DISSERTATION

NOVEL ASSESSMENTS OF COUNTRY PANDEMIC VULNERABILITY BASED ON NON-PANDEMIC PREDICTORS, PANDEMIC PREDICTORS, AND COUNTRY PRIMARY AND SECONDARY VACCINATION INFLECTION POINTS

Submitted by

Marco M. Vlajnic

Department of Systems Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2024

Doctoral Committee:

Advisor: Steven Simske

James Cale
Steven Conrad
Bradley Reisfeld

ABSTRACT

NOVEL ASSESSMENTS OF COUNTRY PANDEMIC VULNERABILITY BASED ON NON-PANDEMIC PREDICTORS, PANDEMIC PREDICTORS, AND COUNTRY PRIMARY AND SECONDARY VACCINATION INFLECTION POINTS

The devastating worldwide impact of the COVID-19 pandemic created a need to better understand the predictors of pandemic vulnerability and the effects of vaccination on case fatality rates in a pandemic setting at a country level. The non-pandemic predictors were assessed relative to COVID-19 case fatality rates in 26 countries and grouped into two novel public health indices. The predictors were analyzed and ranked utilizing machine learning methodologies (Random Forest Regressor and Extreme Gradient Boosting models, both with distribution lags, and a novel K-means-Coefficient of Variance sensitivity analysis approach and Ordinary Least Squares Multifactor Regression). Foundational time series forecasting models (ARIMA, Prophet, LSTM) and novel hybrid models (SARIMA-Bidirectional LSTM and SARIMA-Prophet-Bidirectional LSTM) were compared to determine the best performing and accurate model to forecast vaccination inflection points. XGBoost methodology demonstrated higher sensitivity and accuracy across all performance metrics relative to RFR, proving that *cardiovascular death rate* was the most dominant predictive feature for 46% of countries (Population Health Index), and *hospital beds per thousand people* for 46% of countries (Country Health Index). The novel K-means-COV sensitivity analysis approach performed with high accuracy and was successfully validated across all three methods, demonstrating that *female smokers* was the most common predictive feature across different analysis sets. The new model was also validated with the Calinski-Harabasz

methodology. Every machine learning technique that was evaluated showed great predictive value and high accuracy. At a vaccination rate of 13.1%, the primary vaccination inflection point was achieved at 83.27 days. The secondary vaccination inflection point was reached at 339.31 days at the cumulative vaccination rate of 67.8%. All assessed machine and deep learning methodologies performed with high accuracy relative to COVID-19 historical data, demonstrated strong forecasting value, and were validated by anomaly and volatility detection analyses. The novel triple hybrid model performed the best and had the highest accuracy across all performance metrics. To be better prepared for future pandemics, countries should utilize sophisticated machine and deep learning methodologies and prioritize the health of elderly, frail and patients with comorbidities.

ACKNOWLEDGEMENTS

I would like to thank all of my colleagues at CSU and in particular the Systems Engineering Department. None of this would be possible without my advisor, Steven Simske, and his guidance and mentorship over the years. I would also like to thank my committee members, Dr. James Cale, Dr. Steven Conrad, and Dr. Bradley Reisfeld for all of their support.

I would like to thank my family for being understanding of the late nights and the limited free time I had during this busy time. In particular, I would like to thank my parents, Miodrag and Aleksandra Vlajnic, and my brother, Vanja Vlajnic, for all the support they have provided me with. Their continued belief in me has strengthened my own belief in myself. Most of all I am grateful for my faith and my family's support.

I recognize this work would not have been possible without the support, guidance, and mentorship of those above. Thank you all. I hope I can leave a positive impact on those around me as you all have on me.

DEDICATION

I would like to dedicate this dissertation to my parents, Aleksandra and Miodrag Vlajnic, and my brother, Vanja Vlajnic for all of the sacrifices they made in their own lives to help provide me with opportunities to make my dreams a reality.

TABLE OF CONTENTS

ABSTRACT…… ii
ACKNOWLEDGEMENTS……iv
DEDICATION……v

Chapter 1 Introduction……. 1
1.1 COVID-19 Pandemic Background…….. 1

Chapter 2 Research Objectives…….. 8

Chapter 3 Research Part 1: Dataset……... 9

Chapter 4 Research Part 1: Methodologies……12
4.1 Data Preprocessing with 70/30 train-test-split ratio……12
4.2 Regression Methodologies……..13
4.21 Random Forest Regressor with Distribution Lag and Extreme Gradient Boosting Regressor with Distribution Lag …….... 14
4.22 Ordinary Least Squares Multifactor Regression ……..16
4.3 Clustering Methodologies……...17
4.31 Novel K-means Coefficient of Variance Methodology Approach…..... 17
4.32 Calinski-Harabasz Methodology…….. 19
4.4 Pandemic Risk Scoring Model……... 19

Chapter 5 Research Part 1: Results……20
5.1 Regression Methodology Results……... 20
5.11 Random Forest and Extreme Gradient Boosting Regressor Results…… 20
5.2 Clustering Methodology Results……27
5.21 K-means-Coefficient of Variance Sensitivity and Ordinary Least Squares Multifactor Regression Analysis Results……...27
5.22 Calinski-Harabasz Methodology Results……...31
5.3 Pandemic Risk Scoring Model Results……...36

Chapter 6 Research Part 2: Dataset……38

Chapter 7 Research Part 2: Methodologies……42
7.1 Data Preprocessing of Timeseries Dataset Temporally…….....42
7.2 Forecasting Methodologies……43
7.21 Correlation Analysis…….. 43
7.22 Foundational Forecasting Methodologies…….....44
7.221 Autoregressive Integrated Moving Average (ARIMA)……44
7.222 Facebook Prophet…….....44
7.223 Long Short-Term Memory (LSTM)……..... 44
7.23 Novel Hybrid Forecasting Models……45

7.231 Double Hybrid Model: SARIMA-Bidirectional LSTM…….. 45
7.232 Triple Hybrid Model: SARIMA-Prophet-Bidirectional LSTM……....47
7.3 Accuracy and Performance Assessment…….. 47
7.4 Anomaly and Volatility Analyses……....48
7.41 Isolation Forest……..48
7.42 Generalized Autoregressive Conditional Heteroskedasticity…..... 50
7.5 Vaccination Inflection Point Score……..... 51

Chapter 8 Research Part 2: Results……...52
8.1 Correlation Analysis Results……52
8.2 Forecasting Analysis Results……..... 54
8.3 Accuracy and Performance Assessment…….. 60
8.4 Anomaly and Volatility Analysis Results……60
8.5 Vaccination Inflection Point Score Results…… 63

Chapter 9 Discussion, Conclusions, and Limitations……... 65
9.1 Discussion…….....65
9.2 Limitations……... 76
9.3 Relevance of Research to Systems Engineering……...79
9.31 Verification and Validation (V/V)……... 79
9.32 Tests and Measurement…….....80
9.33 Sensitivity Analysis……..... 81
9.34 Repeatability/resilience/reliability…….. 83
9.4 Future Research…….... 84
9.5 Conclusions……..88

Bibliography…… 89

Chapter 1

Introduction

1.1 COVID-19 Pandemic Background

COVID-19 is an infectious disease, caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) characterized by high morbidity and mortality, and a significant burden on hospital systems and country economies. COVID-19 pandemic will always be remembered as one of the worst healthcare crises in the modern time. It started in December 2019, when the Wuhan Municipal Health Commission (Hubei Province, China) reported a cluster of pneumonia cases of unknown origin, and already in March 2020, the World Health Organization declared a global pandemic [1, 2, 3]. As of the end of 2023, the COVID-19 virus infected over 300 million people, caused deaths for approximately seven million people [1], and had a negative impact of $3.8 trillion on economies around the world and an estimated loss of $202.6 billion in revenue for America's hospitals and healthcare systems [4, 5, 6, 7]. At the end of 2023, COVID-19 is still present with different virus mutations continuing to cause infections and deaths across the world [2, 3, 4].

At the beginning of the pandemic, the world was not prepared, and the health systems were overwhelmed, struggling to manage and triage the high number of infected patients. These patients commonly required advanced care, often in intensive care units, and, unfortunately, for many patients it had a fatal outcome. Scientists around the world were accelerating research to develop new vaccines. Researchers started analyzing the available data in the effort to understand the disease clinically and guide the treatment of patients, as well as how to screen and guide contact tracing and drug development for SARS-CoV-2 virus. They focused their work to address clinical

aspects: identification of drug candidates against SARS CoV-2 virus [27], risk assessment of patients at hospital admission [28]; blood markers as tools for quarantine assessment [29], and vaccine data [32]. Other researchers focused their work on assessing non-clinical factors, such as demographics, travel, environmental factors (temperature, relative humidity, atmospheric pollutants, etc.), capacity and health related county-level factors, vulnerable population scores, national socio-economic factors, and different epidemiological data [12, 30, 31, 33, 34, 35, 36].

Public health experts around the world were trying to forecast the spread of the pandemic, understand critical factors that influence it, and provide guidance to countries on how to respond to the pandemic. Furthermore, they recognized the need to define, predict, and better understand critical country-level factors contributing to morbidity and mortality in a pandemic setting, allowing countries to improve their pandemic readiness, for COVID-19 and any possible new pandemic.

In the effort to better organize and assess data, researchers often utilized existing public heath indices, ratios, and initiatives, such as Case Fatality Ratio and Global Health Security Initiative. They assessed indicators of the magnitude of COVID-19 burden by applying model-derived measures of pandemic severity, statistical models, and corresponding clinical parameters, including excess mortality [9, 10, 11, 15]. The Absolute and Signed Importance Index were used to identify socio-economic factors that contribute to the variability of the pandemic. Using the COVID-19 Vulnerability Index and the pandemic severity Impact Assessment, vulnerable counties were identified and mapped [14, 15]. The average mortality, hospital and critical care unit occupancy, and vaccine impact were measured at the national level using the Resilience Index (r) and the Preparedness and Prevention Index (p) [16]. Poverty was also found to be a significant contributing factor. Multidimensional poverty indices were used on

a global scale, and the COVID-19 poverty vulnerability index was employed at the national level. These indicators demonstrated significant regional and ethnic inequality as well as trends toward rising infection rates and higher mortality rates in vulnerable areas [18, 19, 21].

Many countries and organizations were starting to collect data, making it easier to access different data sources and conduct credible research. At the beginning data sources were limited in duration of data, ranging from several weeks and up to 15 months. For example, Johns Hopkins University (January to July 2020), Oxford COVID-19 Government Response Tracker (January to December 2022), Research and Development data for overall Information Value scores, and World Health Organization-Joint External Evaluation data for Ready Score and four sub scores; *Our World in Data* repository (2021); and data from 3042 counties in the United States (January 2020 to March 2021) [13, 14, 16, 17]. Conducted analyses and published literature varied from descriptive statistics to comprehensive advanced data analytics, sophisticated machine learning and artificial intelligence methodologies. For example, regression models with both independent and proximity dependent outcomes, and variable selection through LASSO [14]; non-parametric, multiple non-linear regression techniques, decision tree-based methods, such as Random Forest and Gradient Boost, Support Vector Machines, K-nearest neighbor and deep neural network models, Convolutional Neural Networks [9, 11, 20, 21, 22, 23, 25, 26]; models based on the Broad Learning System [24]; Hierarchical Condition Category Score; Herfindahl–Hirschman Index, Quantile Regression and Hierarchical Regression Models [17]; and unsupervised machine learning techniques, in particular, hierarchical clustering analysis and agglomerative hierarchical clustering [15]. While the results varied in utility, collectively they helped to advance the existing knowledge and paved the way for further research. Several factors influencing mortality rates were starting to emerge, such as population demographics, gender, age, racial minority, economic and socio-

political factors, and the presence of comorbidities such as obesity and cardiovascular disease. In addition, it was clearly observed that there exist significant variations between countries in terms of size, public governance, expenditures in health system, as well as in testing and reporting. This documented heterogeneity across countries created substantial limitations in standardizing assessment approaches requiring more sophisticated testing methods and more simplified and standardized models [10, 14, 15, 16]. This created an opportunity to identify non-pandemic parameters, factors that are routinely collected at a country level, and assess if they can have a predictive value for pandemic outcomes. Modifying these predictive factors would then stimulate development of appropriate strategies and actions before any future pandemic. Country public health officials, policymakers, and disaster management agencies would then be able to proactively plan and increase pandemic readiness at each country level.

Another important aspect of pandemic readiness is to understand the impact of vaccination. We already know that the experience with the COVID-19 pandemic demonstrated inadequate levels of preparedness across countries worldwide. To properly plan, countries should have a good public health threat surveillance, monitoring, and analytics infrastructure in place [78, 79, 80, 81, 82, 83, 84, 85]. However, half of the countries have a limited capacity to systematically monitor care, including the impact of vaccination [77].

At the population level, vaccination is one of the most effective ways to stop the pandemic from starting and spreading [86]. While the role of the natural acquired immunity is critically important, achieving the herd immunity thresholds creates an environment that can faster control large outbreaks, reduce the number of infected individuals and possible deaths, protect the most vulnerable individuals in the society, and relax other public health measures [87]. There are many factors that can influence if a vaccination campaign will be successful. The speed of development,

level of demand and supply, difficulties with production and distribution of vaccines at the country level and worldwide, as well as the overall vaccination strategy (priority groups for vaccination) and the acceptance of the population (e.g. anti-vaccination movement).

The vaccination efforts for COVID-19 started in December of 2020 for most of the countries in the world. There were several types of vaccines that were available: genetically engineered messenger RNA, viral vector, and protein subunit vaccines. The initial vaccinations from 2020 were followed with booster doses in 2021, 2022 and 2023, for a total of four booster doses, specifically in developed countries [88].

Scientists around the world conducted research to better understand the impact of vaccination on the pandemic outcomes, the correlation of vaccination rates, incidence of COVID-19, and mortality rates. The results confirmed that successful vaccination efforts (e.g. availability of vaccines, public acceptance, strong government programs, etc.) can significantly reduce the negative effects of the COVID-19 pandemic, with a sharp decrease in the fatality rate [89, 90, 91]. Some researchers were able to define the vaccination threshold, identifying that a mean level of administrating about 80 doses of vaccines per 100 inhabitants can sustain a reduction of confirmed cases and number of deaths [83], or when the mean cumulative vaccination rate reaches 29.06 doses per 100 people and 7.88 doses per 100 people, respectively, for spread and mortality [91]. Many researchers also looked at the sentiment around vaccination. Attitudes toward COVID-19 and vaccination, conspiracy beliefs, misconceptions, and complaints about COVID-19 control, were documented as dominant sentiments [93, 94, 95]. Researchers used data from different sources (local, national, and global registries) and different time frames (e.g., periods of 3 or 6 months post initial vaccination). To achieve the needed level of sensitivity of models and analyses, researchers utilized different machine and deep learning

methodologies, such as, neural networks with cut effect [89], Augmented Artificial Neural Network Model for the COVID-19 Mortality Prediction relative to the vaccination rates [96]; Deep Learning Sequence Models for Forecasting COVID-19 Spread and Vaccinations with two recurrent neural network-based approaches, LSTM and GRU [97]; amalgamation of neural network with two powerful optimization algorithms, firefly algorithm and artificial bee colony based feed-forward neural networks to look at the effect of vaccinated population on the COVID-19 prediction [98]; and a multi-path long short term memory (LSTM) neural network for COVID-19 forecasting of new viral variants and vaccination [99]. Other researchers explored other models, structured and unstructured machine learning (ML) models [94], structural topic modeling [95], Latent Dirichlet Allocation (LDA) [100], deep learning and NLP [101, 102]. Cheng applied newly developed ARIMA models to improve the accuracy of weekly COVID-19 case growth rates and forecast COVID-19 spread according to protective behavior and vaccination [103]. Hybrid models, such as HARIMA, a hybrid of ARIMA and HGRNN, a hybrid of the Gaussian Process Regression model and the Generalized Regression Neural Network, were employed by Dhamodharavadhani and associates to predict the vaccination rate [104]. Yi-Tui Chen and colleagues explored the effect of vaccination patterns and vaccination rates on the spread and mortality of the COVID-19 pandemic [91], and Kumar utilized the recurrent neural network (RNN) Convolutional Residual Network (RNNCON-Res) [105]. Nicholson and colleagues used both supervised and unsupervised methodologies to identify the critical county-level factors for studying COVID-19 propagation prior to the widespread availability of a vaccine [112].

Published research has increased collective knowledge and has answered many questions. With limitations of every research, availability of more data and novel methodologies, there is a

need and a responsibility to continue to expand the knowledge around pandemic vulnerability that can allow for better understanding of the dynamics of vaccination, infection rates and mortality. One of the questions that remains unanswered at this time is how to accurately assess and then predict the vaccination inflection points and the time needed to reach the critical cumulative vaccination rate thresholds to identify the pandemic turnaround point.

The real cost of this pandemic is yet to be understood, especially the impact on people and their long-term health, as well as the full impact on the economies worldwide. This increases our collective responsibility to address outstanding research questions and determine the best machine and deep learning methodologies that can be utilized to increase the accuracy of our models. This research focuses on addressing two of these questions, the identification of the pre-pandemic predictors and its role in proactive pandemic readiness planning, and the impact of forecasting vaccination inflection points on pandemic outcomes.

Chapter 2

Research Objectives

This study employs four machine learning approaches, including one innovative strategy, to examine non-pandemic factors in an attempt to improve upon previous work. The performance and accuracy of these methodologies were compared and assessed. These models assessed correlation and predictive value of selected demographic, health, and economic non-pandemic parameters relative to COVID-19 case fatality rates in 26 countries, utilizing a comprehensive 3-year longitudinal dataset. The results of these analyses created a foundation for the development of a novel country specific pandemic risk scoring model. This research was conducted to identify the vaccination inflection points and the time needed to reach the critical cumulative vaccination rate thresholds to observe continuous decrease of the case fatality rates. It was conducted both at an aggregate and at the country level. COVID-19 historical data was utilized to develop models that can be used for future pandemics. Applying advanced AI methodologies to forecast time to country specific vaccination inflection points, as well as, assessing the vaccination rates relative to the case fatality rates, can provide another useful tool to guide countries in their pandemic risk preparedness.

Chapter 3

Research Part 1: Dataset

This research used a dataset from the Oxford University *Our World in Data* Covid 19 Dataset [8]. The data points in this dataset were continuously gathered between January 2020 and the present from the following sources: Johns Hopkins University, the Center for Systems Science and Engineering COVID-19 data, the OXFORD COVID-19 Government Response Tracker, and the European Center for Disease Control. The original dataset contains data from 207 countries and territories from which 26 countries were selected for this research: United States, Canada, Austria, Slovenia, Switzerland, Belgium, Sweden, United Kingdom, Czechia, Slovakia, Iceland, Denmark, Finland, Italy, Ireland, Portugal, France, Netherlands, Luxembourg, Spain, Serbia, Bulgaria, Romania, Latvia, Cyprus, and Estonia. Data for this research was accessed and downloaded on Dec 30, 2022, and this longitudinal dataset was used from the period of January 1, 2020, to December 30, 2022. This research was solely conducted by using publicly available data. By dividing the corresponding values in the total deaths column by the total cases column of the dataset, the case fatality rate (CFR), an epidemiologic statistic defined as the proportion of deaths within an observed population of interest [42], was computed for each of the 26 countries. The variables (features) for the vulnerability assessment were selected based on several criteria. They represented demographic, health, and economic public health parameters, non-pandemic in nature, commonly collected and publicly reported on an annual basis for each country. All the parameters in the research database that fit these criteria were used for this research, with no exclusions. They were grouped into two novel indices, developed for the purpose of the pandemic risk scoring model, and represented in Table 3.1.

The Population Health Index (PHI) represents variables that describe the health status of the overall population living in each country, in terms of age, risk factors, chronic conditions and overall life expectancy. The Country Health Index (CHI) represents variables that describe the health status of a particular country, in terms of population and population density, as well as the economic parameters, such as GDP, poverty and health system. The values of variables, included in the indices, did not change during the observational period (January 2020-December 2022). While most of the country's demographic, health and economic non-pandemic parameters do not change appreciably annually, it is also likely that the COVID-19 pandemic limited regular updates.

Table 3.1: Public health indices (PHI and CHI) definitions from the Our World in Data Dataset Metadata File [8]

Population Health Index (PHI)	Country Health Index (CHI)
***cardiovasc death rate*:** Death rate from cardiovascular disease in 2017 (annual number of deaths per 100,000 people)	***hospital beds per thousand:*** Hospital beds per 1,000 people, most recent year available since 2010
diabetes prevalence: Diabetes prevalence (% of population aged 20 to 79) in 2017	***human development index:*** A composite index measuring average achievement in three basic dimensions of human development– a long and healthy life, knowledge, and a decent standard of living
***female smokers*:** Share of women who smoke, most recent year available	***extreme poverty*:** Share of the population living in extreme poverty, most recent year available since 2010
male smokers: Share of men who smoke, most recent year available	***gdp per capita*:** Gross domestic product at purchasing power parity (constant 2011 international dollars), most recent year available
life_expectancy: Life expectancy at birth in 2019	***population density:*** Number of people divided by land area, measured in square kilometers, most recent year available
aged 65 older: Share of the population that is 65 years and older, most recent year available	***population:*** The population of the country (latest available values)
median age: Median age of the population, UN projection for 2020	

The pandemic risk scoring model presented in Table 3.2 was developed based on the data from the full dataset of 26 countries, for all selected variables in both public health indices. The range for each variable was obtained by observing the minimum and maximum values for each of the 13 features and subsequently, the values were split arithmetically into three even categories. In case of an uneven distribution of countries, the categories were adjusted accordingly. All countries were then classified based on the variable (feature) range and assigned scores across both indices. Tables with country distribution per index and score are provided in the supplement of this dissertation (Tables S198-S199). Countries with the same Pandemic Risk score (Table 3.2) of

predictive features were paired together to accommodate paired country analyses. Tables S198, S199, and S200 in the supplement provide a summary of the distribution of country pairs.

