
Causal Inference
Debiased Machine Learning + Difference in Differences method
Output - Quantitative measure of impact of Covid-19 pandemic on Biomarker trajectory of CVD patients

Fig. 1: Overview of our retrospective data collection and Machine Learning (ML-based) modelling approach for prediction and causal analyses of the impact of the COVID-19 pandemic on the biomarkers of patients with cardiovascular diseases (CVDs).

2) *Identification of the most predictive Features Influencing CVD patient Biomarkers:* Shapley values generated by ML regression using tree-based methods revealed that beyond biological factors, age, socioeconomic status, and race/ethnicity were key drivers highlighting the importance of social determinants of health.
3) *Proposing the DMLDID causal analysis for CVD biomarker prediction:* This innovative approach combines traditional statistical methods with ML to isolate COVID-19 specific effects and advance CVD biomarker prediction. Our study provides evidence of the differential impacts of the COVID-19 pandemic on biomarker values in CVD patients compared to controls.
4) *Rigorous evaluation on a first-of-a-kind large retrospective dataset gathered before and during Covid-19:* which included 80,917 CVD patients and 77,332 non-CVD controls, treated between May 2018 and December 2021. Evaluation on this rare, large dataset suggests that our findings are robust. ML regression analysis reveal CATBoost and XGBoost as the top models for predicting LDL cholesterol and HbA1c (with R^2 values of 0.13 and 0.10), while RF for BMI and BP (with R^2 values of 0.192 and 0.071). Causal analysis using DMLDID revealed significant increases ($p < 0.05$) in BMI and systolic BP for CVD patients during the COVID-19 pandemic compared to non-CVD individuals.

Significance: This study's findings are relevant for Artificial Intelligence (AI) and healthcare professionals due to its valuable insights of COVID-19 on CVD patient biomarkers. Moreover, it proposes a cutting-edge ML-based methodology for rigorously assessing the impacts of the COVID-19 pandemic on CVD patients.

II. Related Work

A. COVID-19 Machine Learning Prediction using Biomarkers

Prior work has leveraged ML to investigate COVID-19 pandemic's effects on patients in general, using their biomarkers as input features for tasks including predicting their health deterioration [20] and mortality risks [21]. In contrast, our study focuses on predicting the values of CVD patient biomarkers during COVID-19 (as target values).

B. ML Research on Predicting COVID-19's Impact on Chronic Disease Patients as a Group that included CVD Patients

ML has been leveraged to investigate COVID-19's psychological impacts on chronic disease (including CVDs) patients [22], predict mortality and COVID-19 severity risk [23]. Barría-Sandoval [23] explored various ML algorithms to predict mortality in 57,623 chronic disease patients, including CVD. The XGBoost, RF, and Multinomial Regression models performed best achieving AUCROC values of 0.624, 0.600, and 0.627 and accuracies of 0.642, 0.641, and 0.633, respectively. Age and place of death were key predictors of mortality.

C. ML Work on Predicting the Impact of COVID-19 on CVD Patients Specifically

These include predicting the effects of COVID-19 on CVD patients while highlighting the role of ethnicity [24], as well as mortality risks in CVD patients with COVID-19 [19]. Castelnuovo et al. [19] used an RF model to predict CVD risk and in-hospital mortality in 3,894 COVID-19 patients. Input features included age, gender, obesity, smoking, and chronic comorbidities such as hypertension. Results identified impaired renal function, elevated C-reactive protein levels, and advanced age as significant mortality risk factors. The RF model achieved 95.2% sensitivity, 30.8% specificity, 83.4% overall accuracy, and an F1 score of 90.4%. In conclusion, while the work reviewed above demonstrated the utility of ML in investigating the impact of the COVID-19 pandemic on at-risk individuals, no prior work has explored ML approaches to predict changes in CVD patient biomarkers during the pandemic. Moreover, our methodology goes beyond ML prediction, exploring causal inference using DMLDID to obviate the effects of confounding factors.

D. *Causal Inference and the DMLDID*

Measuring the impact of interventions and causal inference are important in the field of econometrics. Recently, ML-based methods have been proposed for analyzing causal relationships in data including deep learning models for estimating treatment effects [25], use of the causal forest (that extends random forests) to estimate treatment effects within subgroups [26], and employing the DMLDID method to estimate average treatment effects in a population [14]. The DMLDID methodology addresses confounding factors, minimizes bias and model mispecification to improve estimates of treatment effects [14]. It has been applied to diverse tasks, including, assessing the social impacts of the London Night Tube [27] and estimating the impact of carbon policy on corporate risk-taking [28]. In our study, we innovatively adapt the DMLDID method to the field of medicine and specifically to investigate the average treatment/pandemic and disruptions in care caused by preventive measures, on patients with CVD.

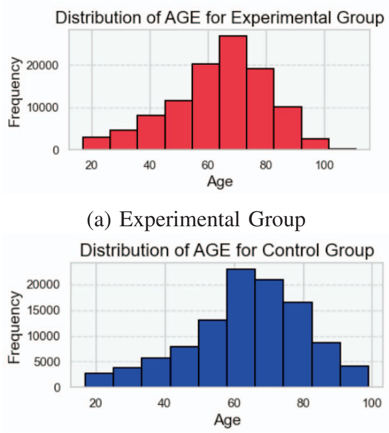
III. Methodology

A. *Overview of Methodology*

Figure 1 is an overview of our ML-based approach for modeling the impact of the COVID-19 pandemic and preventive measures such as lockdowns on biomarkers of CVD patients. *Data Collection:* Electronic Health Records (EHR) of 426,022 patients who were treated at the University of Massachusetts Chan Medical School and Memorial hospital retrieved retrospectively and processed from which 80,917 CVD and 77,332 non CVD patients were selected for our study. Key health biomarkers including HbA1c and LDL_cholesterol, BP and BMI where analyzed in two steps. First, various ML regression models were used to predicted CVD patients biomarker values from their age, gender, income and race. Next, the impact of the pandemic on CVD patients were investigated by comparing their biomarker values to those of non-CVD patients before and during the pandemic. The demographic distribution of patients are described in the Data pre-processing section.

B. *Data Pre-processing*

The 10th edition of the International Classification of Diseases (ICD-10) was used to select patients diagnosed with CVD using ICD-10 codes I25, I48, I50, I63, I65, I67, I73 [29].