Table 3.2: The Pandemic Risk Scoring Model

Population Health Index (PHI) Feature	Range	Score	Country Heath Index (CHI) Feature	Range	Score
cardiovasc death rate per 100,000 people	≤ 204	1	*Hospital beds per thousand*	≥ 5.712	1
	205-323	2		3.966-5.711	2
	≥ 324	3		≤ 3.965	3
diabetes prevalence (%)	≤ 5.49	1	*human development index**	≥ 0.908	1
	5.50-6.99	2		0.857-0.907	2
	≥ 7	3		≤ 0.856	3
male smokers (%)	≤ 27.7	1	*extreme poverty*	< 0.20	1
	27.8-40.3	2		0.20-0.99	2
	≥ 40.4	3		> 1.00	3
female smokers (%)	≤ 20.6	1	*gdp per capita*	> 50,000	1
	20.7-29.3	2		35,000-50,000	2
	≥ 29.4	3		< 35,000	3
life expectancy (years)	≥ 80.89	1	*population density*	< 100	1
	77.97-80.88	2		100-200	2
	≤ 77.96	3		> 200	3
aged 65 older (%)	≥ 19.82	1	*population*	< 10M	1
	16.62-19.81	2		10-50M	2
	≤ 16.61	3		> 50M	3
median age (years)	≥ 44.5	1			
	40.9-44.4	2			
	≤ 40.8	3			

******human development index*: a composite index measuring average achievements in three basic dimensions of human development: a long and healthy life, education level, and a decent standard of living, values for 2019

Chapter 4

Research Part 1: Methodologies

4.1 Data Preprocessing with 70/30 train-test-split ratio

Data utilized in this research was pre-processed according to the standard methodology of assigning the original dataset to training and testing datasets. A 70/30 train-test split was used since a larger training set allows the model to learn more effectively and capture the underlying patterns in the data. This 70/30 train-test split was done for both parts of the analyses. With a larger test set, a more robust estimate of the model's performance on unseen data can be obtained. This is particularly useful when evaluating the model's generalization capabilities and making comparisons between different algorithms or hyperparameter settings [46]. For the first part of the analyses, data was analyzed at the aggregate level for all 26 countries to assess the correlation of the non-pandemic parameters to the case fatality rate variable as a general variable for all countries together. 70% of the training data represented data from the beginning of the pandemic in March 2020 to January 2022. Data was evaluated at the country level using single and paired analyses in the second part of the analyses. The remaining data from February 2022 until December 2022 was part of the 30% testing set. This accounted for the variability of the case fatality rate variable for each country across time. The 70/30 train-test split was conducted utilizing the train-test-split method in the Scikit-learn: Machine Learning library in Python [45]. Random Forest and XGBoost Regressor Models were successfully trained on the training set. To assess how well the machine learning models would perform on new data, ten-fold cross validation was performed. Data cleaning was conducted by resolving the problem of missing and duplicate values, smoothing of noisy data and resolving data inconsistencies, and removing outliers. In this type of dataset, it is

common that some data is missing, both at random and not at random. For this research, it was important that the data on the total number of cases and deaths was complete because it was used for deriving the case fatality rate. This missing data was resolved by taking the mean values of the total number of cases and deaths from the previous day and the next day. Other data was managed in a similar manner. PCA (Principal Component Analysis) was used to resolve the issue of multicollinearity between the features present in PHI and CHI indices, to improve the performance and interpretability of the machine learning models. Data transformation (normalization using Standard Scaler) was applied individually to the training and testing datasets after the train-test split operation was conducted. In addition, examinations of data quality (completeness, dependability, consistency, validity, and lack of redundancy) and feature engineering and feature selection were finished. SMOTE (Synthetic Minority Oversampling Technique) was used for oversampling in the process of data exploration and visualization [36]. Additionally, a correlation matrix of the dataset's many variables was created, and the dataset was explored through the use of graphics and visualization.

4.2 Regression Methodologies

Two sets of machine learning methodologies were applied, the first utilizing Random Forest Regressor (RFR) with distribution lag and Extreme Gradient Boosting (XGBoost) with distribution lag. One of the novel approaches in the second set was the Ordinary Least Squares Multifactor Regression (OLS MFR) model-validated K-means-Coefficient of Variance sensitivity analysis methodology. Research models in this dissertation were selected based on several considerations: 1) characteristics of the dataset (e.g., categorical data type, a non-linear relationship between the independent and dependent variables, a smaller dataset size); 2) constant

or dynamic nature of the variables over the research period; and 3) performance of selected models based on prior research and published literature. All machine learning analyses were done using Python version 3.10.1 and the scikit-learn library version 1.2.0 [45]. In addition, the pandemic risk for individual countries was evaluated utilizing a novel risk assessment scoring model.

4.21 Random Forest Regressor with Distribution Lag and Extreme Gradient Boosting Regressor with Distribution Lag

Random Forest Regressor (RFR) and Extreme Gradient Boosting Regressor (XGBoost) methodologies were applied, both enhanced by distribution lag, to assess which demographic, health and economic factors yield the highest predictors of the COVID-19 case fatality rates per country and to provide the ranking order of predictive features. RFR is a supervised learning algorithm that uses an ensemble method for regression, combining the results of many regression algorithms to enhance the model's accuracy and performance. This model is robust to outliers in the data and works well with a non-linear type of dataset, in addition to making it easier to evaluate the feature importance or the contribution, to the target variable [50]. XGBoost methodology has a built-in cross validation model that helps with overfitting, especially when working with smaller datasets. In addition, the model is more appropriate for real-life datasets, solving for missing values, and showing higher sensitivity and accuracy with a wider distribution of feature importance compared to RFR model [53]. The generated case fatality rate variable was given a distribution lag in order to enhance the performance of the RFR and XGBoost models. The short-term dynamic and long-term cumulative effects of characteristics on a response variable are well-explained by distribution lags [47], [48]. Time lag variables were created for the previous day's, week's, and month's case fatality rate using the shift() method from the Pandas Library in Python. The main

purpose of these variables was to convert the *Our World in Data* COVID-19 timeseries dataset into a supervised learning problem. This enhancement improved and created more robust predictions [47]. The analyses for both methodologies were done on the same dataset in two parts: the first part ranked all 13 predictive features across the dataset of 26 countries (aggregate analysis), and the second part analyzed the ranking order of predictive features per country (single country analysis) and in country pairs (paired analysis). The country pairs were created based on the same Pandemic Risk scores of predictive features (Table 3.2, Supplement Tables S198, S199, and S200). Both analyses reported the ranking order of features per public health indices, PHI, and CHI. Upon implementation and training of the model and evaluation of the model on the test set, the feature importance tables were obtained for each country in the first part of analyses and then for the second part for both indices (PHI and CHI). Ten-fold cross validation was evaluated to identify the best hyperparameters for training the model, to mitigate overfitting and get the best results (defined as the lowest MSE and the highest R^2 score possible) for each country and for each index. To determine which model performs better, the metrics of the two models were compared [best 10-fold cross validation score, mean squared error (MSE), R^2 score, root mean squared error (RMSE), and entropy]. The median value for each of these metrics, selected to minimize the impact of outliers, was calculated for each model and each index (PHI and CHI) and compared to the corresponding values, RFR PHI to XGBoost PHI, RFR CHI to XGBoost CHI. In addition, the distribution of the dominant predictive features that correlate the strongest with the case fatality rate was assessed across countries. For this assessment we utilized the single country analysis from the methodology with the highest accuracy and performance. The purpose of the paired country analysis was to determine whether the top three predictive factors of each country in the pair, which were based on the same Pandemic Risk score, were the same or comparable.

Comparison of single versus paired country analyses was conducted across all predictive features and the results were presented for the most dominant feature per index. Country pairs were selected to represent low, medium, and high ranges of the most correlated predictive feature.

4.22 Ordinary Least Squares Multifactor Regression

OLS Multifactor Regression (OLS MFR) model is an extension of the linear regression algorithm and is appropriate to use with complex real-world data. It is a computationally efficient model allowing for faster model training and interference, introducing multiple independent variables capable of modeling more complex relationships, and reducing the error and bias in the estimates [66], [67]. OLS MFR provides easily interpretable results allowing for insights into relationships between variables, rapid prototyping, and quick analysis. It can be used on a broad range of research questions and data types, handling continuous, discrete, and categorical predictor variables [68]. The accuracy and performance of the K-means-COV methodology approach (more detail provided in Section 4.23) was validated with OLS MFR model. The final ranking order of predictive features across several approaches was compared, and additional validation was carried out by running RFR and XGBoost analyses on the remaining features.

4.3 Clustering Methodologies

4.31 Novel K-means Coefficient of Variance Methodology Approach

A novel model approach for K-means-Coefficient of Variance (K-means-COV) sensitivity analysis was introduced to evaluate predictive features and determine their final ranking order relative to the COVID-19 case fatality rate. By adding a variation coefficient weight vector to reduce the impact of extraneous features, the COV approach was previously utilized to increase the accuracy of K-means clustering [43]. Using K-means-COV methodology for this research introduced several advantages. The K-means clustering technique is easily adaptable to new examples and cases in the dataset [59]. K-means can be applied to various data types and structures, such as numerical, categorical, and mixed data, while different sizes of clusters can be obtained relative to the dataset that is being worked with. The clusters formed by K-means are represented by their cluster centers that provide insights into the characteristics and properties of the data points within each cluster and are easily interpretable. As an unsupervised learning algorithm, K-means does not require labeled data for training. It can discover patterns and structures in data without the need for prior labeling [60]. COV is an efficient model used to compare the variability of different features in the dataset to obtain the strength of correlation of those features. This model allows for the comparison of variability between different variables (features), even if they have different scales or units of measurement. It provides a standardized measure to assess the relative dispersion of data points, making it useful for comparing datasets with diverse characteristics [61]. A thorough understanding of the relationship between independent (input) and dependent (output) variables can be obtained by using K-means-COV sensitivity analysis. This methodology confirms

the prediction outcomes of more conventional and typical machine learning models and tests and evaluates the robustness of the results [62].

In this dissertation, K-means-COV was used in two different approaches. The first part of analyses (aggregate) clustered 26 countries based on the 13 predictive features. The second part ranked the predictive features by clustering countries based on public health indices (PHI and CHI). To determine the optimal number of clusters for the K-Means clustering methodology, the Elbow method was employed. The graph of Within-Cluster Sum of Squares (WCSS) was developed, based on the sum of the squared distance between each point and the centroid in a cluster versus the Number of Clusters. The elbow point, the point at which the rate of decrease of WCSS is minimized, was used to determine the optimal number of clusters for the K-Means algorithm. The country-to-country difference of each clustering feature was calculated and averaged to obtain the mean difference for each clustering feature. Furthermore, the standard deviation, sum of squared deviations from the mean, was calculated by taking the difference between the actual values for each clustering feature for each country-to-country comparison and the mean of the country-to-country difference. The coefficient of variance values for each of the features can be calculated by dividing σ (the standard deviation from the mean of the country-to-country difference) by μ (the mean of the country-to-country difference). The ratio of those two values yielded the coefficient of variance for each respective clustering feature. The results of these analyses yielded predictive features that correlated most highly with the case fatality rate and were then compared to Ordinary Least Squares Multifactor Regression (OLS MFR).

4.32 Calinski-Harabasz Methodology

In 1974, Calinski and Harabasz devised an index called the Calinski-Harabasz clustering approach. When ground truth labels are unknown and the effectiveness of the clustering is validated using dataset-specific quantities and features, this index can be used to assess the clustering model. The Variance Ratio Criterion, or Calinski-Harabasz (CH) Index, gauges an object's cohesiveness (similarity to its own cluster) in relation to its separation (difference from other clusters). Similar to the Elbow Method for determining the ideal number of clusters, cohesion is estimated here based on the distances between data points in a cluster and its cluster centroid, and separation is based on the distances between the cluster centroids and the global centroid [156]. The same strategy was used, employing the Calinski-Harabasz metric and methodology to ascertain the ideal number of clusters for both per index, PHI and CHI clustering analysis (clustering 26 countries based on PHI and CHI Index variables separately), and aggregate clustering analysis (clustering 26 countries on all 13 features combined). The result of clustering 26 countries and determining the optimal number of clusters using Calinski-Harabasz method and metric and the Elbow method yielded the same results at the end of each type of analysis.

4.4 Pandemic Risk Scoring Model

The Pandemic Risk Scoring Model was developed to assess country pandemic readiness. All predictive features were assigned a score, and the total score per index and per country was calculated. The country scores were then classified into risk categories (low, medium, or high), as described in Section 4.21. The distribution of countries based on their total PHI and CHI risk scores is shown in the results section.

Chapter 5

Research Part 1: Results

Two sets of machine learning methodologies were utilized to assess the most accurate methodology in predicting the ranking order of the features correlating the most with the case fatality rate. The first set included RFR and XGBoost, both enhanced with distribution lag, and the second set included a novel approach with K-means-COV and OLS Multifactor Regression. All methodologies were assessed for accuracy and performance and compared in a descriptive way. In addition, countries were assessed for their pandemic risk utilizing a novel pandemic risk scoring model.

5.1 Regression Methodology Results

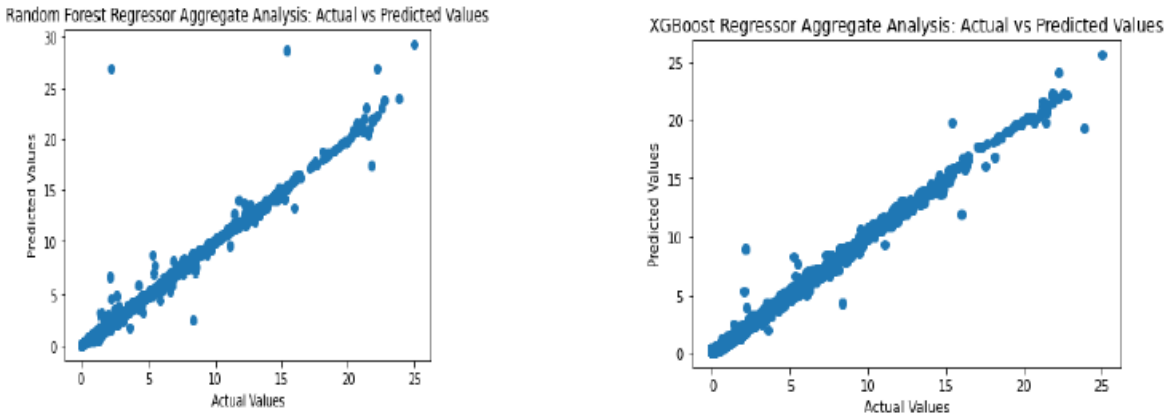
5.11 Random Forest and Extreme Gradient Boosting Regressor Results

The first part of analyses, utilizing RFR and XGBoost methodologies, ranked all 13 predictive features across the dataset of 26 countries (Table 5.1). The performed analyses indicated that the feature *aged 65 older* was the highest-ranking predictive feature for both RFR and XGBoost analyses, while the importance value was higher in the RFR analysis (0.8791) versus XGBoost (0.4394). This feature was followed in importance by *extreme poverty* and *hospital beds per thousand* for RFR, and with *population density* and *extreme poverty* for XGBoost. The accuracy of the performance of the two methodologies was assessed and presented in Table 5.1 and Figure 5.1. The accuracy and performance of both methods in the first part of the analysis was high and similar. XGBoost Regressor model performed better, with three out of five metrics (MSE,

R^2, RMSE) favoring XGBoost model and a better linear relationship of Actual vs Predicted Values [Figure 5.1]. The second part of RFR and XGBoost analyses analyzed the ranking order of predictive features per country (single country analysis) and in country pairs (paired analysis). The country pairs were created based on the same Pandemic Risk scores of predictive features (Section 4.21).

Table 5.1: Rank of predictive features utilizing RFR and XGBoost Methodologies

Random Forest Regressor Analysis		**XGBoost Regressor Analysis**	
Features	**Importance Values**	**Features**	**Importance Values**
aged 65 older	0.8792	*aged 65 older*	0.4395
extreme poverty	0.0305	*population density*	0.1083
hospital beds per thousand	0.0244	*extreme poverty*	0.0975
life expectancy	0.0172	*hospital beds per thousand*	0.0759
median age	0.0139	*life expectancy*	0.0758
population	0.0118	*female smokers*	0.0405
cardiovasc death rate	0.0055	*cardiovasc death rate*	0.0324
female smokers	0.0053	*diabetes prevalence*	0.0313
human development index	0.0046	*median age*	0.0289
gdp per capita	0.0025	*population*	0.0265
male smokers	0.0022	*male smokers*	0.0203
population density	0.0018	*gdp per capita*	0.0181
diabetes prevalence	0.0005	*human development index*	0.0043
Metrics for RFR Aggregate Analysis		**Metrics for XGBoost Aggregate Analysis**	
Best hyperparameters: {'max_depth': 15, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 200}		Best hyperparameters: {'colsample_bytree': 0.8, 'gamma': 0.2, 'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 150, 'subsample': 0.9}	
Best CV score: 0.9655		Best CV score: 0.9561	
MSE: 0.1266		MSE: 0.04643	
R^2 Score: 0.9855		R^2 Score: 0.9946	
RMSE: 0.3558		RMSE: 0.2154	
Entropy Value: 0.0026		Entropy Value: 0.0061	



(a) Random Forest Regressor (b) XGBoost Regressor

Figure 5.1: Comparison of actual vs predicted values for aggregate analyses.

Both analyses reported the ranking order of features per public health indices, PHI and CHI. The summary results of single country analyses, with the most common top three predictive features, are presented in Table 5.2, with detailed information presented in the supplement of this dissertation.

Table 5.2: Summary of RFR and XGBoost single country analysis results

Random Forest Regressor Model				XGBoost Regressor Model			
PHI	**Importance range**	**CHI**	**Importance range**	**PHI**	**Importance range**	**CHI**	**Importance range**
cardiovasc death rate	0.0174- 0.9726	*hospital beds per thousand*	0.0020-0.9750	*cardiovasc death rate*	0.0166-0.9133	*hospital beds per thousand*	0.0322-0.6382
aged 65 older	0.0011- 0.9426	*human development index*	0.0002-0.0780	*life expectancy*	0.00008-0.8916	*population*	0.0001-0.9605
diabetes prevalence	0.0013- 0.5292	*population*	0.0001-0.9654	*diabetes prevalence*	0.0025-0.0908	*human development index*	0.0006-0.1284
	PHI	**CHI**			**PHI**	**CHI**	
Best 10-fold Cross Validation Score	0.9119-0.9988	0.9188-0.9996		**Best 10-fold Cross Validation Score**	0.8763-0.9995	0.9000-0.9997	
	Median: 0.9960	Median: 0.9962			Median: 0.9981	Median: 0.9978	
Mean Squared Error (MSE)	0.0001-2.819	0.0002-5.438		**Mean Squared Error (MSE)**	0.00006-9.026	0.0001-14.92	
	Median: 0.0059	Median: 0.0087			Median: 0.0037	Median: 0.0048	
R^2 Score	0.5821-0.9992	0.6713-0.9994		**R^2 Score**	0.6692-0.9996	0.6747-0.9997	
	Median: 0.9963	Median: 0.9951			Median: 0.9980	Median: 0.9972	
RMSE	0.0112-1.679	0.01587-2.332		**RMSE**	0.0082-3.004	0.0103-3.862	
	Median: 0.07704	Median: 0.0937			Median: 0.0610	Median: 0.0699	
Entropy	0.0001-0.01499	0.0002-0.0143		**Entropy**	0.00007-0.02115	0.0001-0.02727	
	Median: 0.0007	Median: 0.0008			Median: 0.0004	Median: 0.0006	

The single country analyses indicated that *cardiovascular death rate*, *aged 65 older*, and *diabetes prevalence* are the most common top predictive features for PHI utilizing RFR analysis. Similarly, *cardiovascular death rate*, *life expectancy*, and *diabetes prevalence* were the most common features for PHI utilizing XGBoost. The top three predictive features for CHI with RFR analysis were *hospital beds per thousand, human development index,* and population. Both methodologies performed with high accuracy, with XGBoost performing better on all five metrics. In addition, this research looked at the distribution of dominant predictive features, correlating the

most with the case fatality rate across countries (Table 5.3). The single country analyses performed with XGBoost were utilized for this assessment. Results for CHI with XGBoost analysis were similar, with *hospital beds per thousand people, population*, and *human development index.* The accuracy of performance of the two methodologies was assessed and presented in Table 5.2 and Figure 5.2.

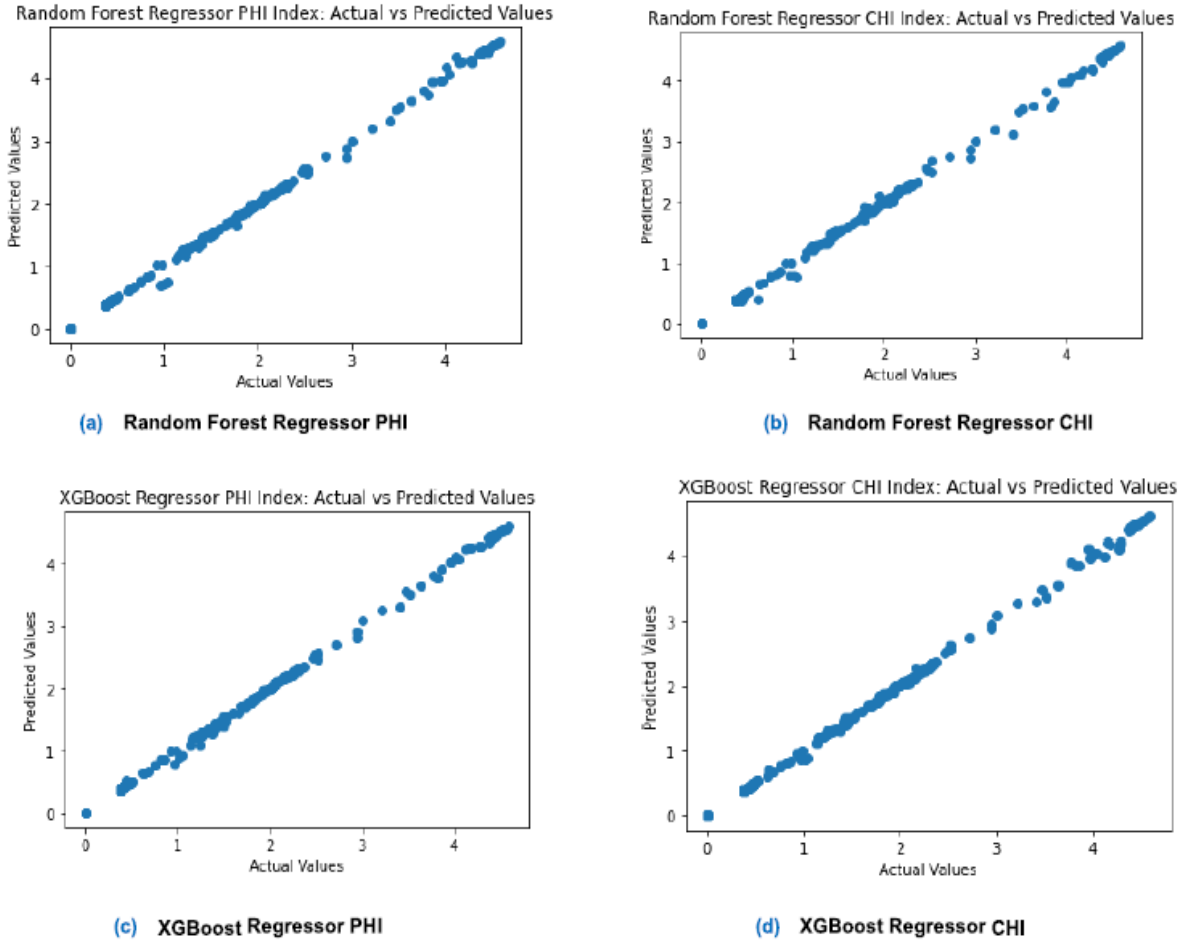


Figure 5.2: Comparison of actual vs predicted values for single country analyses.

Table 5.3: Distribution of dominant predictive features across countries

Distribution of dominant predictive features across countries					
PHI			**CHI**		
feature	**countries (total, %)**	**countries**	**feature**	**countries (total, %)**	**countries**
cardiovascular death rate	12 (46%)	Bulgaria, Czechia, France, Finland, Ireland, Latvia, Portugal, Serbia, Slovakia, Sweden, Switzerland, United States	*hospital beds per thousand*	12 (46%)	Canada, Cyprus, Estonia, Finland, Ireland, Latvia, Portugal, Slovakia, Slovenia, Sweden, Switzerland, UK
aged 65 older	6 (23%)	Austria, Belgium, Canada, Italy, Luxemburg, Slovenia	*population*	10 (39%)	Czechia, Denmark, France, Iceland, Italy, Luxemburg, Netherlands, Romania, Serbia, Spain
life expectancy	6 (23%)	Cyprus, Denmark, Finland, Netherlands, Spain, UK	*population density*	4 (15%)	Austria, Belgium, Bulgaria, US
median age	2 (8%)	Estonia, Romania			

In summary, the *cardiovascular death rate* feature correlates most the strongly with the case fatality rate for 46% of all countries, within the Population Health Index. Similarly, the *hospital beds per thousand* feature has the highest correlation for 46% of countries, within the Country Health Index. The summary results of paired country analyses are presented in Table 5.4, with detailed tables presented in the supplement of this dissertation.

Table 5.4: Summary of RFR and XGBoost paired country analysis results.

Random Forest Regressor Model				**XGBoost Regressor Model**			
PHI	**Importance range**	**CHI**	**Importance range**	**PHI**	**Importance range**	**CHI**	**Importance range**
diabetes prevalence	0.0049-0.9735	*human development index*	0.0011-0.9747	*diabetes prevalence*	0.0129-0.9437	*human development index*	0.0052-0.9287
cardiovascular death rate	0.0008-0.9692	*extreme poverty*	0.0135-0.9739	*cardiovascular death rate*	0.0014-0.6680	*hospital beds per thousand*	0.0005-0.8628
female smokers	0.0008-0.9735	*hospital beds per thousand*	0.0000001-0.7646	*median age*	0.0003-0.8860	*population*	0.0009-0.7469
	PHI	**CHI**			**PHI**	**CHI**	
Best 10-fold Cross Validation Score	0.9443-0.9994	0.9340-0.9993		**Best 10-fold Cross Validation Score**	0.9363-0.9995	0.9274-0.9994	
	Median:	Median:			Median:	Median:	
	0.9974	0.9972			0.9985	0.9979	
Mean Squared Error (MSE)	0.0004-8.727	0.0006-6.845		**Mean Squared Error (MSE)**	0.0004-5.700	0.0008-5.255	
	Median:	Median:			Median:	Median:	
	0.0069	0.0081			0.0052	0.0079	
R^2 Score	0.7734-0.9995	0.8223-0.9994		**R^2 Score**	0.8520-0.9997	0.8635-0.9995	
	Median:	Median:			Median:	Median:	
	0.9979	0.9976			0.9985	0.9978	
Root Mean Squared Error (RMSE)	0.0217-2.954	0.02620-2.616		**Root Mean Squared Error (RMSE)**	0.0204-2.387	0.0293-2.292	
	Median:	Median:			Median:	Median:	
	0.0834	0.09032			0.0724	0.0878	
Entropy	0.0001-0.0275	0.0001-0.01839		**Entropy**	0.0001-0.01429	0.0001-0.0249	
	Median: 0.0006	Median: 0.0007			Median: 0.0005	Median: 0.0009	

Similar to the XGBoost PHI results, the matched country analyses showed that the top three predictive factors for the PHI using RFR technique are female smokers, diabetes prevalence, and cardiovascular death rate. The RFR CHI results list *human development index, extreme poverty*, and *hospital beds per thousand*, while XGBoost lists *human development index*, *hospital beds per thousand*, and *population* as the most predictive features. Both methodologies performed with high accuracy, with XGBoost performing better on all five metrics. The accuracy and performance of both models for the second part of analyses was performed and documented in Table 5.4 and Figure 5.3.

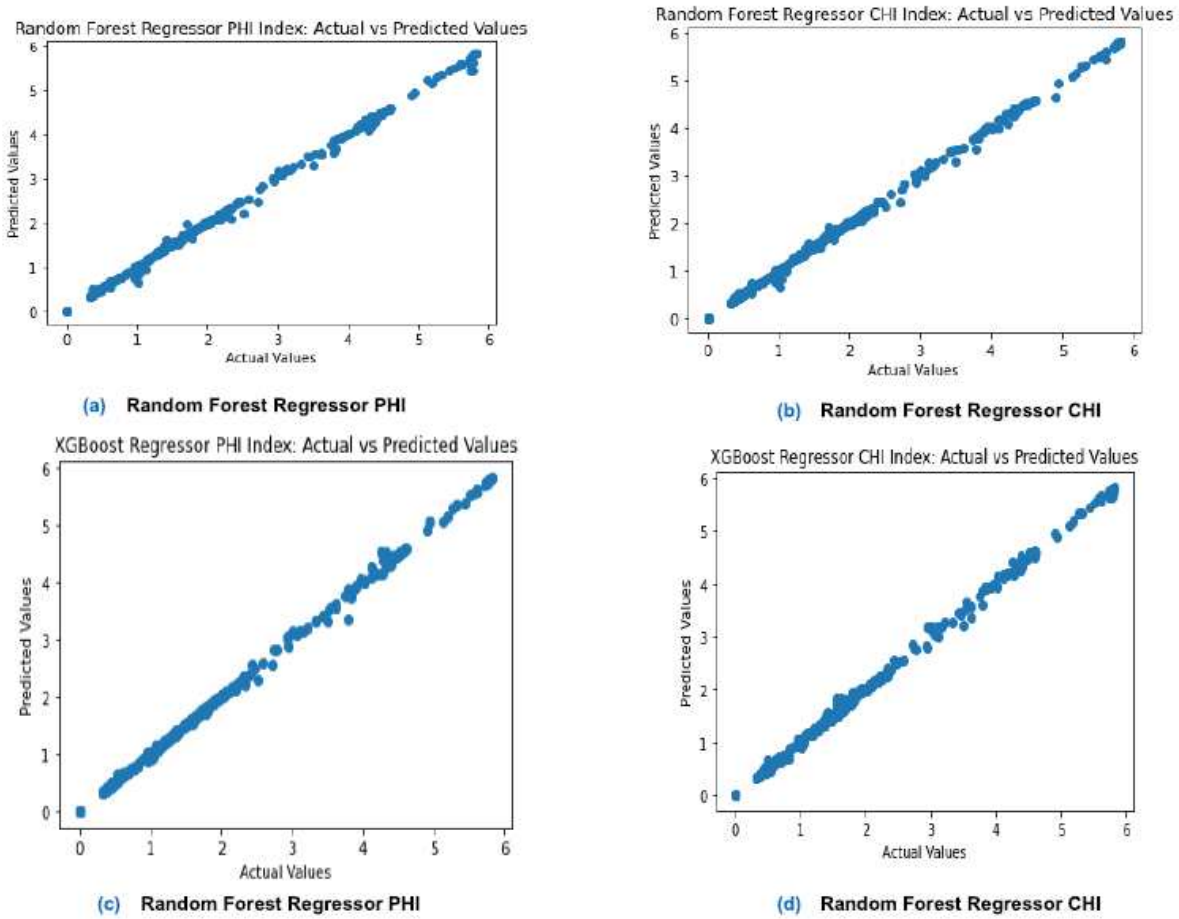


Figure 5.3: Comparison of actual vs predicted values for paired country analyses.

According to the distribution of feature significance values, which was used to evaluate the sensitivity of the two models, XGBoost is a more variable model with a larger distribution across all features. In addition, the similarity of the top three predictive features was assessed from the single country analysis versus paired country analysis focusing on *cardiovascular death rate* as the dominant predictive feature for PHI (Table 5.5). Three country pairs were selected to represent the distribution across low, medium, and high *cardiovascular death rate* ranges.

Table 5.5: Single vs Paired country analyses based on cardiovascular death rate.

Single vs paired country analyses; feature: *cardiovascular death rate* per 100,000 people					
		RFR		**XGBoost**	
Single/paired country	**Actual *cardiovascular death rate***	**PHI**	**CHI**	**PHI**	**CHI**
United Kingdom	122.137	*aged 65 older, diabetes prevalence, median age*	*hospital beds per thousand, population density, population*	*life expectancy, aged 65 older, diabetes prevalence*	*hospital beds per thousand, population, population density*
United States	151.089	*life expectancy, diabetes prevalence, aged 65 older*	*population density, hospital beds per thousand, human development index*	*aged 65 older, median age, life expectancy*	*population density, hospital beds per thousand, population*
United Kingdom/United States	(low range: ≤ 204)	*diabetes prevalence, female smokers, median age*	*human development index, extreme poverty, population*	*diabetes prevalence, median age, male smokers*	*hospital beds per thousand, human development index, population*
Slovakia	287.959	*median age, diabetes prevalence, aged 65 older*	*hospital beds per thousand, human development index, population density*	*life expectancy, median age, aged 65 older*	*hospital beds per thousand, population, human development index*
Slovenia	153.493	*aged 65 older, life expectancy, diabetes prevalence*	*hospital beds per thousand, human development index, population density*	*aged 65 older, life expectancy, diabetes prevalence*	*hospital beds thousand, population, human development index*
Slovakia/Slovenia	(mid-range: 205-323)	*female smokers, diabetes prevalence, male smokers*	*human development index, population, extreme poverty*	*female smokers, diabetes prevalence, median age*	*human development index, population, extreme poverty*
Romania	370.946	*median age, life expectancy, female smokers*	*hospital beds per thousand, human development index, extreme poverty*	*median age, life expectancy, diabetes prevalence*	*population, hospital beds per thousand, human development index*
Serbia	439.415	*diabetes prevalence, aged 65 older, life expectancy*	*population, hospital beds per thousand, extreme poverty*	*aged 65 older, diabetes prevalence, life expectancy*	*population, hospital beds per thousand, population density*
Romania/Serbia	(high range: ≥ 324)	*diabetes prevalence, median age, female smokers*	*population, human development index, extreme poverty*	*diabetes prevalence, median age, female smokers*	*population, hospital beds per thousand, human development index*

Single versus paired country analyses showed similarities in the ranking order of features across the two methodologies. Better accuracy of the XGBoost model over RFR is confirmed by the XGBoost PHI and CHI analyses, which show that two of the three features from the single country analysis were included in the ranking order of features in the paired analysis. As the primary predictive characteristic for CHI, the hospital beds per thousand persons feature was also subjected to single versus paired analysis, with comparable findings (Table S203 in the supplement).

5.2 Clustering Methodology Results

5.21 K-means-Coefficient of Variance Sensitivity and Ordinary Least Squares Multifactor Regression Analysis Results

The second set of machine learning methodologies included a novel K-means-COV sensitivity analysis approach and OLS MFR. The first part of analyses ranked predictive features across the dataset of 26 countries clustered on 13 features utilizing K-means methodology (Table 5.6). The Elbow methodology was utilized to determine that the optimal number of clusters was two (K = 2). Based on the inverse relationship between the COV values and the OLS Feature Importance values for the analyzed features, the K-means clustering-COV sensitivity analysis model had to be repeated several times. Before each new iteration, the three features with the highest COV values (greater than 1) were removed and the process was repeated for the remaining features, for a total of six iterations. The scatter plots of feature importance for both the first and last iterations are presented in Figure 5.4. The final importance rank for the first part of analyses (aggregate analyses) was defined based on the line graph of the last iteration of Coefficient of

Variance versus Ordinary Least Squares (OLS) Multifactor Regression Feature Importance, demonstrating a linear relationship between the remaining features.

Table 5.6: Summary of results across K-means-COV and OLS MFR with comparison, First part of analyses, first and last iteration

First Iteration					
K-means clustering with COV Sensitivity Analysis		**OLS Multifactor Regression**		**K-means clustering with COV vs Feature Importance Values**	
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
hospital beds per thousand	0.6874	*cardiovascular death rate*	0.4007	0.6874	0.4007
human development index	0.7262	*population*	0.2139	0.7262	0.2139
diabetes prevalence	0.8086	*extreme poverty*	0.1649	0.8086	0.1649
female smokers	0.8163	*human development index*	0.1592	0.8163	0.1592
median age	0.8653	*life expectancy*	0.1482	0.8653	0.1482
life expectancy	0.9395	*diabetes prevalence*	0.1462	0.9395	0.1462
aged 65 older	0.9606	*hospital beds per thousand*	0.1419	0.9606	0.1419
GDP per capita	0.9942	*female smokers*	0.124	0.9942	0.124
cardiovascular death rate	0.9947	*male smokers*	0.1163	0.9947	0.1163
population density	0.9989	*aged 65 older*	0.05644	0.9989	0.0564
male smokers	1.057	*GDP per capita*	0.0494	1.057	0.0494
extreme poverty	1.567	*population density*	0.0268	1.567	0.0268
population	1.678	*median age*	0.0217	1.678	0.0217
Last iteration					
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
female smokers	0.6668	*female smokers*	0.1301	0.6668	0.1301
diabetes prevalence	0.7123	*diabetes prevalence*	0.1204	0.7123	0.1204

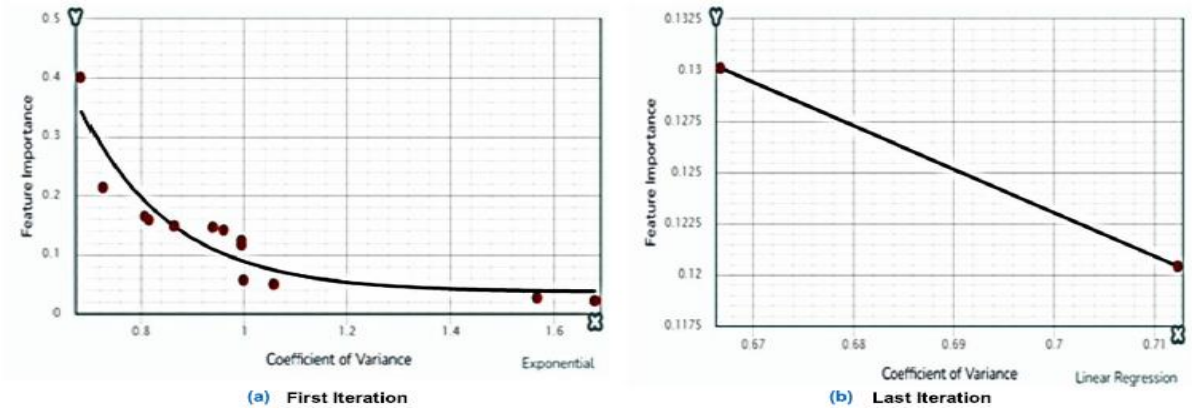


Figure 5.4: K-means-Coefficient of variance (COV) versus ordinary least squares (OLS) multifactor regression feature importance scatter plot, first and last iteration

The second part of analyses with the K-means COV sensitivity and OLS MFR analyses ranked the predictive features by clustering countries based on features grouped into the public health indices (PHI and CHI) utilizing K-means methodology (Table 5.7). The scatter plots of feature importance for both the first and last iterations are presented in Figures 5.5 and 5.6. The line graph of the final COV against OLS MFR Feature Importance iteration, which showed a linear relationship between the remaining features, was used to define the final importance rank for the second portion of analysis (per PHI and CHI).

Table 5.7: Summary of results across K-means-COV and OLS MFR with comparison, Second part of analyses, first and last iteration

First Iteration					
K-means clustering with COV Sensitivity Analysis		**OLS Multifactor Regression**		**K-means clustering with COV vs Feature Importance Values**	
PHI					
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
female smokers	0.7271	*cardiovascular death rate*	0.2521	0.7271	0.2521
diabetes prevalence	0.7448	*life expectancy*	0.2115	0.7448	0.2115
male smokers	0.8146	*male smokers*	0.1238	0.8146	0.1238
median age	0.9009	*median age*	0.1090	0.9009	0.1090
aged 65 older	0.9512	*female smokers*	0.0600	0.9512	0.0600
life expectancy	0.9735	*aged 65 older*	0.0290	0.9735	0.0290
cardiovascular death rate	1.0789	*diabetes prevalence*	0.0063	1.0789	0.0063
CHI					
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
human development index	0.7278	*extreme poverty*	0.2358	0.7278	0.2358
hospital beds per thousand	0.8165	*human development index*	0.1392	0.8165	0.1392
population density	0.9181	*hospital beds per thousand*	0.1052	0.9181	0.1052
GDP per capita	0.9408	*population*	0.0984	0.9408	0.0984
extreme poverty	1.4563	*population density*	0.0359	1.4563	0.03595
population	1.6839	*GDP per capita*	0.0357	1.6839	0.0357
Last Iteration					
K-means clustering with COV Sensitivity Analysis		**OLS Multifactor Regression**		**K-means clustering with COV vs Feature Importance Values**	
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
female smokers	0.7482	*hospital beds per thousand*	0.2608	0.7482	0.2608
hospital beds per thousand	0.8256	*GDP per capita*	0.1778	0.8256	0.1778
GDP per capita	0.8890	*female smokers*	0.1133	0.8890	0.1133

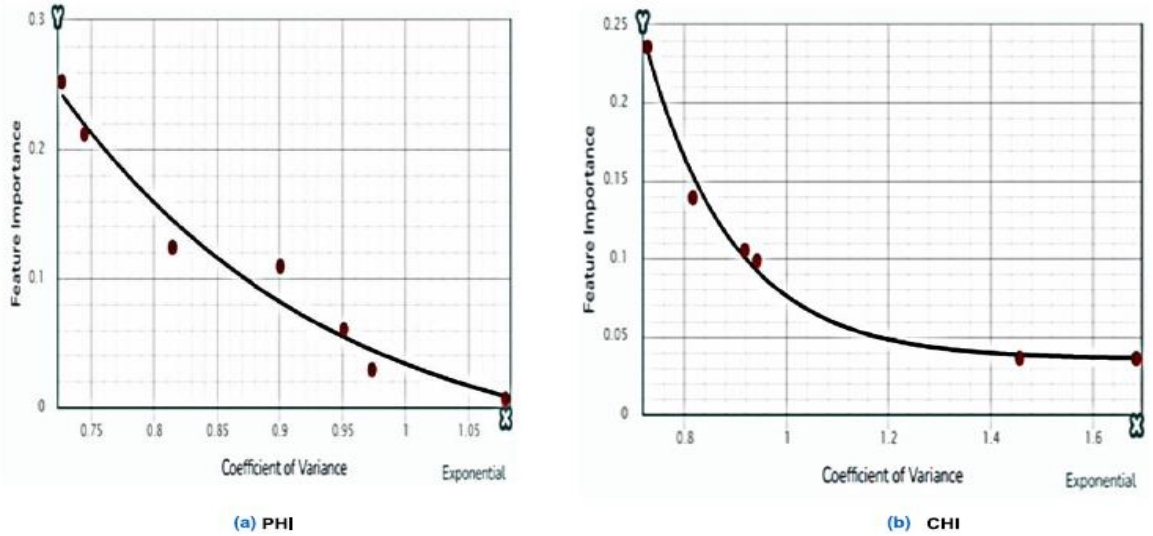


Figure 5.5: K-means-Coefficient of variance (COV) versus ordinary least squares (OLS) multifactor regression feature importance scatter plot, first iteration

Figure 5.6: K-means-Coefficient of variance (COV) versus ordinary least squares (OLS) multifactor regression feature importance line graph, last iteration

Validating the novel K-means-COV sensitivity methodology approach, both sets of analyses using the K-means-COV sensitivity analysis model and Ordinary Least Squares Multifactor Regression confirmed a linear relationship between the final remaining features and the alignment of the ranking order of predictive features. Additional validation was conducted with RFR and XGBoost methodologies utilizing the remaining features from the K-means-COV

analysis, defining the final importance ranks for the aggregate analyses, and showing similar results in the final ranking order of predictive features (Table 5.8).