(a) Experimental Group

(b) Control group

Fig. 2: Age distribution for Experimental and Control Groups

Next, the MatchIt nearest neighbor matching algorithm developed by Ho et al [30] was used to select a representative sample of CVD and non CVD patients for analyses. The MatchIt algorithms was applied to a combined dataset of demographic and control group data, prioritizing patients' age and gender. MatchIt calculates a propensity score (PS) for each patient based on covariate values to create a balanced and representative group from treatment and control group by prioritizing selected covariates. PS calculates the estimated probability of receiving a treatment based on covariates for each observation(patient). This is typically done using logistic regression (equation 1):

$$PS(i) = P(D = 1|X) = \frac{\exp^{\beta_0+\beta_1 X_1+\beta_2 X_2+..+\beta_p X_p}}{1 + \exp^{\beta_0+\beta_1 X_1+\beta_2 X_2+..+\beta_p X_p}} \tag{1}$$

where i is a patient, D is the treatment indicator, X is the covariates and $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_p X_p$ is a linear regression for observed covariates. In the "nearest neighbour" algorithm which we used, each CVD patient was paired with a control patient with the most similar propensity score. The final representative sample used in our analyses, had 80,917 patients with CVD (experimental group), and 77,332 non-CVD patients (control group). Patients in the CVD group had an average age of 70 years with 49.75% males and 50.25% females. The racial composition included 4% Asian, 88.9% White, 6.6% Black/African American, 0.07% Native Hawaiian and 0.3% American Indian or Alaska Native. In the control group, the average age was 65 years, with 48.45% males and 51.55% females. The racial composition included 3% Asian, 92% White, 4.5% Black/African American, 0.05% Native Hawaiian and 0.24% American Indian or Alaska Native. Histograms of the age distributions of both groups after applying MatchIt are shown in Figure 2.

C. *Feature Extraction and Engineering*

Next, our four target CVD biomarkers (HbA1c, LDL cholesterol, BP, and BMI) and input features were extracted. Based on prior evidence in the literature [31], socioeconomic status was included as a feature. Each patient's socioeconomic status was determined as the median income of households in their zip codes based on publicly available information [32]. Age, race, gender and income were input variables. All categorical features were encoded using One-Hot encoding. The dataset was cleaned by removing null values, outliers and medically impossible for each feature.

Finally, we defined a timeframe that spanned pre-, during and post- COVID-19 periods. Measurements taken before January 1st, 2020, were considered pre-pandemic baseline values. March 10, 2020 was considered the beginning of the COVID-19 pandemic. Also, as biomarkers such as HbA1c and LDL cholesterol may take up to two months to change after an altering event, pandemic-influenced biomarker data analyses started on June 10, 2020 (three months after the pandemic started) and spanned one year that ended on June 10th, 2021. This post-pandemic data was compared to the pre-pandemic baseline. Table I presents a description of each target biomarker as well as input variables explored in our analyses. After data pre-processing and feature extraction, our analyses dataset comprised of categorical input variables (age, gender, race, income) along with a continuous target variable/biomarker. To better understand biomarker value distributions, Exploratory Data Analyses was done on each biomarker. Pearson correlation (PC) [33] calculations revealed a non-linear relationship between the target and predictor variables. For instance, for BMI in Fig 3. This finding led us to select non-parametric ML prediction models. PC is calculated as:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \tag{2}$$

where $PC = r = 1, -1, 0$ for positive, negative and no correlation respectively, X_i and Y_i are individual points, $\bar{X}$ and $\bar{Y}$ are the means of the X and Y dataset.

D. Machine Learning Regression Models for Predicting the Impacts of the Pandemic on CVD Patient Biomarker Values

ML models used to predict the values of CVD-relevant biomarkers from input variables included a Feedforward Neural Network model, DT, RF, Extreme Gradient Boosting (XGBoost), AdaBoost and CatBoost. After model training and tuning, final model parameters values are shown in table II. Additional details on each ML model are now provided.

Machine Learning Models:

1) Decision Tree Regression: Decision trees (DT) ML models have a hierarchical tree structure in which each internal node represents a decision based on a feature, and each leaf node represents an outcome [34]. To make predictions, traversal of the decision tree commences at its root and ends at a leaf node. Based on our goal of performing ML regression of continuous target biomarker values, Classification and Regression Trees (CART) algorithms are utilized [35]. Additional details on the DT are provided in Appendix (VII-1).

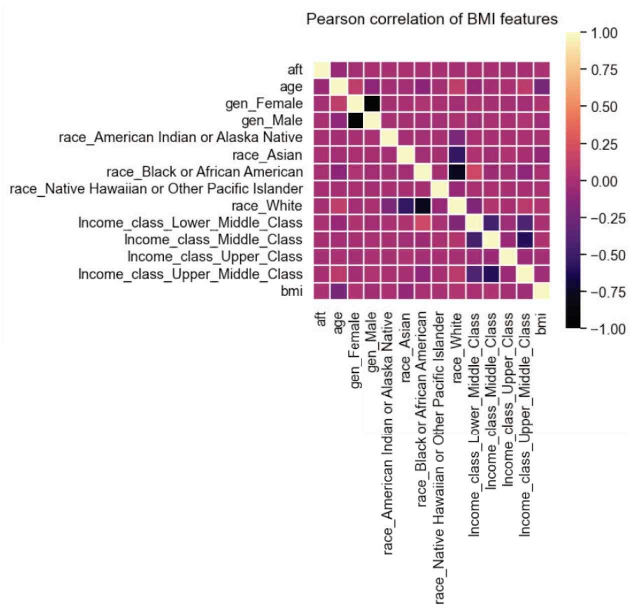


Fig. 3: Pearson Correlation for all features for BMI dataset

2) Random Forest (RF) Regression: A RF model is an ensemble learning method that constructs a collection of DTs. The predictions of individual trees are averaged to yield a final prediction. RF can be used for both classification and regression. Appendix VII-2 contains additional RF algorithm details.

3) Extreme Gradient Boosting (XGBoost): XGBoost is an ML algorithm based on Gradient-Boosted DTs (GBDT) [36]. It combines regression trees with gradient boosting. To optimize the objective function and mitigate overfitting, in each training step, XGBoost feeds back residuals from a base classifier into the subsequent classifier. The combined contributions of all the trees yields the final prediction output, expressed as: $\hat{y} = \sum_{k=1}^{K} gk(X_i)$, where $\hat{y}$ is the predicted target CVD biomarker variable. The ensemble consists of K decision trees, and $gk(X_i)$ represents the output of the k^{th} tree for input features X_i. The appendix VII-3 contains more on XGBoost.

4) Artificial Neural Network (ANN): An ANN is a neural networks model inspired by the human brain's structure [37]. It consists of interconnected nodes/neurons that receive, process, and transmit signals. Each neuron receives one or multiple input signals, assigns them weights, and applies a non-linear activation function to produce an output. ANNs operate in two phases: feed-forward and back-propagation. In this study, a feed-forward neural network trained using backpropagation was used, expressed in equation 3.

$$Z_j^{(i)} = \sum_{k=1}^{n} W_{jk}^{(i)} \cdot A_k^{(i-1)} + b_j^{(i)} \tag{3}$$

where $Z_j^{(i)}$ is the weighted sum of inputs for neuron j in layer i, W_{jk} denotes the weights between neuron k in layer $(i-1)$ and neuron j in layer i. $A_k^{(i-1)}$ is the output (activation

TABLE I: A list of features (target and input variables) used in our dataset, their description, and range used in this study

Variable/Feature	Description	Ranges of Variables/Features used after outlier removal
Target Biomarker Variables		
Glycated Hemoglobin (HbA1c)	A blood test that gives information about a person's average blood sugar levels over the past 2-3 months	> 4 and $\leq 14\%$
LDL Cholesterol	A type of lipoprotein that carries cholesterol in the blood	> 20 mg/dL and ≤ 370
Blood Pressure (BP)	A measure of the force of blood against the walls of the arteries as the heart pumps it around the body (Systolic and Diastolic)	Diastolic: ≥ 1 and ≤ 243 and Systolic: ≥ 1 and ≤ 469
Body Mass Index (BMI)	A numerical measure of a person's body weight in relation to their height	≥ 15 and ≤ 100
Input Variables		
Social Economic Status (Income)	An individual or a family's position within a social and economic hierarchy	Middle class, lower middle class, upper middle class, and upper class
Age	Age distribution used in the dataset	20-39, 40-59, 60-79, ...
Race	Race distribution used in the dataset	American Indian or Alaska native, Asian, Black or African American, Native or Other Pacific Islander, White
Gender	Male = 1, Female = 0	-
Aft	The time period before and during the COVID	Aft = 1 if during the COVID pandemic (after March 10, 2020) and Aft = 0 if before the pandemic (before January 2020)
Exp	The group a patient belongs to	Exp = 1 if patient has CVD and Exp = 0 if patient has does not have CVD

TABLE II: Model parameters and Python libraries used to build models.