Table 5.8: Final ranking order of predictive features across RFR, XGBoost, K-means clustering with COV, and OLS MFR (Aggregate Analyses)

Random Forest Regressor Analysis with remaining features		**XGBoost Regressor Analysis with remaining features**	
Features	**Importance Values**	**Features**	**Importance Values**
diabetes_prevalence	0.9208	*diabetes_prevalence*	0.7662
female_smokers	0.0791	*female_smokers*	0.2337
K-means clustering with COV Sensitivity Analysis		**OLS Multifactor Regression Analysis**	
Features	**Coefficient of Variance**	**Features**	**Importance Values**
female_smokers	0.6668	*female_smokers*	0.1301
diabetes_prevalence	0.7123	*diabetes_prevalence*	0.1203

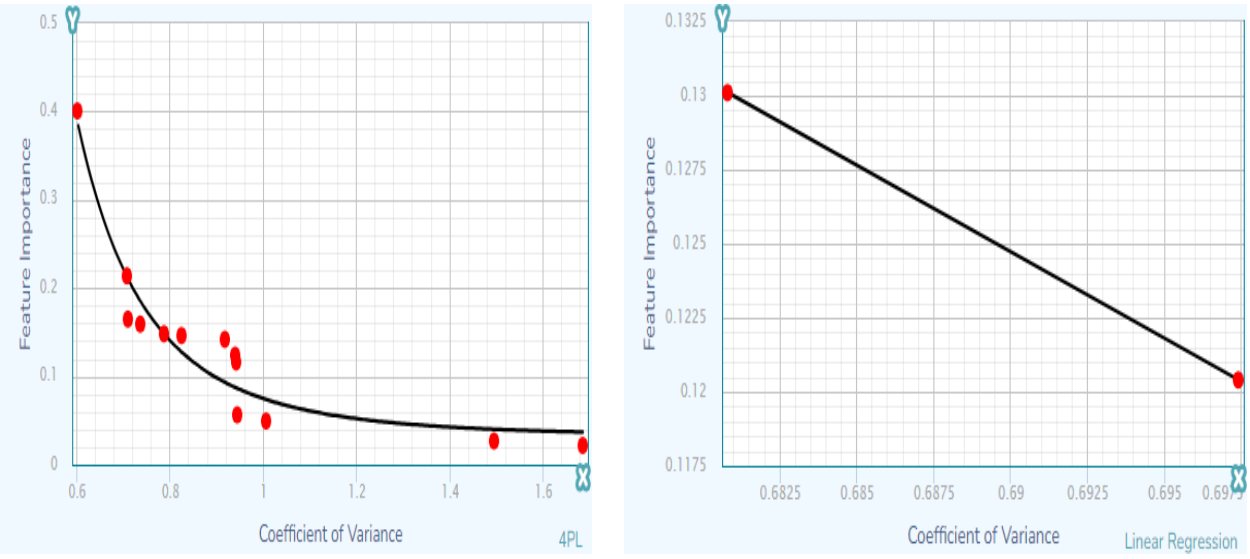
5.22 Calinski-Harabasz Methodology Results

In addition to the novel K-means-Coefficient of Variance clustering methodology used to cluster 26 countries according to all variables together (aggregate analysis) and per PHI and CHI indices, the Calinksi-Harabasz methodology-Coefficient of Variance approach was used to accomplish the same task and the results were compared with the K-means-COV results, discussed in Section 5.21. The first part of analyses ranked predictive features across the dataset of 26 countries clustered on 13 features utilizing Calinksi-Harabasz clustering methodology (Table 5.9). Instead of the Elbow Method that was used before to determine the optimal number of clusters for the 26 countries, the optimal number of clusters was determined using the Calinski-Harabasz methodology and metric. Based on the inverse relationship between the COV values and the OLS Feature Importance values for the analyzed features, the Calinski-Harabasz

clustering-COV sensitivity analysis model had to be repeated several times. Before each new iteration, the three features with the highest COV values (greater than 1) were removed and the process was repeated for the remaining features, for a total of six iterations. The scatter plots of feature importance for both the first and last iterations are presented in Figure 5.7. The final importance rank for the first part of analyses (aggregate analyses) was defined based on the line graph of the last iteration of Calinski Harabasz Clustering-Coefficient of Variance versus Ordinary Least Squares (OLS) Multifactor Regression Feature Importance, demonstrating a linear relationship between the remaining features.

Table 5.9: Summary of results across Calinski-Harabasz clustering-COV and OLS MFR with comparison, First part of analyses, first and last iteration

First Iteration					
Calinski-Harabasz clustering with COV Sensitivity Analysis		**OLS Multifactor Regression**		**Calinski-Harabasz clustering with COV vs Feature Importance Values**	
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
female smokers	0.6022	*cardiovasc death rate*	0.4007	0.6022	0.4007
median age	0.7064	*population*	0.2139	0.7064	0.2139
human development index	0.7107	*extreme poverty*	0.1649	0.7107	0.1649
hospital beds per thousand	0.7352	*human development index*	0.1592	0.7352	0.1592
male smokers	0.7872	*life expectancy*	0.1482	0.7872	0.1482
aged 65 older	0.8251	*diabetes prevalence*	0.1462	0.8251	0.1462
population density	0.9181	*hospital beds per thousand*	0.1419	0.9181	0.1419
diabetes prevalence	0.9402	*female smokers*	0.124	0.9402	0.124
gdp per capita	0.9409	*male smokers*	0.1163	0.9409	0.1163
cardiovasc death rate	0.9444	*aged 65 older*	0.05644	0.9444	0.0564
life expectancy	1.0061	*gdp per capita*	0.0494	1.0061	0.0494
extreme poverty	1.4948	*population density*	0.0268	1.4948	0.0268
population	1.6840	*median age*	0.0217	1.6840	0.0217
Last iteration					
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
female smokers	0.6808	*female smokers*	0.1301	0.6808	0.1301
diabetes prevalence	0.6974	*diabetes prevalence*	0.1204	0.6974	0.1204



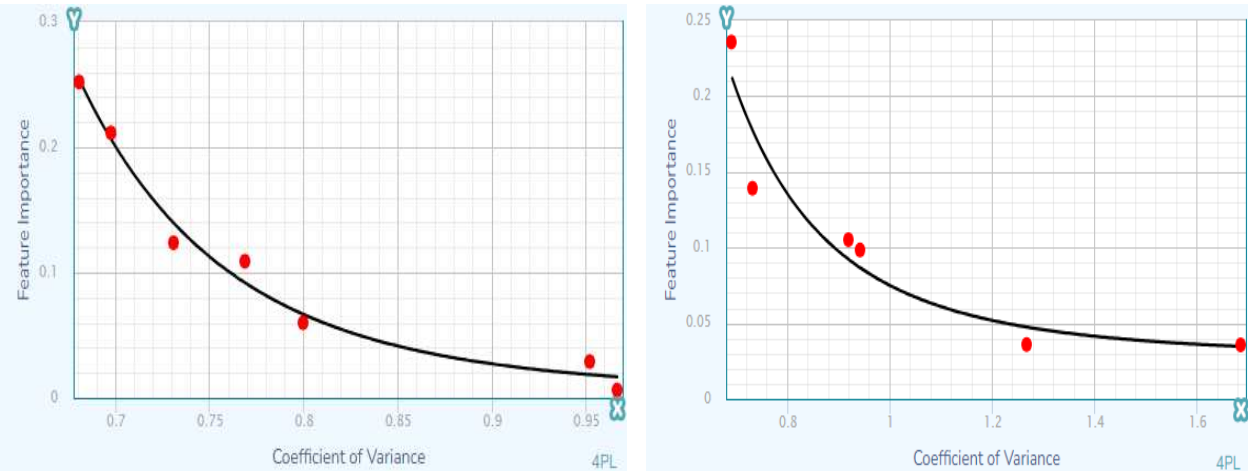
(a) First Iteration **(b) Last Iteration**

Figure 5.7: Calinski-Harabasz Clustering-Coefficient of variance (COV) versus ordinary least squares (OLS) multifactor regression feature importance scatter plot, first and last iteration

The second part of analyses with the Calinski-Harabasz clustering-COV sensitivity and OLS MFR analyses ranked the predictive features by clustering countries based on features grouped into the public health indices (PHI and CHI) utilizing K-means methodology (Table 5.10). The feature importance scatter plots for both the first and last iterations are presented in Figures 5.8 and 5.9. Based on a linear relationship between the remaining features and the line graph of the final iteration of Calinski-Harabasz clustering-COV vs OLS MFR Feature Importance, the final importance rank for the second portion of analyses (per PHI and CHI) was determined.

Table 5.10: Summary of results across Calinski-Harabasz clustering-COV and OLS MFR with comparison, Second part of analyses, first and last iteration

First Iteration					
Calinski-Harabasz clustering with COV Sensitivity Analysis		**OLS Multifactor Regression**		**Calinski-Harabasz clustering with COV vs Feature Importance Values**	
PHI					
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
female smokers	0.6808	*cardiovascular death rate*	0.2521	0.6808	0.2521
diabetes prevalence	0.6974	*life expectancy*	0.2115	0.6974	0.2115
median age	0.7310	*male smokers*	0.1238	0.7310	0.1238
male smokers	0.7686	*median age*	0.1090	0.7686	0.1090
aged 65 older	0.7994	*female smokers*	0.0600	0.7994	0.0600
cardiovascular death rate	0.9519	*aged 65 older*	0.0290	0.9519	0.0290
life expectancy	0.9663	*diabetes prevalence*	0.0063	0.9663	0.0063
CHI					
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
human development index	0.6912	*extreme poverty*	0.2358	0.6912	0.2358
hospital beds per thousand	0.7309	*human development index*	0.1392	0.7309	0.1392
population density	0.9181	*hospital beds per thousand*	0.1052	0.9181	0.1052
gdp per capita	0.9409	*population*	0.0984	0.9409	0.0984
extreme poverty	1.2678	*population density*	0.0359	1.2678	0.03595
population	1.6840	*gdp per capita*	0.0357	1.6840	0.0357
Last Iteration					
Calinski-Harabasz clustering with COV Sensitivity Analysis		**OLS Multifactor Regression**		**Calinski-Harabasz clustering with COV vs Feature Importance Values**	
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
female smokers	0.6480	*hospital beds per thousand*	0.2608	0.6480	0.2608
hospital beds per thousand	0.7010	*gdp per capita*	0.1778	0.7010	0.1778
gdp per capita	0.7635	*female smokers*	0.1133	0.7635	0.1133



(a) PHI (b) CHI

Figure 5.8: Calinski-Harabasz Clustering-Coefficient of variance (COV) versus ordinary least squares (OLS) multifactor regression feature importance scatter plot, first iteration

Figure 5.9: Calinski-Harabasz Clustering-Coefficient of variance (COV) versus ordinary least squares (OLS) multifactor regression feature importance scatter plot, last iteration

Both parts of analyses with Calinski-Harabasz clustering-COV sensitivity analysis model and Ordinary Least Squares Multifactor Regression confirmed a linear relationship between the final remaining features and the alignment of the ranking order of predictive features, validating the novel Calinski-Harabasz clustering approach as well as the novel K-means-COV sensitivity methodology approach. The Calinski-Harabasz clustering ranking order that was obtained for both

parts of analyses were aligned with the ranking order that was obtained from the novel K-means-COV sensitivity methodology approach, confirming the accuracy of the results obtained. Additional validation was conducted with RFR and XGBoost methodologies utilizing the remaining features from the Calinski-Harabasz clustering-COV analysis, defining the final importance ranks for the aggregate analyses, and showing similar results in the final ranking order of predictive features (Table 5.11).

Table 5.11: Final ranking order of predictive features across RFR, XGBoost, Calinski-Harabasz clustering with COV, and OLS MFR (Aggregate Analyses)

Random Forest Regressor Analysis with remaining features		**XGBoost Regressor Analysis with remaining features**	
Features	**Importance Values**	**Features**	**Importance Values**
diabetes_prevalence	0.9208	*diabetes_prevalence*	0.7662
female_smokers	0.0791	*female_smokers*	0.2337
Calinski-Harabasz clustering with COV Sensitivity Analysis		**OLS Multifactor Regression Analysis**	
Features	**Coefficient of Variance**	**Features**	**Importance Values**
female_smokers	0.6808	*female_smokers*	0.1301
diabetes_prevalence	0.6974	*diabetes_prevalence*	0.1203

5.3 Pandemic Risk Scoring Model Results

The Pandemic Risk Scoring Model was developed based on the feature ranges (see Table 3.2). The total score for each country allows classification into low, medium, or high-risk categories per public health index (PHI, CHI). The distribution of countries based on their total PHI and CHI scores is presented in Table 5.12 and Figure 5.10.

Table 5.12: Distribution of countries based on the total PHI and CHI scores.

Pandemic Risk score range	Country Distribution	Pandemic Risk score range	Country Distribution
Population Health Index		**Country Health Index**	
High: 17-21	Slovakia, Serbia, Romania	High: 14-18	Italy, Portugal, Spain, United Kingdom
Medium: 12-16	Czechia, United States, Cyprus, Bulgaria, Estonia, Latvia, Spain, France, Luxembourg, Austria, Ireland, Switzerland	Medium: 10-13	Belgium, Bulgaria, Canada, Cyprus, Czechia, Denmark, Estonia, France, Iceland, Latvia, Luxembourg, Netherlands, Romania, Serbia, Slovakia, Slovenia, Sweden, United States
Low: 7-11	Italy, Portugal, United Kingdom, Netherlands, Belgium, Sweden, Denmark, Canada, Slovenia, Iceland, Finland	Low: 6-9	Austria, Finland, Ireland, Switzerland

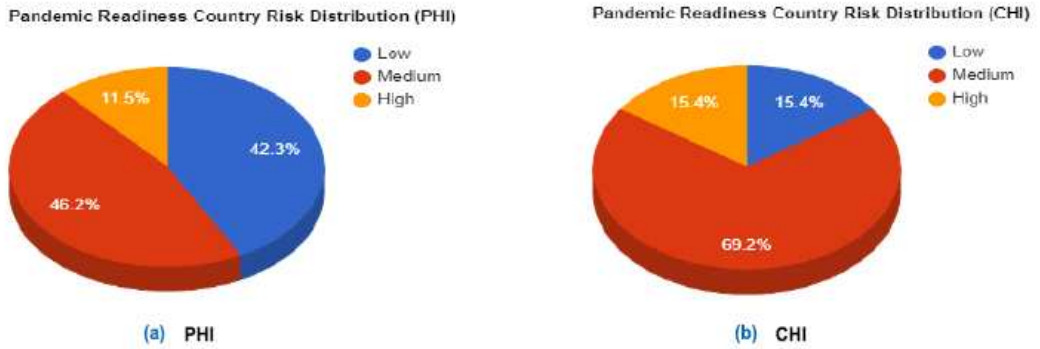


Figure 5.10: Distribution of countries based on total Population Health Index and Country Health Index Pandemic Risk Scoring Model

As shown in Figure 5.10, the majority of the 26 countries were assessed to have a medium pandemic risk (46.2%), while a smaller number of countries are classified in the high (11.5%) or low (42.3%) pandemic risk group under Population Health Index. The Country Health Index shows that 69.2% of countries have medium pandemic risk, while 15.4% of countries have low and high risk. The United States is classified as a medium risk country, for both indices, together with most of the countries in Europe.

Chapter 6

Research Part 2: Dataset

The analyses for the second part of this research used 16 variables. Table 6.1 presents the 14 variables that represent the actual values from the research dataset. Two additional variables, case fatality rate and vaccination rate, were derived. For each of the 26 countries in the dataset, the respective values in the total deaths column were divided by the total cases column to get the case fatality rate (CFR), an epidemiologic statistic defined as the proportion of deaths within an observed population of interest [34]. The number of individuals vaccinated (with at least one dosage) for each of the 26 countries was divided by the country's total population to determine the vaccination rate.

For a more meaningful interpretation of the data variables used to assess the correlation with the vaccination and CFR rates, data variables were organized into novel public health indices, the Population Health Index, PHI [107], and Pandemic Sensitivity Index, PSI (Table 6.1) [157]. The PHI contains the parameters that describe the health of the population such as: cardiovascular death rate, diabetes prevalence, female smokers, male smokers, life expectancy, age 65 and older, and median age. The PSI Index represents variables that are directly impacted by the pandemic, such as total COVID-19 cases and deaths, number of COVID-19 hospital and ICU admissions, Government response stringency index (a composite measure based on nine response indicators including school and workplace closures, and travel bans), reproduction rate of transmission of COVID-19, and positivity rate of COVID-19.

Table 6.1: Public health indices (PHI and PSI) definitions from the Our World in Data Dataset Metadata File [92]

Population Health Index (PHI)	Pandemic Sensitivity Index (PSI)
cardiovascular death rate: death rate from the cardiovascular disease in 2017 (annual number of deaths per 100,000 people)	***stringency_index***: Government response stringency index: composite measure based on 9 response indicators including school and workplace closures, and travel bans.
diabetes prevalence: Diabetes prevalence (% of population aged 20 to 79) in 2017	***positive_rate***: The share of COVID-19 tests that are positive given as a rolling 7-day average
female smokers: Share of women who smoke, most recent years available	***hosp_ patients***: Number of COVID-19 patients in hospital on a given day
male smokers: Share of men who smoke, most recent years available	***icu_patients***: Number of COVID-19 patients in intensive care unit (ICUs) on a given day
life_expectancy: Life expectancy at birth in 2019	***reproduction_ rate***: Real time estimate of the effective reproduction rate of COVID-19
aged 65 or older: Share of the population that is 65 years or older, most recent years available	***total_cases***: Total confirmed cases of COVID-19
median age: Median age of the population, UN projection for 2020	***total_deaths***: Total deaths attributed to COVID-19

This research was conducted to identify the vaccination inflection points and the time needed to reach the critical cumulative vaccination rate thresholds to observe continuous decrease of the case fatality rates. It was conducted both at an aggregate and at the country level. In order to account for fluctuations in the case fatality rate curves, the vaccination inflection points were measured at two distinct intervals. The first vaccination inflection timepoint, primary vaccination inflection point (PVIP) was assessed from the vaccination start date to the date of the first CFR drop post vaccination. From the immunization start date to the sharpest, most notable CFR drop after vaccination, the secondary vaccination inflection point (SVIP) was measured. It represents the timepoint when the cumulative vaccination rate reached a critical threshold showing a continuous decrease of the case fatality rate, signaling the turnaround in the pandemic. Table 6.2 provides an overview of descriptions of critical variables used in this research relative to the vaccination inflection point. COVID-19 historical data was utilized to develop models that can be used for future pandemics.

Table 6.2: Description of derived variables used for vaccination inflection point analyses.

variables	description
vaccination start date	first documented date when vaccination started at the country level
CFR at vaccination start	Case fatality rate at the time on the 1st day of vaccination
CFR + 14 days	case fatality rate at the time when initial immunity from vaccination should be developed
vaccination rate at CFR +14 days	vaccination rate at the time of initial immunity
Primary vaccination inflection point (PVIP)	date when the first case fatality rate reduction is observed post vaccination, measured on the day of the 1st CFR peak post vaccination + one day
CFR at PVIP	case fatality rate at PVIP, measured on the day of the 1st CFR peak post-vaccination + one day
vaccination rate at PVIP	vaccination rate at the PVIP, measured as the vaccination rate on the day of the 1st CFR peak post vaccination + one day
Secondary vaccination inflection point (SVIP)	date when the most significant CFR reduction is observed post vaccination, measured on the day of the CFR peak that is followed by the most significant and continuous CFR reduction post vaccination + one day
CFR at SVIP	case fatality rate at the SVIP, measured as the CFR rate on the day of CFR peak that is followed by the most significant CFR reduction post vaccination + one day
vaccination rate at SVIP	vaccination rate at the SVIP, measured as the vaccination rate on the day of the CFR peak that is followed by the most significant CFR reduction post vaccination + one day

In this research, it was assumed that all vaccines produced by different technologies and manufacturers have the same effectiveness. It was also assumed that distribution of different vaccines in different countries includes a combination of initial two-dose and single-dose vaccines and single dose booster vaccines over the two-year period (Dec 2020-Dec 2022). Since all vaccines require approximately two weeks to produce immunity, the effect of performance of vaccines on CFR was examined two weeks after the start of vaccination.

Several types of vaccines were available at the time of the initial vaccination: genetically engineered messenger RNA Pfizer/BioNTech and Moderna, viral vector vaccines (Janssen/Johnson & Johnson and University of Oxford/AstraZeneca, Sputnik V), protein subunit vaccine (Novavax, Sinovac). The initial vaccinations in 2020 were delivered, in most cases, in sets of 2-doses, with a 3-week period in between (Pfizer/BioNTech, Moderna, Sinovac, Sputnik V). Some initial vaccines were delivered as a single dose vaccine (J&J, AZ/Oxford). Consequently, booster doses were delivered as single dose vaccines, starting in the third quarter of 2021 (Sep 2021 in the US, Oct/Nov 2021 in the EU) and continuing in 2022 (approved boosters in Mar and Sep 2022 in the US) and 2023 (approved in Sep 2023 in US and EU), for a total of four booster doses [88]. Today there are approximately 40 COVID-19 vaccines that were approved by

regulatory agencies for full emergency use authorization. Of those 40, 16 have full authorization in only one country, 12 in ten or fewer countries, and 12 in more than 10 countries [108]. Emergence of new variants may be a challenge for the vaccines, reducing their protective power with the transmissibility of new variants substantially higher than the pre-existing SARS-CoV-2 variants. Booster dose vaccines were introduced to boost the protection power of vaccines and help the individuals with weakened immune systems. Efficacy of most vaccines range from 70-95%, mainly against symptomatic disease [109, 110]. All countries from this dataset (26 countries) are classified in three categories relative to their GDP per capita (>$50,000, $35,000-$50,000, and <$35,000) [92]. Table 6.3 summarizes the distribution of countries. This research was solely conducted by using publicly available data.

Table 6.3: Distribution of countries based on GDP per capita.

GDP per capita	Country distribution
> 50,000	Ireland, Luxembourg, Switzerland, United States
35,000-50,000	Austria, Belgium, Canada, Denmark, Finland, France, Iceland, Italy, Netherlands, Sweden, United Kingdom
< 35,000	Bulgaria, Cyprus, Czechia, Estonia, Latvia, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain

Chapter 7

Research Part 2: Methodologies

7.1 Data Preprocessing of Timeseries Dataset Temporally

Data utilized in this research was pre-processed by assigning the original time series dataset to training and testing datasets temporally. For each country, the training set included data from the beginning of the pandemic (March 1, 2020) until a few weeks post vaccination start. The testing set included the remaining data post vaccination until the end of the dataset (December 30, 2022). Data cleaning was conducted by resolving the problem of missing and duplicate values, resolving data inconsistencies, removing outliers, and smoothing variables used for forecasting (*vaccination_rate* and *case_fatality_rate*), including all exogeneous variables (*stringency_index, aged_65_older, life_expectancy, and positive_rate)*. Smoothing was conducted by using a window of seven days to remove all noisy data. The current day value was calculated using the mean of the previous seven days for each variable. In this type of dataset, it is common that some data is missing, both at random and not at random. For this research, it was important that the data on the total number of cases and deaths was complete since it was used to derive the case fatality rates. This missing data was resolved by taking the mean values of the total number of cases and deaths from the previous day and the next day. Other missing data was managed in a similar manner. Data quality assessments (completeness, reliability, consistency, validity, and no redundancy) were also completed. Exploratory Data Analysis was conducted by exploring graphs and visuals in order to observe trends over time of the vaccination and case fatality rates for each country.

7.2 Forecasting Methodologies

Three foundational forecasting methodologies were applied: Autoregressive Integrated Moving Average (ARIMA), Prophet, and Long-Short Term Memory (LSTM) models. These models were then enhanced and combined to develop novel double and triple hybrids, SARIMA-Bidirectional LSTM and SARIMA-Prophet-Bidirectional LSTM models. They were used to forecast the primary and secondary vaccination inflection points (PVIP and SVIP) relative to the case fatality rates, for each of the 26 countries. All machine learning and deep learning analyses were done using Python version 3.10.1 and the scikit-learn library version 1.2.0 [111]. In addition, the novel Vaccination Inflection Point Score was developed, and countries were classified according to the score.

7.21 Correlation Analysis

The correlation analysis was performed using Ordinary Least Squares Multifactor Regression Methodology to identify the top four variables that correlate the most with vaccination and case fatality rates for implementation into forecasting methodologies. These analyses were performed as an aggregate analysis of 14 variables that were assessed for correlation with vaccination and case fatality rates. All variables were used for the correlation assessment with the vaccination rate. Two variables, *total_cases* and *total_deaths* were not used in the assessment of the case fatality correlation since the CFR is a ratio of these two variables. In order to derive the list of the top four variables most correlated with both vaccination and case fatality rates together, the ranking order was assessed across both target variables (vaccination and case fatality rates).

7.22 Foundational Forecasting Methodologies

Baseline forecasting methodologies were selected based on literature search, model strengths and limitations.

7.221 Autoregressive Integrated Moving Average (ARIMA)

ARIMA (Autoregressive Integrated Moving Average) model is selected for its characteristics of being well-suited for forecasting time series data that exhibits trends and seasonality. It is deemed to be effective in forecasting a variety of real-world phenomena, which has good applicability for COVID-19, showing greater flexibility, accuracy, interpretability, and robustness. The following defines the ARIMA model's parameters: The number of lag observations included in the model is represented by p, the degree of differencing by d, the number of times raw observations are differencing, and the size of the moving average window by q, the order of the moving average [113].

7.222 Facebook Prophet

The Facebook Prophet algorithm is an open-source software developed by Facebook's core Data Science Team. If the time series data has strong seasonal effects, this model works the best. It is a regression model for forecasting, specifically designed to forecast time series data that exhibits trends, seasonality, and coverage for holidays. It is also fast and scalable, and similar to ARIMA, this model is interpretable, robust, flexible, and accurate [114].

7.223 Long Short-Term Memory (LSTM)

LSTM Model is a neural network model that can learn long-term dependencies in time series data, handle nonstationary and noisy data, as well as leverage additional features. It is also

accurate, flexible, and scalable [115, 155].

7.23 Novel Hybrid Forecasting Models

7.231 Double Hybrid Forecasting Model: SARIMA-Bidirectional LSTM

Review of published literature showcases the use of different forecast models and enhancements in COVID-19 research, demonstrating better accuracy and performance in forecasting by hybrid models. For example, ARIMA-LSTM hybrid model was used to predict future COVID-19 transmissions in China where ARIMA-LSTM model was paralleled by weight of regression coefficient performing better than ARIMA alone [117]; the same group also looked at COVID-19 prediction using data from Germany and Japan and utilized three enhanced hybrid models: PSO-LSTM-ARIMA, MLR-LSTM-ARIMA, and BPNN-LSTM-ARIMA. The research showed that BPNN-LSTM-ARIMA had the best prediction accuracy [118]. Priya and colleagues compared time series forecasting models utilizing ARIMA, Facebook Prophet, Holt-Winters Model, and Hybrid ARIMA-ANN (to take advantage of the unique characteristics of ARIMA and ANN models in linear and nonlinear modelling). The Hybrid model showed better accuracy and root mean square error [119]; Morais looked at forecasting daily Covid-19 cases with a hybrid ARIMA and neural network model to capture the linear and non-linear structures of daily Covid-19 cases (MLP-ARIMA) [120]; and Nawi researched a hybrid ARIMA-SVM model [121]. Borges looked at COVID-19 ICU demand forecasting utilizing Prophet-LSTM approach vs a stand-alone approach in Brazil, confirming better performance of the hybrid model [122], and Long researched an efficient forecasting tool for Monkeypox outbreak in the US using ARIMA, Prophet,

NeuralProphet, stacking model, and LSTM models. NeuralProphet achieved the optimal performance [123]. Furthermore, Guha used LSTM and a gated recurrent unit (GRU) in his study to offer two recurrent neural network-based methods for predicting the daily confirmed COVID-19 cases, daily total positive tests, and the total number of vaccinated individuals [97]; Shastri looked at time series forecasting of Covid-19 using deep learning models: the recurrent neural network based variants of long-short term memory (LSTM) such as Stacked LSTM, Bi-directional LSTM and Convolutional [124]; Devaray utilized ARIMA, LSTM, Stacked LSTM (SLSTM) and Prophet approaches [125]; Zhenyu Li researched convolutional neural network combined with the stacked long-short-term-memory network model (CNN-Stack BiLSTM) [126]. The Stacked LSTM (SLSTM) model was also researched by Maaliw [127] and Ali, who also use the bidirectional enhancement to create a stacked Bi-directional long short-term memory (Stacked Bi-LSTM) network that forecasts COVID-19 more accurately [128]. Sah compared different COVID-19 forecasting models, Prophet, ARIMA, LSTM, and stacked LSTM-GRU models demonstrating better prediction results with the hybrid stacked LSTM-GRU model [129]. Other researchers looked at the Ensemble Empirical Mode Decomposition and Deep Learning creating an EEMD-LSTM hybrid model [130] and EEMD method with the Autoregressive Integrated Moving Average Exogenous inputs (ARIMAX) method, which they called EEMD-ARIMAX [131].

Hybrid models for this research were selected based on the literature search, strengths, and limitations of the individual components for forecasting performance, available enhancements to address limitations, and for their specific complementary characteristics that land them well for hybrid application. SARIMA-Bidirectional LSTM hybrid model combines the strengths of two powerful forecasting techniques, ARIMA enhanced with a seasonality component (the S) in SARIMA and enhancing the LSTM model to analyze data in both directions (Bidirectional

component). This hybrid combines a linear and non-linear model, benefits from forecasting time series data that exhibits trends and seasonality and at the same time, an ability to learn long-term dependencies in time series data, as well as capture both forward and backward dependencies. SARIMA-Bidirectional LSTM complements the strength of each model and is expected to achieve better forecasting accuracy than either model individually [116].

7.232 Triple Hybrid Forecasting Model: SARIMA-Prophet-Bidirectional LSTM

With the addition of a Facebook Prophet forecasting model, which is especially made to forecast time series data that displays patterns, seasonality, and holidays, the triple hybrid SARIMA-Prophet-Bidirectional LSTM forecasting model improves upon the previously described hybrid model. The new triple hybrid combines the strengths of all three forecasting techniques with an ability to capture short-, medium-, and long-term dependencies, handle non-stationary and noisy data, and leverage additional features. Due to the complementary nature of the hybrid model components and a better fit for the data being researched, it would be expected that the new models would achieve better forecasting accuracy than either model individually.

7.3 Accuracy and Performance Assessment

Accuracy and performance assessment was conducted across all the models (foundational and hybrid models) evaluating vaccination and case fatality rates: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Entropy, relative to the actual data. In addition, the accuracy of the forecasting results of each model was compared with actual historical data from the Our World in Data dataset, specifically, to the actual time needed to reach

the vaccination inflection points for each country.

7.4 Anomaly and Volatility Analyses

Anomaly and Volatility analysis and assessments were conducted across all-time series analysis and forecasting models utilizing Isolation Forest and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) models, both well-studied in this field. These methodologies were selected based on the review of published literature that showcase their good performance as well as being valuable algorithms for anomaly and volatility detection in the context of COVID-19 vaccination forecasting [132]. The best performing model for predicting the time to COVID-19 vaccine inflection point for each country was chosen using the results of anomaly and volatility detection.

7.41 Isolation Forest

The last part of the research was focused on the assessment of anomaly and volatility detection analysis across the time series analysis models. These analyses were conducted to identify unusual or unexpected patterns in data, to prevent overfitting, improve the accuracy, performance, and reliability of machine learning models and complex systems. It is often used in Systems Engineering to detect unusual activity in system logs, performance bottlenecks in systems, and anomalous patterns in system data and to improve overall reliability, efficiency, and security of complex systems. The first algorithm used in this research is Isolation Forest.

Isolation Forest can detect anomalies in an unsupervised manner. This model is used to compare the accuracy of different forecasting models and considered to be efficient, scalable, and

robust to outliers. It works by randomly selecting features and splitting values to create partitions of the data. This process is repeated until isolation of the anomalies. It is particularly well-suited for high-dimensional data, which is the case with COVID-19 vaccination data, which includes features such as vaccination rate, case fatality rate, population density, and socio-economic factors. It is also relatively insensitive to outliers, which can be a problem for other anomaly detection algorithms. Isolation Forest can be used to detect anomalies in the vaccination and case fatality rates. This can be useful for identifying periods where the vaccination and CFR rate are significantly higher or lower than expected, adjusting, or improving the forecasts for the vaccination inflection point [133].

Isolation Forest measured three parameters: Precision, Recall, and F1-score. Precision measures the proportion of detected anomalies that are actually true anomalies, where high precision (closer to 1) is very accurate in its anomaly detections, with few false positives. A good threshold for Isolation Forest is 0.7 or higher. Recall measures the proportion of true anomalies that are correctly identified by the model, high recall (closer to 1) means the model is sensitive and can capture most anomalies. A good threshold for Isolation Forest is 0.7 or higher. F1-score combines precision and recall into a single metric, balancing their trade-off. A high F1-score (closer to 1) indicates a good balance between precision and recall, suggesting a reliable anomaly detector. Isolation Forest results at 0.7 or higher for all parameters are considered to be good results [133].

7.42 Generalized Autoregressive Conditional Heteroskedasticity (GARCH)

The second algorithm used to compare the accuracy of different forecasting models is the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model. The GARCH model is a powerful tool employed to capture and model volatility patterns in the residuals. This model considers the conditional variance and accounts for the time-varying volatility and is especially well suited for time-series analysis, which is the case with COVID-19 vaccination and case fatality rate data.

The GARCH model was used to forecast the volatility of the COVID-19 vaccination and CFR rate. This helped to identify periods where the vaccination and CFR rates are likely to increase or decrease more rapidly than expected. If the model detects anomalies, this could indicate that the vaccination and CFR rates are not following the expected patterns [134].

For anomaly and volatility detection, respectively, in the context of COVID-19 vaccine inflection point forecasting, isolation forest and GARCH models are suitable. They are both efficient, important for anomaly and volatility detection in large datasets, and robust to outliers. This can be a problem in COVID-19 vaccination data due to factors such as data entry errors and reporting delays. These models are also flexible, due to ease of adaptation to a variety of different anomaly detection tasks. The GARCH model also has several limitations, such as sensitivity to the choice of parameters, less robust performance for very short time series datasets, and the inability to capture all types of anomalies.

The GARCH model measures three parameters: Volatility Persistence, Relative Importance of ARCH Term, and Relative Importance of GARCH Term. Volatility Persistence represents the degree to which shocks to volatility persist over time, with an acceptable range

between 0.7 and 1. Values below 1 are considered acceptable, ensuring stationarity of the volatility process. However, values closer to or exceeding 1, indicate stronger persistence, meaning shocks have longer-lasting impacts on volatility and might suggest issues like integrated volatility or model misspecification. The range that is typical and acceptable for Relative Importance of ARCH Term is 0 to 0.4. Relative Importance of GARCH Term captures the persistence of volatility shocks over time with an acceptable range of 0.3 to 0.9 [134].

7.5 Vaccination Inflection Point Score

Vaccination Inflection Point score was developed to categorize countries based on their actual time to achieving secondary vaccination inflection point, representing the time of the most significant CFR reduction post vaccination, and therefore, identifying the critical threshold signaling the turnaround in the pandemic. Countries were categorized into three groups corresponding to scores 1, 2, and 3, with a score of 1 indicating that the country needing the shortest amount of time to reach their secondary vaccination inflection point. This tool can help with the interpretation of changes in the pandemic dynamic, serve as a learning tool for the importance of the contribution of vaccination to achieving faster herd immunity, and improving the overall pandemic risk of countries.

Chapter 8

Research Part 2: Results

8.1 Correlation Analysis Results

The correlation analysis was performed using Ordinary Least Squares Multifactor Regression Methodology. These analyses were performed as aggregate analysis with 14 variables. The correlation was assessed first with the vaccination rate as the target variable, followed by the case fatality rate. The top four variables most correlated with vaccination rate were: *stringency_index, life_expectancy, positive_rate, and total_deaths*. The top four variables for the case fatality rate were: *stringency_index, aged_65_older, life_expectancy, and positive_rate*. Using the ranking order of factors across both vaccination and case fatality rates, the top four variables that are most connected with both vaccination and case fatality rates together were determined. The final ranking order of the four variables was: *stringency_index, aged_65_older, life_expectancy, and positive_rate,* representing the exogeneous variables that were used in the primary and secondary vaccination inflection point forecasting analyses. The stringency index and positive rate were variables representing the PSI index and aged 65 and older and life expectancy represented the PHI index.

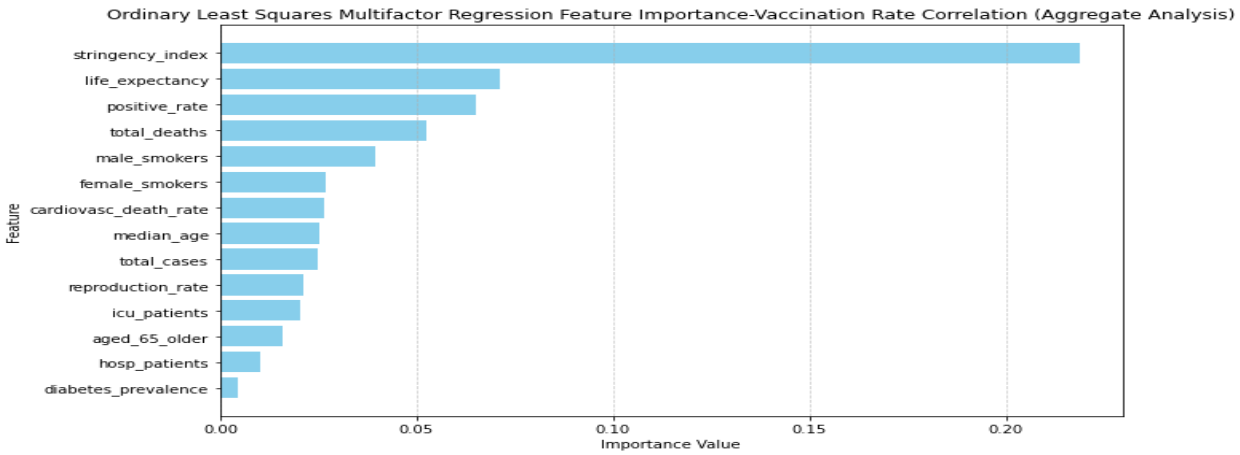


Figure 8.1: Correlation analysis for vaccination rate (aggregate analysis)

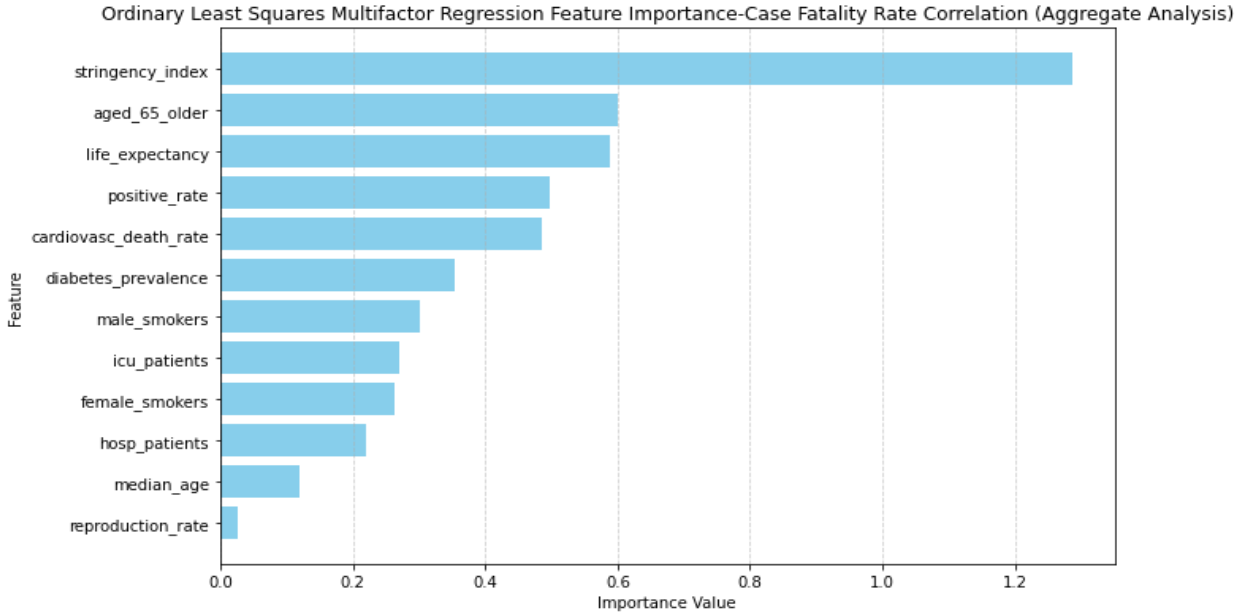


Figure 8.2: Correlation analysis for case fatality rate (aggregate analysis)

8.2 Forecasting Analysis Results

The summary of the conducted analyses is presented in Table 8.1, first as an aggregate and then per GDP per capita category (>$50,000, $35,000-$50,000, and <$35,000). Overall, all countries started their vaccination campaigns within the 43 days, starting with Latvia on December 4, 2020, and ending with the UK starting on January 10, 2021. Countries with the higher GDP initiated their vaccination efforts faster than the other countries (15 days vs 26 and 35 days), however, countries with the mid-range GDPs reached the PVIP and SVIP faster than the other two groups, with high and low GDP per capita. In contrast to 76 and 131.7 days, PVIP was reached in 37.5 days, and SVIP in 299.2 days as opposed to 336.5 and 380.4 days.

When median numbers were employed, similar results were seen, with the mid-range GDP countries once again outperforming the high- and low-range GDP countries. They also had the lowest achieved vaccination rate (74.7% vs. 70.6% and 63.3%) and the shortest time required to reach both PVIP (34 days vs. 80.5 and 82 days) and SVIP (316 days vs. 343.5 and 365 days). Overall, all countries reached an average vaccination rate of 67.8% (mean) and 71.25% (median) at the time they observed the significant CFR drop post-vaccination (SVIP). The highest vaccination rate was achieved in Portugal (89%) and the lowest in Bulgaria (28%).