Model	Python Libraries Used	Parameters
Neural Network	TensorFlow	hidden_layer_sizes: 100, max_iter: 1000, random_state: 42
Decision Tree	scikit-learn	Max_depth: 1000, random_state: 42
Random Forest	scikit-learn	n_estimators: 1000, random_state: 42
XGBoost	XGBoost	n_estimators: 300, random_state: 42
AdaBoost	scikit-learn	n_estimators: 300, learning_rate: 0.1, random_state: 42
CatBoost	CatBoost	n_estimators: 300, learning_rate: 0.1, depth: 6, l2_leaf_reg: 1, random_state: 42

function) of neuron k in layer $(i-1)$, n is the count of variables, $b_j^{(i)}$ is the bias term for neuron j in layer i. The appendix VII-4 contains more details on the ANN.

5) Adaptive Boosting (AdaBoost) Regression: : AdaBoost is an ensemble learning technique for either ML classification or regression. A boosting algorithm is used to regress features and target variables [38]. The boosting process assigns greater weights to erroneously classified instances with the objective of enhancing predictive performance. Multiple weak learners are aggregated, expressed as $w_i^{(k+1)} = w_i^k \beta(1 - l_k)$. Consider a dataset $S = (x_1, y_1), \ldots, (x_n, y_n)$, where n represents pairs of observations. Each observation is assigned a weight w_i. The likelihood of including an observation in the training set at iteration k is determined for each i based on its assigned weight. The average loss l_k for model k across all observations i is calculated as the weighted sum of these probabilities. The appendix VII-5 contains more details on AdaBoost.

6) CatBoost Regression: CatBoost is an ML algorithm designed for gradient boosting on DTs [39]. It builds an ensemble of sequential DTs with each tree trying to correct the errors of preceding trees, which results in a strong predictive model. It performs best on tasks with categorical features but can also handle numerical features.

Consider the dataset $S = \{(x_1, y_1)_i\}_{i=1}^N$, where $X_i = (X_{i,1}, \ldots, X_{i,M}), X_m = (X_{m,1}, \ldots, X_{m,N})$ is a vector comprising a mix of categorical and numerical features where m is the total number of features. The associated label is $Y_i \in \mathbb{R}$. Initially, the dataset undergoes random permutation, which results in overfitting. Subsequently, the mean label value is computed for samples in the same category within the permutation using the value immediately preceding the current one. This permutation is expressed as:

$$\frac{\sum_{l=1}^{p-1} [x_{\epsilon l,j} = x_{\epsilon p,j}] \cdot Y_{\epsilon l} + \alpha \cdot P}{\sum_{l=1}^{p-1} [x_{\epsilon l,j} = x_{\epsilon p,j}] + \alpha} \tag{4}$$

$\epsilon = \epsilon_1, \epsilon_2, \ldots, \epsilon_N$. The observation $x_{i,k}$, represents an element in a feature matrix or dataset, which is reordered and substituted with $x_{\epsilon p,j}$. The estimation of $x_{\epsilon p,j}$ follows the procedure outlined in equation 4. The prior value is represented by P, and its associated weight is characterized by the parameter α, where $\alpha > 0$. The prior P is the mean label value for regression. P and α reduce noise caused by categories that have few entries. Appendix VII-6 has more details of catboost.

E. Causal Inference Approach to Estimate the Effect of the Pandemic on Patients with CVD

We combined Double/Debiased Machine Learning (DML) with Differences in Differences (DID) [14] or DMLDID, to estimate the impact of COVID-19 pandemic on CVD patients. DML performs causal inference (ATT) on our non-parametric dataset [12], while DID compares changes between CVD patient and control groups before and during COVID-19 [40].

We analyzed a dataset with N patient observations, where each observation is indexed as $i = 1, \ldots, N$, and covers two time periods; before and during the pandemic, denoted as $t = 0, 1$, respectively. Let Y_{it} represent the observed outcome/biomarker for patient i at time t, and D_{it} be the treatment indicator at time t. In the pre-treatment period, all patients have $D_{i0} = 0$, and thus we can simplify the notation by writing $D_i = D_{i1}$. In our case, $D_i = 1$ if patient i has CVD during the pandemic, and $D_i = 0$ otherwise for patient i. Let $Y_{it}(1)$ be the potential outcomes for patient i at time t if the patient has CVD during the pandemic and $Y_{it}(0)$ otherwise. A challenge with causal inference is that $Y_{it}(1)$ and $Y_{it}(0)$ cannot be observed both for the same patient. Instead, $Y_{i1} = D_i Y_{i1} + (1 - D_i) Y_{i1}(0)$ is observed in the post-treatment period. Put simply, we only $Y_{i1}(1)$ can be accessed for the treated units (CVD patients) and $Y_{i1}(0)$ for the controls. The goal is to estimate the average treatment effect on the treated (ATT) (equation 5).

$$\theta = E[Y_{i1}(1) - Y_{i1}(0) | D_i = 1] \tag{5}$$

However, simply comparing outcomes before and after treatment for treated units can introduce bias by neglecting temporal trends. Also, post-treatment differences between treated and control groups can also be confounded. DID considers time-invariant unobserved confounding and temporal trends by using data from both treated and control groups in both pre- and post-treatment periods.