Table 8.1: Summary of results across 26 countries

	Aggregate data for 26 countries	Countries with GDP per capita >$50,000	Countries with GDP per capita $35,000-$50,000	Countries with GDP per capita < $35,000
		4 countries (15.4%)	11 countries (42.3%)	11 countries (42.3%)
vaccination start	43 days (Dec 8, 2020 - Jan 10, 2021	15 days (Dec 13 - Dec 28, 2020)	26 days (Dec 8, 2020 - Jan 3, 2021)	35 days (Dec 4, 2020 - Jan 8, 2021)
	mean (range)			
time to reach PVIP*	83.27 days (15-367)	76 days (49-94)	37.5 days (15-75)	131.7 days (16-367)
vaccination rate at PVIP	13.1% (0.1-50)	9.7% (0.9-24.8)	5% (0.4-33.1)	18.6% (1.5-50.1)
time to reach SVIP**	339.31 days (161-560)	336.5 days (296-363)	299.2 days (161-371)	380.4 days (319-560)
vaccination rate at SVIP	67.8% (28-89)	71% (66.4-76.5)	74.2% (63.8-81.8)	60.3% (28-89.1)
	median (range)			
time to reach PVIP	57.5 days (15-367)	80.5 days (49-94)	34 days (15-75)	82 days (16-367)
vaccination rate at PVIP	6.05% (0.1-50)	6.6% (0.9-24.8)	2.4% (0.4-33.1)	9.5% (1.5-50.1)
time to reach SVIP	355.5 days (161-560)	343.5 days (296-363)	316 days (161-371)	365 days (319-560)
vaccination rate at SVIP	71.25% (28-89)	70.6% (66.4-76.5)	74.7% (63.8-81.8)	63.3% (28-89.1)

*PVIP: Primary vaccination inflection point
**SVIP: Secondary vaccination inflection point

Analysis of vaccinations by age group in Our World in Data (except for three countries) showed similar distribution by age [92]. The elderly population (60-70, 70-80, and 80+ years of age) achieved the highest vaccination rates in all, but three countries (Latvia, Romania, and Bulgaria), followed by the middle age group (18-24, 25-59). The smallest vaccination rates were observed in the youngest age group (0-17). The data for the US and UK were not available in the Our World in Data dataset, however, data from official government sites demonstrated the same patterns observed with the rest of the countries [153, 154], supplement Tables S10, S11, and S12. There were no official records available for Serbia at the time of this research. This confirms earlier statements that most countries prioritize elderly and frail population in their vaccination campaigns. Looking at the countries based on their GDP per capita grouping, the mid-range group on average achieved higher vaccination rates of the elderly population than the countries with higher and lower GDP per capita. These findings support the better performance of the countries in the mid-range GDP group, demonstrating the importance of prioritizing the needs of the elderly population (age 65 and older and life expectancy) in a pandemic setting. It should be assumed that other factors, such as acceptance and robustness of the vaccination campaign and vaccination mandates imposed by governments played a significant role as well [143].

Table 8.2 presents the ranking order of the countries based on the time to reach SVIP. The UK was the first country to observe the SVIP, taking only 161 days (CFR 3.3%, vaccination rate 63.8%) to reach the same point as Romania, which took 560 days (CFR 2.2%, vaccination rate 41.6%).

Supplemental Tables (Table S1, S2A-B, S3A-B) present all results of all forecasting models, the three foundational (ARIMA, PROPHET, LSTM) and the two hybrid forecasting models (double hybrid: SARIMA-Bidirectional LSTM, and triple hybrid: SARIMA-Prophet-

Bidirectional LSTM). The baseline data for each country, as well as the actual historical data from the COVID-19 pandemic are also documented in these supplemental tables.

Table 8.2: Ranking of the Countries based on the time to reach SVIP.

Rank	Country	Time (days) to reach SVIP	Vaccination start date	Date SVIP reached	Vaccination rate at SVIP
1	United Kingdom	161 days	Jan 10 2021	Jun 20 2021	63.88%
2	Iceland	201 days	Dec 30 2020	Jul 19 2021	71.64%
3	Denmark	274 days	Dec 8 2020	Sep 8 2021	73.99%
4	Belgium	292 days	Dec 28 2020	Oct 16 2021	74.77%
5	Netherlands	293 days	Jan 8 2021	Oct 28 2021	69.99%
6	Ireland	296 days	Dec 28 2020	Oct 20 2021	76.58%
7	Italy	316 days	Dec 27 2020	Nov 8 2021	79.32%
8	Portugal	319 days	Jan 1 2021	Nov 16 2021	89.10%
9	France	323 days	Dec 27 2020	Nov 15 2021	76.88%
10	Switzerland	329 days	Dec 21 2020	Nov 15 2021	66.42%
11	Finland	333 days	Jan 3 2021	Dec 2 2021	77.16%
12	Spain	337 days	Jan 4 2021	Dec 7 2021	80.84%
13	Cyprus	353 days	Jan 6 2021	Dec 25 2021	71.53%
14	Sweden	358 days	Jan 3 2021	Dec 27 2021	72.39%
15	Luxembourg	358 days	Dec 28 2020	Dec 21 2021	70.99%
16	Serbia	361 days	Jan 8 2021	Jan 4 2022	48.20%
17	Estonia	362 days	Dec 27 2020	Dec 24 2021	63.29%
18	United States	363 days	Dec 13 2020	Dec 11 2021	70.30%
19	Bulgaria	365 days	Dec 29 2020	Dec 29 2021	28.08%
20	Canada	370 days	Dec 14 2020	Dec 19 2021	81.80%
21	Austria	371 days	Dec 27 2020	Jan 2 2022	75.10%
22	Czechia	372 days	Dec 27 2020	Jan 3 2022	65.12%
23	Slovenia	374 days	Dec 27 2020	Jan 5 2022	59.07%
24	Slovakia	386 days	Jan 3 2021	Jan 24 2022	45.73%
25	Latvia	395 days	Dec 4 2020	Jan 3 2022	71.06%
26	Romania	560 days	Dec 27 2020	Jul 10 2022	41.64%

In the dataset used for this research, 65% of countries started their vaccination efforts in December 2020, and 35% started in January 2021. The primary vaccination inflection point representing the first observed reduction in the CFR post vaccination was reached at 83.27 days (mean, range 15-367 days), with 42% of countries seeing the initial impact in less than 50 days, 38.4% in 50-100 days, and 19.2% above 100 days (Figure 8.3). This reduction was achieved with the initial vaccination rate of 31.1% (mean, range 0.1% to 50%), with 27% of countries reaching the vaccination rate of >25%, 15.3% reaching the rate between 11-25%, and 57.7% reaching the rate of <10% (Figure 8.4). Romania had the longest wait to first reduction at 367 days (CFR 3.2%,

vaccination rate 27.8%), whereas Finland observed the fastest PVIP in just 15 days (CFR 1.6%, vaccination rate 1.1%).

The secondary vaccination inflection point (SVIP), representing the most significant reduction in CFR post vaccination, signaling the start of the continuous CFR reduction and turnaround in the pandemic, was reached at 339.31 days (mean, range 161-560 days), with 23.1% of countries observing this impact in less than 300 days, 53.8% from 300-370 days, and 23.1% in more than 370 days (Figure 8.5). This reduction was achieved with the cumulative vaccination rate of 67.8% (mean, range 8%-89%), with 50% of countries reaching the vaccination rate between 50-75% (Figure 8.6). Most of the countries reached a significant drop in the CFR in 2021 (73%), out of which 61.5% reached it in the 4th quarter of 2021, 11.5% in the 3rd quarter of 2021, and 27% in early 2022. The highest vaccination rate at this inflection point was achieved in Portugal (89%) on November 16, 2021.

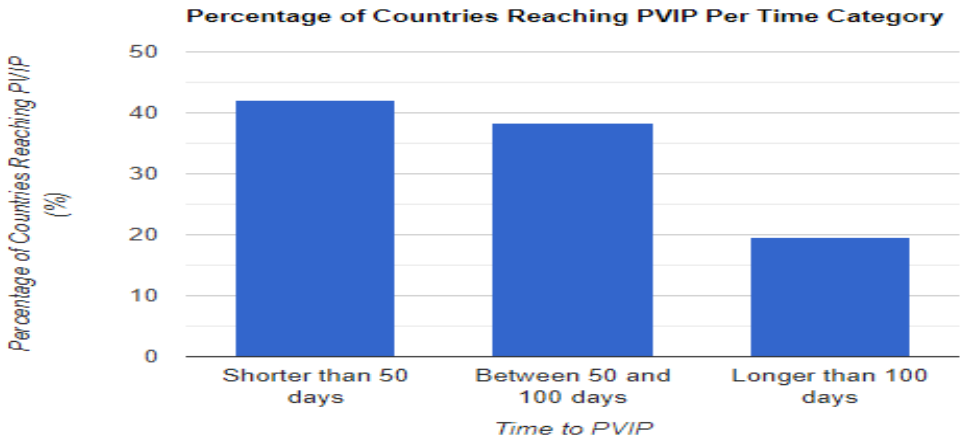


Figure 8.3: Percentage of Countries Reaching PVIP Per Time Category

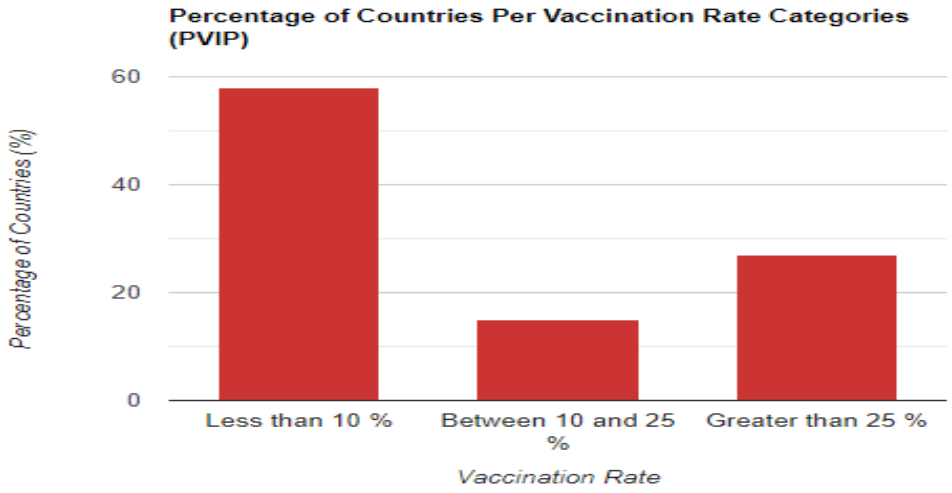


Figure 8.4: Percentage of Countries Per Vaccination Rate Categories (PVIP)

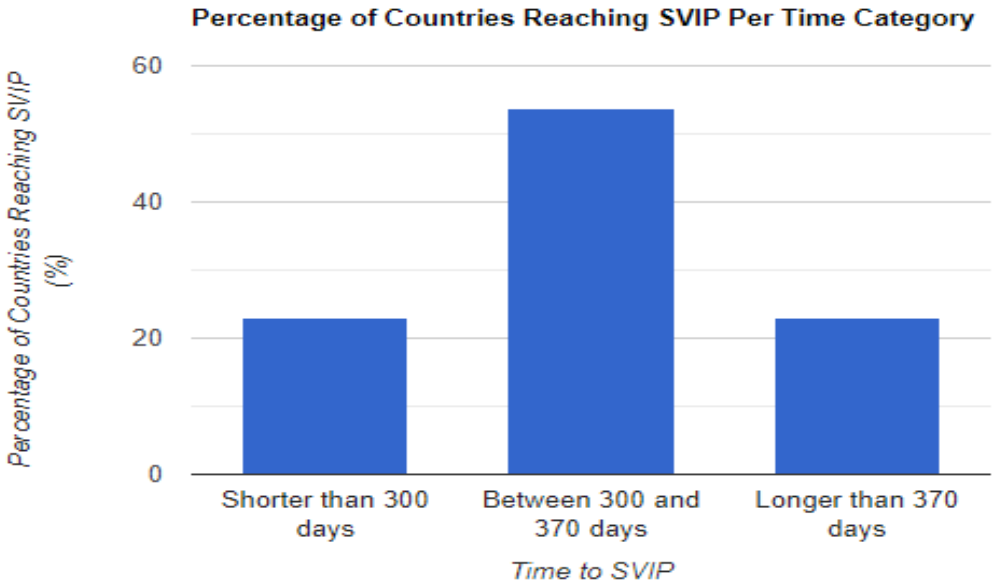


Figure 8.5: Percentage of Countries Reaching SVIP Per Time Category

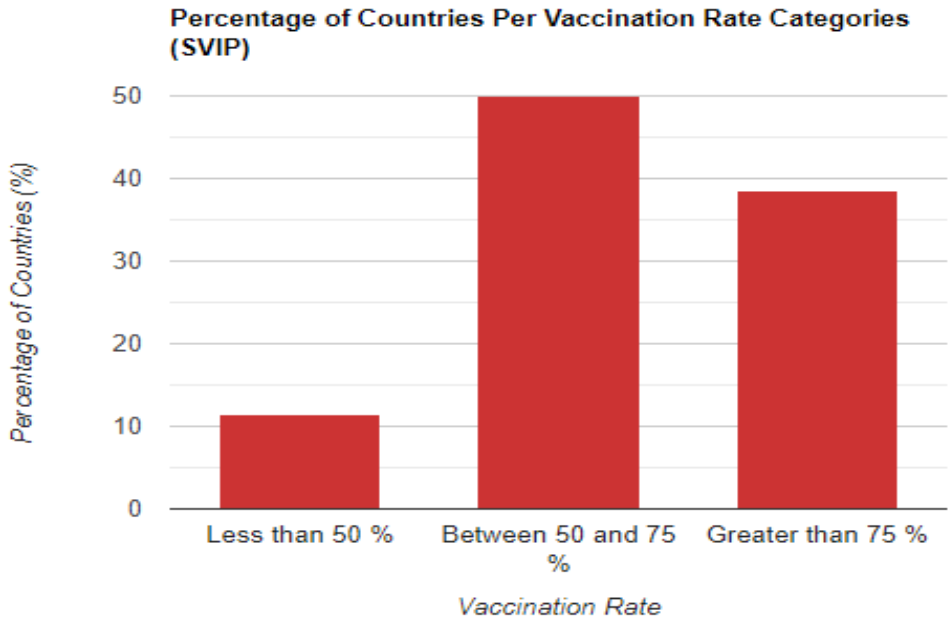


Figure 8.6: Percentage of Countries Per Vaccination Rate Categories (SVIP)

Overall, at the time of the SVIP, all countries with the exception of three, showed a reduction in the CFRs relative to the CFRs at the beginning of the vaccination. The highest CFR at the time of the SVIP was in Bulgaria (4.1%), followed by the UK (3.32%), and Italy (2.75%). The CFRs of the remaining countries were less than 2.0%, with the exception of Belgium and Romania. The countries with the lowest CFRs on record were Iceland (0.45%) and Cyprus (0.42%). Bulgaria, Latvia and Slovakia had the CFRs at the SVIP that were higher than the CFR at the vaccination start date, however, all three countries showed a reduction in the CFRs from the PVIP to the SVIP, indicating a positive impact of the vaccination.

8.3 Accuracy and Performance Assessment

When assessing vaccination and case fatality rates, accuracy and performance evaluations were carried out for both fundamental and hybrid models. The metrics that were compared to the real data were Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Entropy. Tables 8.3 and 8.4 showcase the mean and median results for all calculated metrics indicating the superior performance of the triple hybrid model SARIMA-Prophet-Bidirectional LSTM.

8.4 Anomaly and Volatility Analysis Results

All time-series analysis and forecasting models were subjected to anomaly and volatility assessments using the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) and Isolation Forest models. In the Isolation Forest model, precision, recall, and F1-score values above 0.7 indicate good performance. As presented in Tables 8.5 and 8.6, both mean and median values were above 0.7, indicating that all forecasting methodologies are performing well and accurately, validating performance of all forecasting models. In the GARCH model, Volatility Performance between 0.7-1, Relative Importance ARCH Term between 0-0.4, and Relative Importance of GARCH Term between 0.3 - 0.9, indicate good performance. Tables 8.7 and 8.8 presented that both mean and median values are within typical and acceptable ranges for all three indicators, suggesting that all forecasting methodologies are performing well and accurately, validating performance of all forecasting models.

Table 8.3: Vaccination rate forecasting metrics.

Metric	Country	ARIMA	Prophet	LSTM	SARIMA-Bidirectional LSTM Double Hybrid	SARIMA-Prophet-Bidirectional LSTM Triple Hybrid
Mean Absolute Error (MAE)	Mean:	0.273440346	0.269496	0.246185113	0.197624417	0.04688918
	Median:	0.2801870395	0.264832	0.190341017	0.091510384	0.008675794
Mean Squared Error (MSE)	Mean:	0.147081	0.185847	0.184565242	0.12681021	0.022887863
	Median:	0.104444003	0.161433	0.076258508	0.0723305005	0.0000841
Root Mean Squared Error (RMSE)	Mean:	0.333789454	0.321616	0.298212928	0.210208282	0.053976834
	Median:	0.327567389	0.322921	0.31088121	0.187344609	0.009161993
Entropy	Mean:	0.197672158	0.172476	0.093079818	0.111602614	0.10330825
	Median:	0.157320015	0.181081	0.02683001	0.0245333975	0.02033074

Table 8.4: Case fatality rate forecasting metrics.

Metric	Country	ARIMA	Prophet	LSTM	SARIMA-Bidirectional LSTM Double Hybrid	SARIMA-Prophet-Bidirectional LSTM Triple Hybrid
Mean Absolute Error (MAE)	Mean:	0.423921554	0.240282	0.243192654	0.240685426	0.059632661461538
	Median:	0.271977023	0.211509	0.206126957	0.163648243	0.033323647
Mean Squared Error (MSE)	Mean:	0.430524623	0.225626	0.165527034	0.147206818	0.008526044
	Median:	0.106648124	0.213829	0.0766076	0.062206746	0.001274246
Root Mean Squared Error (RMSE)	Mean:	0.500104124	0.275272	0.271744391	0.303519309	0.063672814
	Median:	0.326569872	0.273384	0.2643577545	0.258133644	0.037807365
Entropy	Mean:	0.199711931	0.19532	0.083534021	0.042455902	0.034290933
	Median:	0.217112239	0.193974	0.0327498355	0.009875701	0.0095256035

Table 8.5: Isolation Forest: Anomaly Detection for vaccination rate.

Isolation Forest-Anomaly Detection Results (Vaccination Rate Forecasting)			
Country	**Precision**	**Recall**	**F1 Score**
United States	0.957	0.739	0.8007
Austria	0.9117	0.8159	0.9772
Serbia	0.7469	0.9786	0.8216
Canada	0.9969	0.8121	0.7191
Belgium	0.7331	0.9941	0.8332
Bulgaria	0.7079	0.8811	0.7067
Czechia	0.9397	0.8437	0.7551
Denmark	0.9514	0.8081	0.7378
Estonia	0.7033	0.9454	0.9022
Finland	0.8061	0.8942	0.7076
France	0.7569	0.9209	0.7978
Iceland	0.7113	0.7931	0.7926
Ireland	0.8135	0.8411	0.7034
Italy	0.7614	0.8705	0.8592
Latvia	0.7289	0.8352	0.786
Luxembourg	0.7961	0.7753	0.8538
Netherlands	0.7212	0.7771	0.9907
Portugal	0.8517	0.9156	0.8336
Romania	0.9704	0.7877	0.7137
Slovakia	0.7616	0.904	0.9632
Slovenia	0.8737	0.7902	0.7825
Spain	0.8443	0.868	0.9747
Sweden	0.8305	0.7749	0.973
Switzerland	0.9417	0.7178	0.7002
United Kingdom	0.915	0.8303	0.8497
Cyprus	0.8723	0.9244	0.7774
Mean:	**0.8309**	**0.8476**	**0.8197**
Median:	**0.822**	**0.8382**	**0.7993**

Table 8.6: Isolation Forest: Anomaly Detection for case fatality rate.

Isolation Forest-Anomaly Detection Results (Case Fatality Rate Forecasting)			
Country	**Precision**	**Recall**	**F1 Score**
United States	0.8875	0.7503	0.857
Austria	0.926	0.7301	0.8791
Serbia	0.9929	0.8165	0.8089
Canada	0.797	0.9341	0.7126
Belgium	0.9709	0.7439	0.8172
Bulgaria	0.7817	0.9051	0.9914
Czechia	0.9896	0.9788	0.7884
Denmark	0.8408	0.9157	0.7323
Estonia	0.8036	0.8121	0.8191
Finland	0.9038	0.8677	0.8376
France	0.8538	0.9745	0.7094
Iceland	0.9427	0.9422	0.9627
Ireland	0.8745	0.7822	0.769
Italy	0.9391	0.907	0.9095
Latvia	0.9576	0.9067	0.7518
Luxembourg	0.8179	0.7254	0.996
Netherlands	0.747	0.8287	0.8018
Portugal	0.7876	0.918	0.8372
Romania	0.8899	0.752	0.7196
Slovakia	0.9771	0.9905	0.9455
Slovenia	0.921	0.9264	0.7131
Spain	0.7366	0.9903	0.7089
Sweden	0.7791	0.9988	0.8754
Switzerland	0.8953	0.7339	0.7227
United Kingdom	0.7583	0.7996	0.8798
Cyprus	0.9981	0.7523	0.7227
Mean:	**0.8757**	**0.8609**	**0.818**
Median:	**0.8887**	**0.8864**	**0.8131**

Table 8.7: GARCH: Volatility Detection for vaccination rate.

GARCH-Volatility Detection Results (Vaccination Rate Forecasting)			
Country	**Volatility Persistence**	**Relative Importance of ARCH Term**	**Relative Importance of GARCH Term**
United States	0.8282	0.2504	0.852
Austria	0.7063	0.1871	0.6404
Serbia	0.8652	0.3208	0.8104
Canada	0.8244	0.026	0.8845
Belgium	0.714	0.0902	0.4021
Bulgaria	0.7451	0.1982	0.3019
Czechia	0.8241	0.1447	0.5649
Denmark	0.7312	0.0737	0.3504
Estonia	0.8505	0.069	0.6422
Finland	0.7691	0.2895	0.4325
France	0.7524	0.1025	0.4491
Iceland	0.8171	0.063	0.63
Ireland	0.7553	0.0197	0.8928
Italy	0.7087	0.2291	0.3192
Latvia	0.7707	0.1126	0.7709
Luxembourg	0.7432	0.1899	0.4111
Netherlands	0.7774	0.0394	0.5235
Portugal	0.75	0.0377	0.7828
Romania	0.8011	0.0934	0.3291
Slovakia	0.7535	0.0961	0.6822
Slovenia	0.7218	0.2805	0.3256
Spain	0.7524	0.3012	0.4183
Sweden	0.7409	0.1956	0.8429
Switzerland	0.7716	0.1138	0.4444
United Kingdom	0.8017	0.0097	0.7839
Cyprus	0.7527	0.3958	0.6299
Mean:	**0.7703**	**0.1511**	**0.5814**
Median:	**0.7544**	**0.1132**	**0.5974**

Table 8.8: GARCH: Volatility Detection for case fatality rate.

GARCH-Volatility Detection Results (Case Fatality Rate Forecasting)			
Country	**Volatility Persistence**	**Relative Importance of ARCH Term**	**Relative Importance of GARCH Term**
United States	0.7459	0.055	0.5546
Austria	0.7973	0.295	0.6177
Serbia	0.7003	0.1989	0.4148
Canada	0.7955	0.0907	0.5704
Belgium	0.837	0.1485	0.5992
Bulgaria	0.7926	0.1119	0.4248
Czechia	0.8525	0.2854	0.815
Denmark	0.8412	0.0142	0.5472
Estonia	0.7639	0.2216	0.5843
Finland	0.7499	0.1335	0.3002
France	0.8511	0.2473	0.8511
Iceland	0.7068	0.2207	0.3123
Ireland	0.7093	0.3246	0.6198
Italy	0.8087	0.0608	0.8999
Latvia	0.7109	0.0828	0.3823
Luxembourg	0.8555	0.3423	0.6214
Netherlands	0.8424	0.1738	0.3842
Portugal	0.809	0.216	0.7016
Romania	0.7274	0.1294	0.3674
Slovakia	0.8668	0.1789	0.4596
Slovenia	0.846	0.3062	0.7968
Spain	0.8044	0.3552	0.3832
Sweden	0.7273	0.3213	0.602
Switzerland	0.7126	0.3562	0.6668
United Kingdom	0.8057	0.3061	0.5669
Cyprus	0.7119	0.356	0.3568
Mean:	**0.7835**	**0.2128**	**0.5539**
Median:	**0.7964**	**0.2184**	**0.5687**

8.5 Vaccination Inflection Point Score Results

Vaccination Inflection Point score was developed to categorize countries based on their actual time to achieving secondary vaccination inflection point, representing the time of the most significant CFR reduction post vaccination. Countries were categorized into three groups with scores 1, 2, and 3, with a score of 1 indicating the country needing the shortest amount of time to reach their secondary vaccination inflection point.

Table 8.9 and Figure 8.7 present the distribution of countries per VIP score. This data indicates that the majority of countries (53.8%) reached the SVIP between 300-370 days (score 2). The median results indicate that the countries with the shortest time to SVIP, score 1 (72.75%), have numerically greater vaccination rates than score 2 (71.75%) and score 3 (62%), notwithstanding the wide variation in the achieved vaccination rates among the various countries.

Table 8.9: Distribution of countries per SVIP score

SVIP score	days to SVIP	distribution of countries	% of countries	vaccination rate range
1	< 300	Denmark, Belgium, Iceland, Ireland, Netherlands, UK (6)	23.10%	63.8-76%. median 72.75%
2	300-370	US, Serbia, Canada, Bulgaria, Estonia, Finland, France, Italy, Luxemburg, Portugal, Spain, Sweden, Switzerland, Cyprus (14)	53.80%	28-89%. median 71.75%
3	>370	Austria, Czechia, Latvia, Romania, Slovakia, Slovenia (6)	23.10%	41.6-75.1%. median 62%

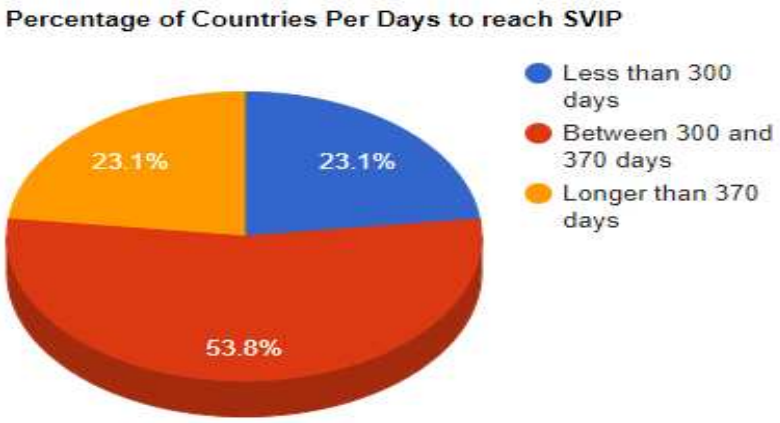


Figure 8.7: Percentage of Countries Per Days to reach SVIP.

Chapter 9

Discussion, Conclusions, and Limitations

9.1 Discussion

Since the start of the COVID-19 pandemic in 2020, many scientists at academic, clinical and research sites conducted significant research to enhance the understanding and improve the status of the pandemic. The devastating effect of the COVID-19 pandemic increased the urgency to develop testing models and standardized tools that are more sophisticated but also simpler and can increase the implementation and usability at the country level. These tools and models that use the pre-pandemic parameters, which are commonly collected and documented for every country, can be utilized to conduct pandemic risk assessments earlier. These assessments can serve as proactive indicators stimulating active discussion and development of pandemic readiness strategies by country public health officials, policy makers, and disaster management agencies. It is also equally important to assess, once the pandemic starts, how to predict and then influence the pandemic outcomes. Vaccination is one substantial factor that can significantly influence the start, strengths and the duration of a pandemic.

The first part of this research describes the application of two machine learning methodologies, Random Forest and Extreme Gradient Boost Regressor enhanced with the distribution lag model, and a novel machine learning approach using K-means-Coefficient of Variance sensitivity analyses, validated by Ordinary Least Squares Multifactor Regression model, RFR, XGBoost, and the Calinski-Harabasz method. These analyses were done to rank demographic, health, and economic parameters (non-pandemic predictive features) for 26 countries relative to their importance and correlation with the COVID-19 case fatality rates. These

non-pandemic features were grouped into two novel public health indices, Population Health Index and Country Health Index. This allowed for a more meaningful interpretation of the results in the appropriate context of public health. The grouping also provided the basis for the novel approach of K-means clustering for COV sensitivity analysis. In addition, it created the foundation for the novel Pandemic Risk Scoring model, classifying countries into low, medium, or high pandemic vulnerability risk categories.

The most appropriate models for feature importance assessment were found to be the RFR and XGBoost models, which are widely used by academics [9, 20, 22, 23, 27, 35]. The K-means-Coefficient of Variance sensitivity analysis was developed as a more sensitive novel machine learning approach, and the Ordinary Least Squares Multifactor Regression methodology was introduced as a validation model. All four methodologies were applied in a similar approach, first looking at the ranking order of all 13 predictive features at the aggregate level, followed by more complex analyses, single and paired country analyses, reporting the ranking order of features per public health indices, PHI and CHI. K-means clustering methodology was utilized with K-means-COV sensitivity analysis methodology. Calinski-Harabasz methodology was another clustering methodology that was applied in order to validate the results obtained by RFR and XGBoost models, as well as, the novel K-means-COV sensitivity analysis methodology. RFR and XGBoost were compared with performance metrics (best 10-fold cross validation score, mean squared error, R^2 score, root mean squared error, and entropy). The median value for each of these metrics, selected to minimize the impact of outliers, was calculated for each model and each index (PHI and CHI) and compared to the corresponding values, RFR PHI to XGBoost PHI, RFR CHI to XGBoost CHI. The comparison of RFR and XGBoost analyses confirmed that the XGBoost

methodology has a higher sensitivity, with the distribution of feature importance values being wider across all the features, and a higher accuracy across all performance metrics.

The K-means-Coefficient of Variance sensitivity analysis was developed, assessed, and validated with OLS MFR methodology. Additional validation was conducted with RFR, XGBoost and the Calinski-Harabasz methodologies showing similar results in the final ranking order of predictive features. The novel approach of K-means-COV sensitivity methodology can bring an additional value to the field of Systems Engineering since the sensitivity analysis provides a deeper understanding of the relationships between input and output variables. K-means-COV sensitivity analysis tested the robustness and validated the results, feature importance and rankings, of RFR and XGBoost Regressor models. The important relationships between model inputs (all data input features) and the target variable (case fatality rate), both on the aggregate level and per index (PHI and CHI), led to the development of better and more robust prediction models.

The study showed that XGBoost Regressor outperformed RFR in terms of accuracy and performance. The XGBoost single country analysis identified *cardiovascular death rate*, *life expectancy*, and *diabetes prevalence* as the top three predictive features in the Population Health Index, while the number of *hospital beds per thousand*, *total population*, and *human development index* were the most common predictive features with the highest correlation with the COVID-19 case fatality rate, in the Country Health Index. The most dominant predictive feature across all counties (46%) was *cardiovascular death rate* for the PHI, and *hospital beds per thousand* people (46%) for CHI. In addition, single versus paired analysis showed similarities in the ranking order of predictive features for both PHI and CHI.

These results align with the expectations that any chronic disease will be worsening during a pandemic, with additional stress, possible higher exposure to different infectious diseases due to

close proximity of other people, challenging access to health care providers, and possible limited supplies of medications. It is well-known that the *cardiovascular death rate* is a predictor of health status at a country level. In addition, cardiovascular diseases remain the leading cause of death worldwide [37], [38], with higher rates indicating countries with a more vulnerable population with underlying chronic conditions. Similarly, it is expected to see the same vulnerability applying to countries with a higher diabetes prevalence [41].

Most developed countries usually have a higher life expectancy, with a higher percentage of their populations being older, frail, with more comorbidities and chronic conditions, and therefore more susceptible to acute infections. These conditions would be expected to be exacerbated during a pandemic. For the Country Health Index predictive features, lower values in the number of *hospital beds per one thousand* people indicate a lower pandemic readiness level overall, since the ability of countries to compensate for an increased number of patients needing hospital admissions and urgent care, often needed in a pandemic setting, is lower. Similarly, a higher *population density* is an indicator of potentially higher infection transmission rates, due to the closer proximity of individuals in higher population density areas.

The novel K-means-COV model approach, validated with OLS MFR, RFR, XGBoost and the Calinski-Harabasz methodologies, identified the percentage of *female smokers* and *diabetes prevalence* as the most predictive features correlating with case fatality rate of COVID-19 in the first part of analyses. In the second part of analyses, with countries clustered based on the public health indices, *female smokers, hospital beds per thousand*, and *GDP per capita* had higher predictive values. The predictive features were identified with both K-means-COV and OLS MFR methodologies, however, with a different ranking order. Smoking is also a well-known risk factor for the health of individuals, as well as the overall population, and is traditionally observed with a

higher percentage of male smokers than female smokers. The percentage of female smokers is now being recognized as a more sensitive indicator, since the numbers are steadily increasing, especially in the countries where smoking cessation measures are not rigorously implemented [39], [40]. Diabetes prevalence, similar to cardiovascular death rate mentioned earlier, is another stable indicator of underlying chronic conditions of the population that will have a higher vulnerability in any pandemic setting. The number of *hospital beds per thousand* feature was ranked high with this model, similar to the RFR and XGBoost models, as well as *GDP per capita*, indicating the economic prosperity of a country. Countries with the higher *GDP per capita* would be expected to invest more in building a stable health system infrastructure that would better sustain pressures of a pandemic.

For the United States, the XGBoost single country analysis identified *aged 65 older, median age,* and *life expectancy* to be the top three predictive features most highly correlating with the case fatality rate for COVID-19 in the PHI index. The actual post-pandemic COVID-19 data collaborates with these findings, with the highest mortality rate observed in elderly people [44]. Currently in the US, 15.4% of the overall population is 65 or older, with a median age of 38.3 years and a life expectancy of 78.9 years, representing a society that has an advanced health system and a higher quality of health care. For the XGBoost CHI index, *population density, hospital beds per thousand,* and *population* were identified as the highest predictive features. Presently 338 million people live in the US, with a population density of 35.6 and 2.8 hospital beds per thousand people.

The novel Pandemic Risk Score model is a tool to classify the pandemic risk of countries into three levels: low, medium, or high-risk categories, per public health index. Countries with a lower to medium overall Pandemic Risk Score for PHI will have a better overall pandemic

response, representing countries with a population that is younger and healthier (e.g., Norway, Finland, Sweden, United Kingdom, Denmark, Netherlands, Portugal, United States, Switzerland, etc.). Countries with the low to medium scores for CHI, have a higher human development index and a stronger health system that will have less difficulties to accommodate hospitalization of a larger number of patients. These parameters are indicators of a higher standard of living, higher economic status, and low poverty (e.g., Norway, Austria, Finland, Ireland, Switzerland, United States, United Kingdom, Slovenia, etc.). The US has a score of 14 for the Population Health Index and 12 for the Country Health Index, which represents medium pandemic risk.

Data from this research using non-pandemic parameters commonly collected annually by countries, indicates that 42.3% of the countries have a low pandemic risk for PHI, and only 15.4% for CHI. Therefore, the majority of countries worldwide have a high or medium pandemic risk. These findings highlight the need for proactive management of pandemic readiness at a country level, including strategic planning and resourcing to guide the efforts of public health officials and governments.

The second part of this research addresses the importance and impact of vaccination and the time needed to reach the critical cumulative vaccination rate thresholds (vaccination inflection points) to observe continuous decrease of the case fatality rates, signaling the turnaround point in the pandemic. It was conducted both at the aggregate and country levels, it expands into pandemic predictors, and utilizes COVID-19 actual data to refine forecasting tools for future outbreaks.

The analysis of the actual COVID-19 historical data shows that all countries had the highest fatality rates during the first year of the pandemic. The initial lowering of the case fatality rates was achieved with the pandemic measures, such as masks and social distancing, school and workplace lockdowns, and testing and tracking. The subsequent lowering of the CFR was observed

with the introduction and implementation of first vaccines in December of 2020. Several types of vaccines were available at the time of the initial vaccination: genetically engineered messenger RNA, viral vector vaccines, and protein subunit vaccine [135]. The initial vaccinations were delivered as single dose or 2-dose vaccines, followed by single dose booster vaccines to improve already established immunity. The first booster dose was approved for use in the third quarter of 2021, followed by two in 2022, and one in 2023, for a total of four booster doses, in developed countries [88]. As of today, there are approximately 40 different vaccines that were approved by regulatory agencies for full emergency use authorization across different countries [108].

In the dataset used for this research, 65% of countries started their vaccination efforts in December 2020, and 35% started in January 2021. We looked at the time to reach two distinct timepoints, the primary vaccination inflection point (PVIP) represented as the first reduction in the case fatality rate post vaccination start, and the secondary vaccination inflection point (SVIP), represented with a most significant and continuous case fatality rate drop post vaccination. Looking at the mean values, the PVIP was reached on day 83.27 at the vaccination rate of 31%, and the SVIP was reached at day 339.31 at the average vaccination rate of 67.8%. All four parameters had a very large range, signalizing the presence of outliers. Median values indicate a shorter time to reach the PVIP (57.5 days), lower vaccination rate at PVIP (6.05%), a longer time to reach the SVIP (355.5 days) and a higher overall vaccination rate at SVIP (71.25%). The results demonstrated that countries with the mid-level GDP per capita were the most successful in implementing their vaccination campaigns and had the shortest times to reach both vaccination inflection points looking at mean as well as median values. Regarding the individual countries, Finland was the first country to reach the PVIP in only 15 days with the vaccination rate of 1.1%, while Romania had the longest wait, reaching the PVIP in 367 days with a vaccination rate 27.8%.

In just 161 days (vaccination rate 63.8%), the UK saw the largest CFR reduction (SVIP), while Romania took 560 days (vaccination rate 41.6%) to reach the same threshold. The highest vaccination rate at SVIP was achieved in Portugal (89%) on November 16, 2021.

The Secondary Vaccination Inflection Point score categorizes countries based on their actual time to achieving secondary vaccination inflection point, where the lowest score represents the shortest time, and therefore, the best score. The majority of countries reached the SVIP between 300-370 days. Countries that had a shorter time to SVIP and, therefore, a lower score, had the highest median vaccination rates. This tool can be utilized to better understand and interpret the changes in the dynamics of the pandemic.

Looking specifically at the US (Supplement Table S1), PVIP was achieved after 94 days post-vaccination, at a vaccination rate of 24.8% and case fatality rate of 1.8%. The most significant reduction in case fatality rate for the SVIP was achieved at 363 days after the vaccination start date or 269 days after the PVIP date. The vaccination threshold at the SVIP was 70.3% with the case fatality rate of 1.59%. When the vaccine was started, the CFR was 1.82; however, by the time it reached the SVIP, it was only 1.59. In most countries, including the US, priority for COVID-19 vaccination was given to health care workers, residents and personnel of long-term care facilities, elderly patients, and patients with certain comorbidities. For an easier interpretation and comparison of different forecasting models, the US was grouped with Switzerland, Luxembourg, and Ireland in the high GDP per capita countries (>$50,000). The US launched the immunization effort first, but it took longer than other countries to achieve both vaccination inflection points. These results were most likely influenced by the impact of widespread anti-vaccine campaigns, scientific misinformation, and overall lack of readiness of certain parts of the population to support government efforts [93, 94, 95].

As mentioned earlier, the impact of vaccination is dependent on many factors, such as speed of implementation of the campaign, demand and supply of vaccines, acceptance of the vaccine by the targeted population, and others. It is important to mention that the direct impact of vaccination at the population level will often lag and the data may show some initial misalignments that can be explained. For example, the most significant reduction in case fatality rate in the US was observed in December 2021, signaling the turnaround of the pandemic in the US, with the steady decline of the ratio of total infections and death cases. However, in the next few months in 2022 there was a significant increase in new infections and deaths [92]. While the vaccination rate in the US at that time was reaching 70%, it can be assumed that the increase in new cases was caused by several factors, such as the delay in immunity development post vaccination, breakthrough infections, lack of booster vaccination, higher vulnerability of the unvaccinated population, relaxation of pandemic measures and, most significantly, the emergence of new variants with limited immunity coverage from existing vaccines (e.g., omicron variant BA.2.86 that emerged in Nov of 2021).

To understand the findings and applications of this research, it is important to examine the results within the appropriate context and look at the potential variables that may have influenced the results. There are four most important factors that influence the vaccination, and the case fatality rates in any country. Two are non-pandemic variables (not immediately influenced by the pandemic): percentage of people in the population who are 65 years of age or older, and the life expectancy of the population. Two additional variables are pandemic variables: percentage of people who had a confirmed COVID-19 infection with testing, and the level and scope of the pandemic measures that were implemented. The four variables were ranked in order of importance: *stringency_index* (the level of implemented pandemic measures), *aged_65_older*, *life_expectancy*,

and *positive_rate*. It would be expected that the same factors would be the most important in a potential new pandemic as well. This would be influenced by the increased vulnerability of the elderly and sick patient populations to any infectious disease, the importance of the infection transmission rates, and the speed of implementation of response measures. For any future pandemic, it would be important to have vaccines available at the outbreak, to have a fast roll out of the vaccination campaigns and cover the most susceptible populations first. These measures would significantly decrease the speed of progression and duration of a pandemic. In addition, these critical learnings highlight the need to take care of the most vulnerable parts of the populations and implement appropriate procedures for testing, vaccination, and other public health measures. This research may have been influenced by the inherited challenges of the vaccination process, described more in the Limitation section.

All foundational forecasting methodologies utilized in this research (ARIMA, Prophet, and LSTM) performed well and demonstrated good accuracy and precision, with only small numerical differences in results relative to the actual values. Different enhancement features were utilized to improve limitations of foundational models providing customization on specificities of data allowing for more robust analyses. Combining models into hybrids of foundational or foundational with enhancement models to meet the needs of the data is a newer approach that required validation. The two hybrid forecasting models (double hybrid: SARIMA-Bidirectional LSTM, and triple hybrid: SARIMA-Prophet-Bidirectional LSTM) utilized in this research are both novel models. The hybrid models were validated by comparing their results to the foundational models alone, to each other, and to the actual historical data. They both performed well with high accuracy and precision, and better than the foundational models. However, the performance and accuracy of the triple hybrid SARIMA-Prophet-Bidirectional LSTM model was superior to other models.

In addition, the anomaly and volatility detection analyses, conducted using Isolation Forest and GARCH models, validated performance of all forecasting methodologies, reporting all indicators within the typical and acceptable ranges. All foundational and hybrid models used for forecasting showed comparable results at the primary and secondary vaccination inflection timepoints and performed with high accuracy relative to the actual data.

Ability to predict the vaccination inflection point and measuring its immediate, as well as the most pronounced impacts, allows for a deeper understanding of the dynamics between the vaccination and case fatality rates. It is determined that countries can achieve a maximum vaccination rate of 70% with milder measures, and that 90% can be reached only with strict mandates imposed by governments [143]. This highlights the need to plan, organize and execute efficient vaccination campaigns, and improve surveillance and monitoring to substantially reduce morbidity and mortality and avoidance of breakdown of health care systems in countries to control potential new pandemics [144, 145, 146, 147, 148, 149, 150, 151, 152]. In addition, it is important to remember that the data for this research was trained based on the specificities of the COVID-19 pandemic. For the use of these forecasting models for future pandemics, they may need to be re-trained with the data specific to the new pandemic.