1) The traditional linear DID model: is expressed in equation 6 [27]. Parameters to be estimated include: μ, α, D_i, δ and τ, and ϵ_{it} is the error term. Y_{it} is the outcome for patient i at time t, D_{it} is a binary treatment indicator that equals 1 for the treated group and 0 for the control group. The coefficient τ captures the treatment effect and is the key parameter of interest in the traditional DID model.

$$Y_{it} = \mu + \alpha * D_i + \delta * t + \tau * D_{it} + \epsilon it \tag{6}$$

To select a treatment indicator for DID (D_{it}), a novel variable, *Exp* x *Aft* was created to capture the interaction between patient group and the timing of the pandemic. *Exp* = 1 or 0 for the CVD patient experimental group or control group respectively, while *Aft* = 1 or 0 for during or before the pandemic respectively. When both *Exp* (group) and *Aft* (pandemic period) equal 1, *Exp* x *Aft* is set to 1 ($D_{it} = 1$); otherwise, it is assigned a value of 0 (see Table 1). Thus, our treatment group (D_{it}) is defined as patients with CVD during the pandemic (*Exp* = 1 and *Aft* = 1). Other combinations of *Exp* and *Aft* are our control group because they do not fit the criteria of our treatment group. The *Exp* x *Aft* variable is replaced with the treatment indicator (D_{it}) in subsequent sections. DID makes the “parallel trend assumption”, which asserts that without the treatment, treated and control groups would have shown similar and parallel trends in their average outcomes over time (equation 7).

$$\mathbb{E}[Y_{i1}(0) - Y_{i0}(0) | D_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0) | D_i = 0] \tag{7}$$

However, in practice, other factors may have influence, leading to deviations from a parallel trend. Covariates are often added to account for this. However, the traditional DID model tends to perform sub-optimally with high-dimensional confounding factors and variable treatment effects. Model mis-specification is another common challenge, as the relationship between covariates and outcomes may not be linear. Misrepresentation can result in biased and unreliable estimates of the pandemic’s impact. Therefore, given the non-linear nature of our covariate-outcome relationship, we adopted the DMLDID [14], which combines DML [12] with DID [40] to estimate the causal impact of the pandemic on CVD patient biomarker values.

2) DML-based DID estimator: : As described by [27], calculating the ATT according to Abadie [40] in repeated outcomes (longitudinal) data where $\{Y_{i0}, Y_{i1}, D_i, X_i\}$ is observed (simplified to without subscript i $\{Y_0, Y_1, D, X\}$), requires two assumptions to be met; one, the conditional parallel trend which states that given specific covariates, the treated and control groups would have followed similar trends over time without the treatment

$$\mathbb{E}[Y_1(0) - Y_0(0) | X, D = 1] = \mathbb{E}[Y_1(0) - Y_0(0) | X, D = 0] \tag{8}$$

Secondly, the propensity scores for treated units must be within the range of propensity scores for control units.

$$P(D = 1) > 0 \quad \text{and} \quad P(D = 1 | X) < 1 \tag{9}$$

Thus, the ATT (θ_0) can be conceptualized as:

$$\theta_0 = \mathbb{E}[\frac{Y_1 - Y_0}{P(D = 1)} * \frac{D - P(D = 1|X)}{1 - P(D = 1|X)}] \tag{10}$$

and estimated as:

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_{i1} - Y_{i0}}{\hat{p}} * \frac{D_i - \hat{g}(X_i)}{1 - \hat{g}(X_i)} \tag{11}$$

where $\hat{p} = P(D_i = 1)$, and propensity score estimation $\hat{g}(X_i) = P(D_i = 1 | X_i)$. However, Abadie’s semi-parametric DID estimator is asymptotically normal and $\sqrt{N}$-consistent when the propensity score is calculated using traditional non-parametric methods $\sqrt{N}$. However, when using an ML estimator for $\hat{g}(X_i)$, $\hat{\theta}$ may not exhibit $\sqrt{N}$-consistency. This occurs because the score function based on equation 11 has a non-zero directional derivative with respect to the propensity score, and ML estimators typically converge at a slower rate than $\sqrt{N}^{-\frac{1}{2}}$ due to regularization bias [14]. DML-ID addresses this using cross-fitting and the Neyman-orthogonal score [12]:

$$\varphi = \frac{Y_1 - Y_0}{P(D = 1)} * \frac{D - P(D = 1|X)}{1 - P(D = 1|X)} - \theta_0 - c \tag{12}$$

where c = $\frac{D - P(D=1|X)}{P(D=1) * (1 - P(D=1|X))} \mathbb{E}[Y_1 - Y_0 | X, D = 0]$ is an adjustment term.

Neyman orthogonality condition assumes that the moment conditions (equations that capture the expected relationships between the outcome variable, the treatment variable, and other covariates or control variables) are orthogonal or independent of certain nuisance parameters (parameters not of interest) [12]. Thus, using the Neyman-orthogonal score permits the reliable estimation of the parameter of interest, and the application of cross-fitting effectively eliminates bias caused by overfitting. DMLDID procedures can be summarized as:

- **Step 1:** Divide the data sample S into K random partitions using K-fold splitting. To keep it simple, make each partition S_k the same size $n = \frac{N}{K}$. For each partition, create an auxiliary subset $S_k^c = S \setminus S_k$ from the original sample.
- **Step 2:** Use S_k^c to build estimates for p_0, g_0, and l_0, where $l_0 = \mathbb{E}[Y_1 - Y_0|X, D = 0]$. The estimate $\hat{p}_k = \frac{1}{n}\sum_{i \in S_k{}^c} D_i$, $\hat{l_k}$, $\hat{g_k}$ are calculated using ML methods that converge faster than $N^{-\frac{1}{4}}$, such as Random Forests. Extremely high or low propensity scores are usually removed. For estimating the nuisance parameter l_0, the Random Forests algorithm is used in this study as it can handle various types of variables and complex model relationships.
- **Step 3:** Create an interim debiased ATT estimator using the remaining data in partition S_k using the formula:

$$\tilde{\theta}_k = \frac{1}{n}\sum_{i \in S_k} \frac{D_i - \hat{g}_k(X_i)}{\hat{p}k(1 - \hat{g}k(X_i))} \cdot (Y_{i1} - Y_{i0} - \hat{l}_k(X_i)) \quad (13)$$

- **Step 4:** Combine the estimators from all K partitions to obtain the final Doubly Robust Estimator for the Average Treatment Effect: $\tilde{\theta}_{DML} = \frac{1}{K}\sum_{k=1}^{K} \tilde{\theta}_k$.

With the DMLDID approach, we can interpret the pandemic's impact on CVD patients, while also accounting for the influence of complex parameters (nuisance) related to the pandemic, crucial for avoiding biased effect estimates.

F. ML Regression Model Performance Metrics

Regression metrics were used to evaluate our regression models including Mean Square Error (MSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_t)^2 \quad (14)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \bar{Y}_t| \quad (15)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_t)^2}{\sum_{i=1}^{n}(Y_i - \tilde{Y}_t)^2} \quad (16)$$

where $\bar{Y}_t$ is the predicted value, Y_i the actual value, $\tilde{Y}_t$ the mean value and n the index of an observed sample.

G. Interpreting ML Models Using Shapley Values

Shapley values [41] quantify the influence of individual features on the outcome of a prediction. Specifically, Shapley values determine the predictive value of a feature f, by calculating its average incremental contribution across all potential feature subsets. The SHAP explanation method uses additive feature attribution based on Shapley values, expressed as:

$$g(z')\psi_0 + \sum_{j=1}^{m}\psi_j z'_j \quad (17)$$

where $g(z')$ is the model's prediction when all features are set to specific values z'. m is the total number of features in the model. The constant term ψ is the average prediction value across all samples, ψ_0 is the Shapley value for the baseline prediction while ψ_j is the Shapley value associated with feature j. The best performing predictive model was used to generate our final results for predicting each CVD biomarker. Shapley values provided insight into the importance of each feature to predicting CVD patient biomarker values during COVID-19.

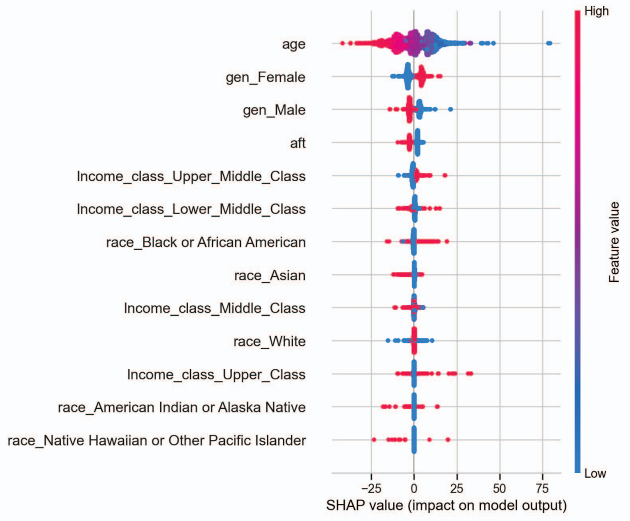


Fig. 4: Feature importance of Explanatory variables for LDL Cholesterol based on SHAP values. Features (rows) are ordered by decreasing overall importance to the prediction. Each dot represents an individual. The plot for each feature shows the SHAP value of each observation on the x-axis, with color representing the value of the feature from low (blue) to high (red). The absolute value represents the magnitude of the feature's contribution, while its sign indicates whether the contribution is positive or negative. Age was the most important feature for predicting LDL cholesterol.

IV. Results

A. Correlation Between Model Features and Biomarkers

Pearson correlation analysis showed no significant linear relationship between the input features and the CVD biomarkers (Figure 3 and Figures 10, 11, 12 in Appendix VII-B), justifying our use of non-parametric ML models.

B. ML Regression of CVD Patient Biomarker Values

1) LDL Cholesterol: Of the six ML models compared, CATBoost consistently performed best across all evaluation metrics for predicting LDL cholesterol values of CVD patients, achieving the lowest MSE and MAE of 1250.023 and 27.749, respectively (Table III). Moreover, CATBoost also had the highest explained variability, with an R^2 value of 0.134, implying that 13.4% of the variance in LDL cholesterol levels is accounted for by the model.

TABLE III: Model predictive results for LDL Cholesterol

Model	Decision Tree	Random Forest	XGBoost	Neural Network	ADABoost	CATBoost
Mean Squared Error (MSE)	1312.17	1277.728	1271.33	1288.77	1417.45	1250.023
Mean Absolute Error (MAE)	28.24	27.90	27.81	28.15	30.35	27.749
R-squared (R^2)	0.09	0.1148	0.119	0.107	0.018	0.134

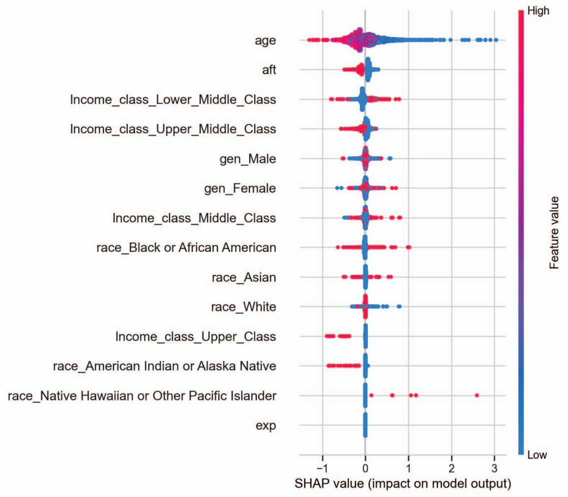


Fig. 5: SHAP feature importance values of input features for predicting HbA1c values. Age was the most important feature for predicting HbA1c levels

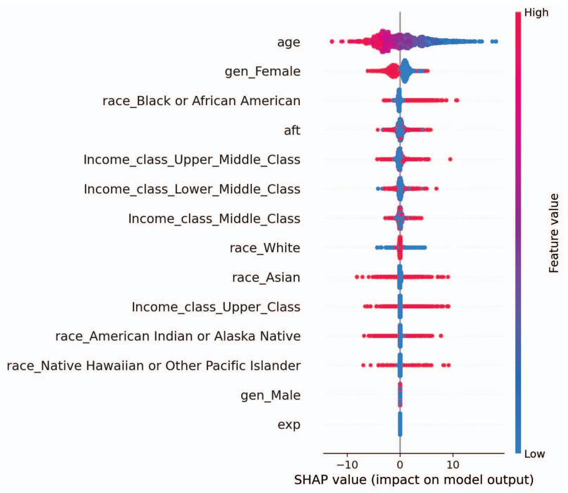


Fig. 6: Feature importance of explanatory variables for BP levels (systolic) based on SHAP values. Age was the most important feature for predicting BP levels

Due to its superior performance on LDL cholestrol regression, CATBoost was selected for SHAP analysis. Figure 4 depicts SHAP values of age, gender, race, and socioeconomic status features generated by CATBoost for predicting LDL cholesterol values of CVD patients during COVID-19. Age was the most important feature with the highest SHAP value, with an inverse relationship with LDL cholesterol values. This implied that as age increased, LDL cholesterol values decreased. Other features had lower importance for predicting LDL cholesterol: higher LDL was associated with an increase in the number of females, upper-middle-class income category, and Black or African American patients. Onset of the COVID-19 pandemic (denoted as the "aft" feature) was moderately linked to lower LDL cholesterol, suggesting the potential effect of the pandemic on this biomarker.

2) HbA1c: CATBoost had the lowest MSE (2.113) of all six models for predicting HbA1c, followed closely by XGBoost and RF (Table IV). Notably, NN exhibited the lowest R^2 (0.004), suggesting limited explained variance despite competitive error metrics. ADABoost had relatively higher errors (MSE=2.580, MAE=1.300) and lower R^2 (0.062) compared to other models.

The XGBoost ML model was used for the SHAP analysis because it had the highest R^2 value (0.115) with the best explanation of variance for HbA1c. Figure 5 shows age as the most important feature. Based on SHAP values, lower HbA1c levels was associated with increasing age, upper-middle and upper income, and identification as American Indian or Alaska Native. Other races mostly exhibited unclear relationships with HbA1c values. The "aft" feature, denoting data collected during the COVID-19 pandemic, was linked to lower predicted HbA1c values.

3) Systolic Blood Pressure (BP): Here, XGBoost, DT, and RF had similar MSE and MAE performance (Table V). R^2 values for these models were also comparable. RF was used to calculate feature importance as it had the highest R^2. The SHAP values shown in 6 indicate that age was the most important feature for predicting BP. Decrease in predicted BP values was associated with increasing age and number of the female gender, while the Black or African American race had increased BP predictions, with a less clear impact on other races. Income had no clear interaction with predicted. BP levels. The pandemic ("aft" feature) had no clear interaction with BP, though there was a slight increase in values measured during the pandemic. These findings highlight the complex influence of demographic and socioeconomic factors on BP predictions, with age, gender, race, income class, and the pandemic time period all having significant effects.

4) Body Mass Index: DT and RF had the lowest MSE values (best performance) and DT, RF and XGBoost had the lowest MAE values for predicting BMI (Table VI). RF was selected to calculate SHAP feature imporance since it had the highest R^2 value, explaining nearly 20% of variance. Figure 7 shows the SHAP values of all features for predicting BMI. Age had the highest SHAP value (highest importance) and

TABLE IV: Model predictive results for HbA1c

Model	Decision Tree	Random Forest	XGBoost	Neural Network	ADABoost	CATBoost
Mean Squared Error (MSE)	2.227	2.180	2.150	2.323	2.580	2.113
Mean Absolute Error (MAE)	1.1178	1.11	1.102	1.164	1.3	1.12
R-squared (R^2)	0.083	0.103	0.115	0.004	0.062	0.108

TABLE V: Model predictive results for Blood Pressure

Model	Decision Tree	Random Forest	XGBoost	Neural Network	ADABoost	CATBoost
Mean Squared Error (MSE)	493.520	493.505	494.352	519.113	536.038	508.729
Mean Absolute Error (MAE)	17.66	17.66	17.68	18.18	18.63	17.987
R-squared (R^2)	0.071	0.071	0.069	0.0227	-0.0092	0.042

TABLE VI: Model predictive results for Body Mass Index

Model	Decision Tree	Random Forest	XGBoost	Neural Network	ADABoost	CATBoost
Mean Squared Error (MSE)	47.897	47.8997	48.004	53.212	65.998	48.74
Mean Absolute Error (MAE)	5.17	5.17	5.184	5.473	6.405	5.354
R-squared (R^2)	0.19	0.192	0.190	0.102	0.113	0.145

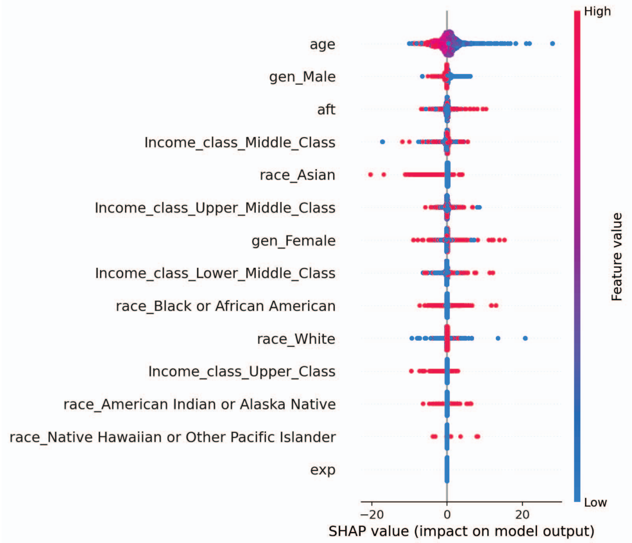


Fig. 7: Feature importance of Explanatory variables for BMI Levels based on SHAP values. Age was the most important feature for predicting BMI levels

increasing age was associated with lower BMI values. Lower BMI was associated with increasing number of males and Asians. Income class did not have a clear interaction with predicted BMI. Data collected during the COVID-19 pandemic ("aft" feature), did not have a clear direction, although BMI increased during the pandemic.

C. The Causal Effect of the COVID-19 Pandemic on Patients with and without CVD

TABLE VII: DML estimates of the impact of the pandemic on CVD patients for both biomarkers

	Effect / Coefficient	Std Err	P-value	T-test
BMI	16.437	6.6026	0.0128	2.489
BP	1.179	0.028	<0.01	41.758

To estimate the impact of the pandemic on CVD patients' BMI and BP biomarker values, DMLDID analysis was used, comparing CVD patients with a control group. Only BMI and BP were analyzed because patient values were available for both patient groups. HbA1c and LDL cholesterol biomarker values were unavailable for control group patients, as they were only tested for CVD patients. The RF model was used as it performed best in predicting BP and BMI. The results in Table VII shows that at a significance level of 0.05, the coefficient for the DID analysis on BMI indicates a significant effect (t = 2.489, p = 0.0128). The estimated coefficient of 16.437 suggests an average increase in BMI in the CVD group compared to controls during the intervention. DID analyses on BP reveals a highly significant impact (t = 41.758, $p < 0.01$). The coefficient of 1.179 indicates that during the pandemic, CVD patients had higher systolic BP values than the control group by an average of 1.179mmHg. These findings suggest that the onset of the pandemic has a discernible effect on both BMI and BP, highlighting the effectiveness of the DID approach in evaluating the impact of the pandemic.

V. Discussion

This study explored the impact of COVID-19 pandemic on four key biomarkers in CVD patients. Contrary to our initial hypothesis, which anticipated health deterioration due to factors like limited access to medical facilities and pandemic-related measures such as lockdowns, the biomarker analysis yielded varying findings. Surprisingly, HbA1c and LDL cholesterol values improved during the pandemic, but BMI and BP increased as hypothesized. The lack of a control group for HbA1c and LDL cholesterol limits comparisons. However, comparing experimental (CVD patients during the pandemic) with a control group revealed significant impacts on BMI and BP. The findings suggest differential effects of the pandemic on various CVD biomarkers. We now expound on our findings.

LDL Cholesterol: decreased in CVD patients after the onset of the pandemic, and it was best predicted using CATBoost. The SHAP analysis identified age as the most important feature, with lower predicted LDL cholesterol associated with increasing age, consistent with prior findings [42]. Other features exhibited varied importance on prediction, aligning with studies linking aging, chronic diseases, and COVID-19 to higher mortality rates [1], [43].

Glycated Hemoglobin (HbA1c): Contrary to our expectations, HbA1c values decreased (improved) in CVD patients during the pandemic, similar to the findings for LDL cholesterol. SHAP analysis highlighted age as the most important feature, followed by income and race, aligning with prior work linking higher HbA1c with socioeconomic disparities and stress [44]. However, the unexpected decrease in HbA1c with increasing age contrasts with typical trends, as HbA1c tends to rise with age in individuals without diabetes [45].

Blood Pressure (BP): RF performed best in predicting BP. We found slightly higher BP readings for CVD patients compared to the control group (pre- and post-pandemic), aligning with previous studies linking elevated BP with CVD [46]. Causal analysis via DMLDID revealed a value of 1.179, congruent with the hypothesis that the stressful pandemic would worsen BP. Contrary to prior findings [47], we found that BP values decreased with advancing age.

Body Mass Index (BMI): RF also performed best in predicting BMI. BMI values of CVD patients increased during the pandemic compared to controls (DMLDID: 16.437), which aligns with prior studies that found that high BMI values are generally correlated with CVD and specifically among chronic diseases patients who contracted COVID-19 [48]. This is consistent with our hypotheses that the BMI of CVD patients would increase during the pandemic, possibly due to less access to routine medical care.

Rationale for Biomarker Changes during the Pandemic

Our biomarker analyses revealed both expected and unexpected results, emphasizing the need for nuanced analytical methods to understand the pandemic's differential effects. While BMI and BP worsened as anticipated in CVD patients during the pandemic, we unexpectedly observed improvements in LDL cholesterol and HbA1c compared to a control group. We explore three possible explanations for these unexpected changes: lifestyle adjustments due to COVID-19, patient selection biases favoring health-conscious individuals, and potential study limitations.

Lifestyle changes, including modified eating habits and heightened health awareness, could have counterbalanced challenges in accessing medical care during the pandemic, particularly for patients vulnerable to pandemic-related complications. These individuals may have been motivated to enhance their health, aiming to prevent complications. Previous research indicates that major life events can lead to behavioral changes associated with health improvements [49], [50].

Patient selection bias may have yielded a dataset with a more health-conscious population. Put another way, patients who placed enough importance on seeking medical treatment and requested lab measurements both during and before the pandemic may be more likely to prioritize their health than patients who did not regularly measure their biomarkers and hence not represented in the dataset.

Study limitations that may have affected our results include the relatively small effect sizes relative to the large sample size. The control group's composition may also be a limitation, as individuals without CVD might may have other other health issues. In future, the control group may be reconstructed using on different approaches that reduce selection biases.

VI. Conclusion and Future work

Our multifaceted ML-based analysis revealed CATBoost, XGBoost, and RF regression models outperformed other models in predicting LDL cholesterol, HbA1c, BMI and BP trajectories among CVD patients. These models consistently identified age as the most important variable on CVD biomarkers, which aligns with established knowledge that aging affects biological mechanisms. Interestingly, income was another key factor for LDL cholesterol and HbA1c biomarkers and race and ethnicity had notable impacts BP and BMI biomarkers highlighting the need to consider social determinants of health in biomarker analyses.

DMLDID analysis revealed significant changes in CVD patient biomarker values during the pandemic, estimating higher changes in their BMI and systolic BP compared to controls. This suggests that interventions could be targeted and highlights a differential impact of the pandemic on CVD patients, possibly due to pre-existing conditions or altered behaviors during the pandemic. However, the specific mechanisms causing these group differences require further investigation.

Our findings provide valuable insights to healthcare professionals and quantitative evidence of pandemic-related CVD patient biomarker trajectories. Our methodology should be of interest to ML/AI researchers who could repurpose and apply it to analyze other health conditions. DMLDID introduces cutting-edge methods for estimating how a pandemic affects individuals with chronic illnesses.

Future research should delve deeper into the identified relationships, particularly the mechanisms behind pandemic-related changes in high-risk groups. Additionally, incorporating longitudinal data and exploring causal inference methods could improve our understanding of how specific factors influence biomarker trajectories over time. Ultimately, a comprehensive approach that combines traditional statistics and ML techniques holds promise for predictive analytics, and advancing personalized medicine in the face of health challenges caused by the pandemics such as COVID-19.

For Appendix, see the link: Appendix

References

[1] R. Verity *et al.*, "Estimates of severity of coronavirus dis. 2019: a model-based analysis," *Lancet Infec. Dis.*, vol. 20, no. 6, pp. 669–677, 2020.

[2] T. C. Tsai *et al.*, "Association of community-level social vulnerability with US acute care hospital intensive care unit capacity during COVID-19," *Healthcare*, vol. 10, p. 100611, Mar. 2022.

[3] Mozaffarian *et al.*, "Heart disease and stroke statistics—2016 update: a report from the american heart association," *circulation*, vol. 133, no. 4, pp. e38–e360, 2016.

[4] Mattioli *et al.*, "COVID-19 pandemic: the effects of quarantine on cardiovascular risk," *Eur J. Clin Nutr.*, vol. 74, pp. 852–855, June 2020. Number: 6 Publisher: Nature Publishing Group.

[5] N. Bilgin Dogan and E. Ozel, “The missing stemis and lifestyle changes during the covid-19 pandemic,” *Asia Pacific Journal of Public Health*, vol. 33, no. 2-3, pp. 296–298, 2021.

[6] I. Y.-H. Chu, P. Alam, H. J. Larson, and L. Lin, “Social consequences of mass quarantine during epidemics: a systematic review with implications for the covid-19 response,” *Journal of travel medicine*, vol. 27, no. 7, p. taaa192, 2020.

[7] M. E. Dupre *et al.*, “Access to routine care and risks for 30-day readmiss. in patients with CVD,” *Am. Heart J.*, vol. 196, pp. 9–17, 2018.

[8] Aparisi *et al.*, “Low-density lipoprotein cholesterol levels are associated with poor clinical outcomes in COVID-19,” *Nutr. Metab Cardiovas*, vol. 31, pp. 2619–2627, Aug. 2021.

[9] A. S. Ikram and S. Pillay, “Admission vital signs as predictors of COVID-19 mortality: a retrospective cross-sectional study,” *BMC Emergency Medicine*, vol. 22, p. 68, Apr. 2022.

[10] Mehrabadi *et al.*, “Detection of COVID-19 Using Heart Rate and Blood Pressure: Lessons Learned from Patients with ARDS,” in *Proc. EMBC 2021*, (Mexico), pp. 2140–2143, IEEE, Nov. 2021.

[11] A. Goto *et al.*, “Hemoglobin a1c levels and the risk of cardiovascular disease in people without known diabetes: a population-based cohort study in japan,” *Medicine*, vol. 94, no. 17, p. e785, 2015.

[12] Y. D. Chernozhukov, Victor Peng *et al.*, “Double/debiased machine learning for treatment and structural parameters,” 2018.

[13] J. B. Dimick and A. M. Ryan, “Methods for evaluating changes in health care policy: the difference-in-differences approach,” *Jama*, vol. 312, no. 22, pp. 2401–2402, 2014.

[14] N.-C. Chang, “Double/debiased machine learning for difference-in-differences models,” *Economet J*, vol. 23, no. 2, pp. 177–191, 2020.

[15] D. Cucinotta and M. Vanelli, “WHO Declares COVID-19 a Pandemic,” *Acta Bio Medica : Atenei Parmensis*, vol. 91, no. 1, pp. 157–160, 2020.

[16] L. Tan *et al.*, “Toward real-time and efficient cardiovascular monitoring for COVID-19 patients by 5G-enabled wearable medical dev: a deep learning app,” *Neural Comp. App*, vol. 35, pp. 13921–13934, July 2023.

[17] C. Hu *et al.*, “Early prediction of mortality risk among patients with severe COVID-19, using machine learning,” *Intl. J Epi*, vol. 49, pp. 1918–1929, Dec. 2020.

[18] B. Pramenković *et al.*, “Machine Learning Techniques for Predicting Outcomes of COVID-19 for Patients with preexisting Chronic Diseases,” in *CMBEBIH 2021* (A. Badnjevic and L. Gurbeta Pokvić, eds.), IFMBE Proceedings, (Cham), pp. 867–882, Springer Intl. Pub., 2021.

[19] D. Castelnuovo *et al.*, “Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and ml-based findings from the multicentre Italian CORIST Study,” *Nutr. Metab. Cardiovas.*, vol. 30, pp. 1899–1913, Oct. 2020.

[20] A. Saab *et al.*, “Early Prediction of All-Cause Clinical Deterioration in General Wards Patients: Development and Validation of a Biomarker-Based Machine Learning Model Derived From Rapid Response Team Activations,” *J Patient Saf.*, vol. 18, p. 578, Sept. 2022.

[21] M. Mahdavi *et al.*, “A machine learning based exploration of COVID-19 mortality risk,” *PLOS ONE*, vol. 16, p. e0252384, July 2021.

[22] K. Singh *et al.*, “Health, psychosocial, and economic impacts of the COVID-19 pandemic on people with chronic conditions in India: a mixed methods study,” *BMC Public Health*, vol. 21, p. 685, Apr. 2021.

[23] Barría-Sandoval *et al.*, “Interpretable machine learning for mortality modeling on patients with chronic diseases considering the COVID-19 pandemic in a region of Chile: A Shapley value based approach,” *Research in Statistics*, vol. 1, Aug. 2023.

[24] F. Allery *et al.*, “Towards mitigating health inequity via machine learning: a nationwide cohort study to develop and validate ethnicity-specific models for prediction of cvd risk in covid-19 patients,” 2023.

[25] B. Koch *et al.*, “Deep learning for causal inference,” *SocArXiv. October*, vol. 10, 2021.

[26] Y. Zhang *et al.*, “Estimating heterogeneous treatment effects in road safety analysis using generalized random forests,” *Accident Analysis & Prevention*, vol. 165, p. 106507, 2022.

[27] Y. Zhang *et al.*, “Quantifying the social impacts of the london night tube with a double/debiased machine learning based difference-in-differences approach,” *Transport Res. A-Pol*, vol. 163, pp. 288–303, 2022.

[28] L. Xing, D. Han, and X. Hui, “The impact of carbon policy on corporate risk-taking with a double/debiased machine learning based difference-in-differences approach,” *Financ. Res. Lett.*, vol. 58, p. 104502, 2023.

[29] World Health Organization, “International classification of diseases (icd).” https://www.who.int/standards/classifications/classification-of-diseases. Accessed: Dec. 13, 2023.

[30] E. Stuart *et al.*, “Matchit: nonparametric preprocessing for parametric causal inference,” *J. stat. software*, 2011.

[31] R. B. Hawkins *et al.*, “Socio-economic status and covid-19–related cases and fatalities,” *Public health*, vol. 189, pp. 129–134, 2020.

[32] “US ZIP Codes:.” https://www.unitedstateszipcodes.org. 2023.

[33] P. Sedgwick, “Pearson’s correlation coefficient,” *Bmj*, vol. 345, 2012.

[34] B. De Ville, “Decision trees,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 5, no. 6, pp. 448–455, 2013.

[35] S. Pathak, I. Mishra, and A. Swetapadma, “An Assessment of Decision Tree based Classification and Regression Algorithms,” in *Proc (ICICT)*, pp. 92–95, Nov. 2018.

[36] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proc. SIGKDD*, pp. 785–794, 2016.

[37] K. Worden *et al.*, “Artificial Neural Networks,” in *Machine Learning in Modeling and Simulation: Methods and Applications* (T. Rabczuk and K.-J. Bathe, eds.), Computational Methods in Engineering & the Sciences, pp. 85–119, Cham: Springer International Publishing, 2023.

[38] K. H. Abegaz and Etikan, “Boosting the Performance of Artificial Intelligence-Driven Models in Predicting COVID-19 Mortality in Ethiopia,” *Diagnostics, MDPI*, vol. 13, Jan. 2023. Number: 4.

[39] S. Olszewski *et al.*, “Regression Modeling for Monitoring Organochlorine Pesticide Residues,”

[40] A. Abadie, “Semiparametric difference-in-differences estimators,” *Rev. Econ. Stud.*, vol. 72, no. 1, pp. 1–19, 2005.

[41] S. M. Lundberg *et al.*, “From local explanations to global understanding with explainable AI for trees,” *Nature Machine Intelligence*, vol. 2, pp. 56–67, Jan. 2020. Number: 1 Publisher: Nature Publishing Group.

[42] A. Ferrara *et al.*, “Total, ldl, and hdl cholesterol decrease with age in older men and women: The rancho bernardo study 1984–1994,” *Circulation*, vol. 96, no. 1, pp. 37–43, 1997.

[43] O. Dyer, “Covid-19: Pandemic is having “severe” impact on non-communicable disease care, WHO survey finds,” *BMJ*, vol. 369, p. m2210, June 2020. Publisher: British Med. J. Pub. Group.

[44] M. E. Hilliard *et al.*, “Stress and a1c among people with diabetes across the lifespan,” *Curr. Diabetes Rep.*, vol. 16, pp. 1–10, 2016.

[45] L. N. Pani *et al.*, “Effect of aging on a1c levels in individuals without diabetes: evidence from the framingham offspring study and the national health and nutrition examination survey 2001–2004,” *Diabetes Care*, vol. 31, no. 10, pp. 1991–1996, 2008.

[46] A. C. Flint *et al.*, “Effect of systolic and diastolic blood pressure on cardiovascular outcomes,” *New Eng. J. Med.*, vol. 381, no. 3, pp. 243–251, 2019.

[47] L. J. Laffin, H. W. Kaufman, Z. Chen, J. K. Niles, A. R. Arellano, L. A. Bare, and S. L. Hazen, “Rise in blood pressure observed among us adults during the covid-19 pandemic,” *Circulation*, vol. 145, no. 3, pp. 235–237, 2022.

[48] Y. D. Peng *et al.*, “Clinical characteristics and outcomes of 112 cardiovascular disease patients infected by 2019-nCoV,” *Zhonghua Xin Xue Guan Bing Za Zhi*, vol. 48, pp. 450–455, June 2020.

[49] F. Astin *et al.*, “Managing lifestyle change to reduce coronary risk: a synthesis of qualitative research on peoples’ experiences,” *BMC cardio. disorders*, vol. 14, no. 1, pp. 1–16, 2014.

[50] Z. S. Verdun, “Impact of a health shock on lifestyle behaviours,” 2020.

[51] E. Pekel, “Estimation of soil moisture using decision tree regression,” *Theor. Appl. Climatol*, vol. 139, pp. 1111–1119, Feb. 2020.

[52] G. Biau and E. Scornet, “A random forest guided tour,” *TEST*, vol. 25, pp. 197–227, June 2016.

[53] E. G. Dada *et al.*, “Ensemble Machine Learning for Monkeypox Transmission Time Series Forecasting,” *Applied Sciences*, vol. 12, p. 12128, Jan. 2022. Number: 23 Publisher: MDPI.

[54] D. K. Barrow and S. F. Crone, “A comparison of AdaBoost algorithms for time series forecast combination,” *Intl J. Forecasting*, vol. 32, pp. 1103–1119, Oct. 2016.

[55] R. Asad *et al.*, “Achieving Personalized Precision Education Using the Catboost Model during the COVID-19 Lockdown Period in Pakistan,” *Sustainability*, vol. 15, p. 2714, Jan. 2023. Number: 3 Pub: MDPI.

[56] J. H. Friedman, “Stochastic gradient boosting,” *Comp. Stat. Data An.*, vol. 38, pp. 367–378, Feb. 2002.

[57] J.-e. Chen *et al.*, “Debiased/Double ML for Instr. Variable Quantile Regressions,” *Econometrics, MDPI*, vol. 9, p. 15, June 2021. No. 2.

[58] J. B. Dimick and A. M. Ryan, “Methods for Evaluating Changes in Health Care Policy: The Difference-in-Differences Approach,” *JAMA*, vol. 312, pp. 2401–2402, Dec. 2014.