The results of this research, both addressing the impact of the non-pandemic factors on the pandemic outcomes, and the determination and the timing to reach primary and secondary vaccination inflection points can guide countries in the assessment of the pandemic risk and inform public health policy makers in creating measures to minimize the impact of any potential future pandemic and it impact on people, environment, and socio-economic systems.

9.2 Limitations

This research has several limitations that can be utilized to guide further research. The methodology used in the first part of this research has several limitations. For example, the RFR model with multiple decision trees may be fast to train but slower and ineffective for real time predictions and can result in overfitting for datasets in presence of outliers. It may provide feature importance rankings while not providing complete visibility into the coefficients, as in linear regression algorithms [49]. Additionally, when training on a smaller dataset or with many decision trees, the XGBoost Regressor approach may become overfit. This model is also computationally intensive, with multiple hyperparameters which must be tuned [52]. Both RFR and XGBoost do not include lagged features that can increase the dimensionality of the dataset, especially when using multiple lag time steps. For this dissertation, both methods were enhanced with the distribution lag, leading to more robust predictions and helping in preventing overfitting [47]. K-means-COV sensitivity methodology requires the specification of the number of clusters (K) and is sensitive towards outliers [57]. The COV approach yields a relative variability measure, but it does not provide information about the nature or sources of the variability, which makes it more difficult to comprehend the results and necessitates additional domain knowledge and context [58]. OLS Multifactor Regression model is also sensitive to outliers and to overfitting, which can reduce the prediction accuracy of the model [65]. This model assumes that the data is linear with no multicollinearity between the features of the dataset. To address non-linearity in the dataset for this research, the data was appropriately scaled and normalized using the StandardScaler() method in Python before the OLS Multifactor Regression methodology was ran [45]. In addition, some limitations may be identified in: A) Further enchantment of methodologies: 1) addressing overfitting by increasing dataset size, introducing new cross

validation techniques (instead of the 10 Fold cross validation used for this research), and alternative model enhancements [47], [51], [54]; 2) Inclusion of additional machine learning methodologies [Support Vector Machines (SVM), K-Nearest-Neighbors (KNN), and Perceptron (a neural network model)] can be utilized to improve efficiency with less hyperparameters, while still obtaining high prediction accuracy rates [55]; 3) Introducing additional performance enhancements of regression models (e.g., bagging, boosting, or stacking) [56]; 4) Implementing different K-means clustering methods to create a more flexible and interpretable clustering structure (hierarchical clustering) [63]; 5) Enhancing COV sensitivity methodology to provide more stable estimates of variability of outliers (median absolute deviation or trimmed mean), adaptation to time series data to consider temporal dependencies and autocorrelation, (rolling or time-varying COV) [64]; 6) Utilizing statistical formulas on new public health indices (PHI and CHI) to assess how they perform in comparison to cutting-edge machine learning algorithms [such as SVM, KNN, Support Vector Regression, Linear, Logistic, Multinominal and Ordinal Logistic Regression, Chi-Squared Test, and Analysis of Variance (ANOVA)]. B) Enhancement of the dataset: 1) Increasing the size, general completeness, and accuracy of the dataset, since the collection of data in the *Our World In Data* dataset is voluntary for all involved countries, limiting analyses to countries with more complete data; 2) Increasing accuracy of data utilized to calculate the case fatality rate in this dataset, since the death rates for COVID-19 may be severely underreported worldwide; 3) Increasing the number of non-pandemic parameters (predictive features) beyond what is available in the current dataset; and 4) Expanding the predictive features to include pandemic parameters in addition to non-pandemic parameters, increasing the sensitivity of the analyses.

The second part of this research, focused on vaccination has several limitations that may have

influenced the results due to the inherited challenges of the vaccination process. Published literature highlights the challenges introduced by the disparity in the distribution of COVID-19 vaccines, where majority of the vaccines were initially delivered to high – and upper middle-income countries vs lower-income countries [136]. This was evident by the differences in times to first vaccination inflection point, demonstrating that lower-income countries had a higher case fatality rate and needed a longer time to observe the case fatality rate reduction as a result of vaccinations, than the higher-income countries. Lack of availability of sufficient doses of vaccines, less organized execution of vaccine campaigns, including the order of vaccination (elderly and immunocompromised population) may have also influenced the results across countries. In addition, factors affecting vaccination acceptance, confidence in safety and efficacy and the risk of side effects, preference for natural immunity, scientifically sounding misinformation, as well as different cultures and political systems, also played a role in the observed vaccination patterns, spread of infection, and mortality of COVID-19 [137, 138, 139, 140, 141, 142]. This research has several limitations that can be utilized to guide further research, such as: (1) inherited limitations and variabilities of the vaccination campaigns in different countries (supply, distribution, new variants reducing the effectiveness of current vaccines; (2) differences in the health system infrastructures, speed and scope of implementation of other pandemic measures across countries; (3) limitations of the Our Word In Data dataset (e.g., size, completeness, and accuracy, due to the voluntary data reporting and possible underreporting of infection and death cases; and (4) selection of machine and deep learning methodologies and enhancements.

Identified limitations can serve as a call to action and can be utilized to guide further research.

9.3 Relevance of Research to Systems Engineering

9.31 Verification and Validation (V/V)

Verification and Validation (V/V) are two crucial processes in systems engineering. They are performed to check and confirm that a system meets its intended purpose and functions as designed. Both processes are independent, however they are at the same time complementary. They are usually performed throughout the entire system development lifecycle.

Regarding the application of Verification and Validation (V/V) to this research, the verification part includes the aspect of the repeated similar results for similar countries and the validation part being the consistent behaviors for similar (related) tests. In this research, multiple tests were done from aggregate analysis to determine which predictive factors correlate the most with case fatality rate. All thirteen variables were analyzed together, and then per index (PHI and CHI analyses), for single and paired country analyses to be able to determine if the final ranking order of variables that correlate the most with case fatality rate would be consistent across all the different tests. Similar countries were also assessed for similarity of the results. After performing all tests (Random Forest Regressor, XG-Boost Regressor, OLS Multifactor Regression Analysis, K-means-COV, and Calinski-Harabasz-COV), the results obtained at the end were consistent with each other, demonstrating consistent behaviors of performed tests and showcasing that diabetes prevalence and female smokers are the two most correlated predictive factors with case fatality rate.

In the second part of this research, the forecasting methodologies, both foundational (ARIMA, Prophet, LSTM with enhancements) and hybrid (SARIMA-Bidirectional LSTM and

SARIMA-Prophet-Bidirectional LSTM), also demonstrated consistent behavior. The obtained results across countries were similar and consistent.

9.32 Test and Measurement

In the first part of this research, utilizing regression analyses, one parametric model was used (Ordinary Least Squares Multifactor Regression) assuming a specific form for the relationship between variables and relying on assumptions about the error terms (residuals) being normally distributed and independent. In addition, two non-parametric models (Random Forest and Extreme Gradient Boosting Regressor with Distribution lag enhancements) were used. Non-parametric regression models usually make fewer assumptions about the data or the form of the relationship and are more flexible in capturing complex relationships. In this research, Random Forest and XGBoost produced similar results with XGBoost having a higher accuracy. The OLS MFR was used to validate the results. Results of all tests were similar and consistent.

Regarding the clustering analyses, the K-means-Coefficient of Variance is parametric agnostic, however, in this research, K-means methodology was used to cluster data, followed by the calculation of COV within each cluster, in essence, following a non-parametric approach (no assumptions about the data distribution within clusters. The Calinski-Harabasz model and the coefficient of variance (COV) used together are considered to be a non-parametric approach for evaluating clustering results, due to their reliance on relative measures and not requiring assumptions about the underlying data distribution. In this research, the results of both clustering analyses were similar and consistent.

The second part of this research utilized forecasting models. While forecasting analysis itself is not a type of parametric test, parametric and non-parametric approaches can be used as tools within forecasting analysis for model evaluation and selection in specific scenarios. Which

model is selected depends on the data characteristics and the forecasting problem. For example, for parametric models, ARIMA (Autoregressive Integrated Moving Average), SARIMA, and Prophet models assume a specific structure for the time series process. In addition, smoothing models assume specific decay patterns in the data. For non-parametric models, these models make fewer assumptions and can be more flexible for complex data patterns. Neural network models, such as LSTM and Bidirectional LSTM models, used in this research, can capture non-linear relationships without strict assumptions. Hybrid models (double hybrid SARIMA-Bidirectional LSTM and triple hybrid SARIMA-Prophet-Bidirectional LSTM combine the benefits of both approaches.

In this research, all foundational as well as hybrid forecasting methodologies with parametric, non-parametric, or a combination of both models, performed with high accuracy and delivered similar results, very close to the actual historical COVID-19 data.

9.33 Sensitivity Analysis

Two different types of Sensitivity Analyses were performed in this research. The first one was regarding K-means-Clustering-Coefficient of Variance sensitivity analysis which was performed in order to assess the predictive factors and establish the final order in which they should be ranked in relation to the COVID-19 case fatality rate. The COV technique was previously used to improve K-means clustering performance by including a variation coefficient weight vector to lessen the impact of irrelevant characteristics. For this study, the K-means-COV technique brought about several benefits. The COV model is an effective tool for comparing the variability of several variables within a dataset in order to determine the degree of correlation between those features. Even if two variables (features) have different scales or units of measurement, this approach enables the comparison of their variability. When comparing

datasets with different properties, it offers a standardized way to evaluate the relative dispersion of data points. K-means-COV sensitivity analysis can provide a comprehensive understanding of the link between independent (input) and dependent (output) variables. This methodology tests and assesses the robustness of the results, confirming the prediction outcomes of more traditional and usual machine learning models. K-means-COV was used in two different approaches. The first part of analyses (aggregate) clustered 26 countries based on the 13 predictive features. The second part ranked the predictive features by clustering countries based on public health indices (PHI and CHI).

The second type of sensitivity analysis in this research that was applied was the Calinski-Harabasz Method combined with Coefficient of Variance (COV) Model. The Calinski-Harabasz index was used to evaluate the clustering model when the ground truth labels are unknown and the efficacy of the clustering was verified using quantities and features relevant to the dataset. A measure of an object's cohesiveness (similarity to its own cluster) in relation to its separation (different from other clusters) is the Variance Ratio Criterion, also known as the Calinski-Harabasz (CH) Index. Cohesion was assessed here based on the distances between data points in a cluster and its cluster centroid, while separation was based on the distances between the cluster centroids and the global centroid, akin to the Elbow Method for figuring out the optimal number of clusters. Using the Calinski-Harabasz metric and methodology, the same approach was taken as in K-means-clustering to determine the optimal number of clusters for each index, PHI and CHI clustering analysis (which clustered 26 countries according to the variables in the PHI and CHI Index separately), and aggregate clustering analysis (which clustered 26 countries based on all 13 features combined). At the conclusion of both clustering analyses, the outcomes of figuring out the ideal number of clusters using the Elbow Method or the Calinski-Harabasz

method and metric and then, clustering 26 countries based on the optimal number of clusters obtained, were the same.

The results were cohesive across the multiple tests and analyses that were performed. The female smokers variable was shown to be the most frequently occurring predictive feature across various study sets, as demonstrated by the excellent performance and successful validation of the unique K-means-COV sensitivity analysis methodology approach across all three methodologies (Random Forest Regressor, XGBoost Regressor, and OLS Multifactor Regression). The Calinski-Harabasz methodology was also utilized to validate the novel K-means-COV model approach. The results were most sensitive to the different country clusters that were obtained across the two different clustering analysis methodologies performed in this research (K-means Clustering and Calinski-Harabasz clustering, both combined with Coefficient of Variance analysis). All of the machine learning methods that were assessed demonstrated excellent accuracy and predictive value.

9.34 Repeatability/resilience/reliability

Repeatability is very important in Systems Engineering because of the goal to achieve the same system behavior under the same conditions. Consistent behavior under controlled conditions provides confidence that the system is functioning as designed, creating accurate models for simulation and analysis. This predictable system behavior can be assessed during testing and verification phases. In this research, the same input was given in the controlled setting producing similar and consistent results, demonstrating high repeatability, while any system can demonstrate high repeatability in controlled tests, it may behave differently in real-

world scenarios under different conditions. This research was done in the background of the COVID-19 pandemic data. In a setting of a different pandemic, it may show variations.

In contrast to repeatability, reliability focuses on consistent performance under the expected operating conditions, meaning that the system can function for longer periods of time without any failures. This is essential to consider in real-world setting that introduces complexities, variations and uncertainties.

Resilience is another important concept in Systems Engineering. It ensures that a system can continue to function or recover quickly when it encounters disruptions, unexpected events, or stresses. By designing for resilience, engineers can create systems that are more reliable, adaptable, and better equipped to handle the complexities of the real world.

As described in the Verification and Validation section, in the first part of this research that utilizes regression analyses, the results were similar. They were also logically consistent, with minor differences that are explainable due to the small variable sample size and similar correlation coefficients. In the second part of this research, the results were also similar, consistent, and very close to actual values of historical COVID-19 data. Anomaly and volatility analyses also validated the accuracy of the models, as described in Section 8.4.

9.4 Future Research

Throughout human history, different kinds of health issues, global epidemics and pandemics created numerous problems. The difficulty of these problems in many cases were at the level of extinction. Several factor, such as the modernization, exceptional mobility, and rapid evolution of the human society, created tremendous difficulties how to distinguish, manage, and

successfully conquer health problems and issues. In the previous century, the Spanish Flu pandemic had more casualties than World War I. In this century, it was the COVID-19 global pandemic that literally shut down human society. The consequences were not just economic, but also existential. In the future, especially because of rising temperatures and climate changes, air and water contaminations, and melting of permafrost, numerous new viruses and bacteria may be released. These microorganisms, hidden from us for millions of years, will be presented to us, modern humans, as new. Not a single living cell or an organism existing now has had any encounter with them before, therefore, there will be no known natural immunity to these microorganisms. Once again, we may have rapidly spreading global pandemics with devastating consequences.

This highlights the importance of research work in public health, the understanding of learnings of prior pandemics, and the implementation of findings. We have the responsibility to analyze the COVID-19 pandemic in order to prepare better for what is yet to come. The research questions addressed in this dissertation are aimed to simplify and better understand predictors of pandemic responses at the country level. These findings can also be utilized to create simplified monitoring tools at the global level, and to classify expected "behavior" of countries when facing pandemic challenges, looking at the pre-pandemic factors or understanding the vaccination inflection points better. All can and should be used for risk mitigation and pandemic readiness planning.

This research opens the possibilities of additional research. Machine and deep learning methodologies used in this research, were compared, enhanced, and built to improve accuracy and performance. Double and triple hybrid forecasting methods open the door to improved predictive analysis to better assess trajectory of future pandemics. In addition to accuracy,

performance, and predictability of the models, variability plays a significant role. The term "variability" in systems engineering describes the natural fluctuation or irregularity in a system or the data it generates. Variability presents a major challenge when interpreting pandemic data for various reasons. Most common data sources for the pandemic were obtained from hospitals, public health organizations, and research institutes. It should be acknowledged that there may be discrepancies in these sources due to different data collection methods. Additionally, pandemics are greatly impacted by human behavior. Variations can be observed in factors such as travel patterns, population density, and adherence to public health initiatives. Furthermore, biological factors, such as the transmissibility and virulence of viruses, might vary due to their own mutations. As a result of that, each person's immunological reaction is unique. This variability may result in the misinterpretation of data. If variability is not taken into account, trends may not be reflective of the total population. Also, if models developed on pandemic data fail to take variability into account, they may not be very useful. Public health initiatives based on highly variable data will not be as successful as planned, leading to suboptimal strategies.

In order to better understand future pandemics, some experiments that can be carried out after analyzing COVID-19 data include creating simulation models that take into account variables such as human behavior, viral characteristics, and healthcare capacity in order to forecast the effects of various interventions in various scenarios. Long-term research on vaccine effectiveness can also monitor the efficacy of various vaccinations against emerging variations and declining immunity. In addition, social network analysis can be carried out to examine the dissemination of information and disinformation on social networks and to create more effective communication plans for future outbreaks. Additionally, creating Decentralized Trial Networks creates networks of medical facilities to carry out quick clinical trials for novel medications and

vaccinations in the event of a pandemic occurring in the future.

While deep learning and machine learning can be effective tools for assessing pandemic data, in order to produce more reliable and generalizable conclusions it is important to take variability into account by creating ensemble approaches that aggregate predictions from several models trained on distinct data subsets. Additionally, by applying various K-means clustering techniques to produce a more adaptable and comprehensible clustering structure (hierarchical clustering), unsupervised learning techniques can be used to find hidden patterns and anomalies in large datasets, comprehend the dynamics of the pandemic, and forecast future outbreaks.

A probabilistic technique that enables continuous model updates, as new data become available, is introduced by utilizing additional time series analysis models, such as the Bayesian Structural Time Series (BSTS) model. This is very helpful in fast moving circumstances like pandemics. Additionally, it offers a quantification of uncertainty, recognizing the underlying variety in the data. Furthermore, based on past data and taking seasonality and outside influences into account, merging these models to create double and triple hybrids offers a far more accurate forecast and prediction of future trends in cases, hospitalizations, and resource needs.

Explainable AI (XAI) techniques, such SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), are another set of machine learning methodology worth investigating. More specifically, the SHAP technique rates the significance of information that the model uses to generate predictions. This can highlight potential biases resulting from variability in individual data points and assist in determining which parts of the data are most significant. By building more easily understood models around those particular data points, the LIME approach provides an explanation for each unique model forecast. This can demonstrate how variations in particular data points could impact the results of the model.

Building more resilient systems for pandemic planning, response, and recovery requires systems engineers and data scientists to acknowledge variability and use these approaches and procedures.

9.5 Conclusions

In conclusion, this research will add to the overall body of knowledge base in the areas of machine and deep learning, as well as in public health. The first part of this research confirms that machine learning techniques, RFR, XGBoost, MFR, as well as a novel K-means-COV sensitivity analyses, are powerful tools for assessment and ranking of the strongest predictors of pandemic vulnerability. In the area of public health, Population Health Index and Country Health Index, the two novel indices, as well as the novel Pandemic Risk Scoring model, provide an additional approach for assessing country pandemic vulnerability based on traditional non-pandemic parameters and can serve as a powerful indicator and a call to action.

The second part of this research demonstrates that the novel hybrid time series forecasting models, combining foundational models with enhanced features, provides better performance and higher accuracy over traditional foundational models. Anomaly and volatility detection analyses were effectively used to validate the triple hybrid SARIMA-Prophet-Bidirectional LSTM model, which demonstrated improved performance and accuracy compared to previous models. In addition, it shows that 42% of countries had seen an immediate effect of vaccination in <50 days, and 23.1% of countries reached the most pronounced impact in <300 days, suggesting the need for improvements. Using cutting-edge AI techniques to estimate the period until vaccination inflection points unique to each country and comparing vaccination rates to case fatality rates can offer another helpful tool to help countries prepare for pandemic risk.

Bibliography

[1] World Health Organization. (Jan. 5, 2020). *Pneumonia of Unknown Cause—China*. [Online]. Available: https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON229

[2] The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team ‘‘The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)—China, 2020,’’ *China CDC Weekly*, vol. 2, no. 8, pp. 113–122, 2020, doi: 10.46234/ccdcw2020.032.

[3] *WHO Director-General’s Opening Remarks at the Media Briefing on COVID-19*, World Health Org., Geneva, Switzerland, 2020.

[4] World Health Organization. (2023). *WHO Coronavirus (COVID-19) Dashboard*. Accessed: Apr. 27, 2023. [Online]. Available: https://covid19.who.int/

[5] Hospital for Special Surgery and HSS News. (Nov. 5, 2021). *HSS Study Identifies Risk Factors for ‘Long-Haul’ COVID-19 in People with Rheumatic Diseases*. Accessed: Apr. 27, 2023. [Online]. Available: https://news.hss.edu/hss-study-identifies-risk-factors-for-long-haulcovid-19-in-people-with-rheumatic-diseases/

[6] American Hospital Association. (2021). *American Hospital Association Homepage: AHA*. Accessed: Apr. 27, 2023. [Online]. Available: https://www.aha.org/

[7] United Nations. (2020). *UNDESA World Social Report 2020 | DISD*. Accessed: Apr. 27, 2023. [Online]. Available: https://www.un.org/development/desa/dspd/world-social-report/2020-2.html

[8] E. Mathieu, H. Ritchie, L. Rodés-Guirao, C. Appel, C. Giattino, J. Hasell, B. Macdonald, S. Dattani, D. Beltekian, E. Ortiz-Ospina, and M. Roser. (2020). *Coronavirus Pandemic (COVID-19)*. [Online]. Available: https://ourworldindata.org/coronavirus

[9] S. Markovic, I. Salom, A. Rodic, and M. Djordjevic, ‘‘Analyzing the GHSI puzzle of whether highly developed countries fared worse in COVID-19,’’ *Sci. Rep.*, vol. 12, no. 1, Oct. 2022, Art. no. 17711, doi: 10.1038/s41598-022-22578-2.

[10] D. S. Kennedy, V. Vu, H. Ritchie, R. Bartlein, O. Rothschild, D. G. Bausch, M. Roser, and A. C. Seale, ‘‘COVID-19: Identifying countries with indicators of success in responding to the outbreak,’’ *Gates Open Res.*, vol. 4, p. 62, Sep. 2021, doi: 10.12688/gatesopenres.13140.2.

[11] M. M. I. Bhuiyanm, M. M. M. Ahmed, A. Alvi, M. S. Islam, P. Mondal, M. A. Hossain, and S. N. M. A. Hoque, ‘‘On predicting COVID-19 fatality ratio based on regression using machine learning model,’’ in *Advanced Information Networking and Applications* (Lecture Notes in Networks and Systems), vol. 450. Cham, Switzerland: Springer, 2022, doi: 10.1007/978-3-030-99587-4_28.

[12] K. L. Foster and A. M. Selvitella, ‘‘On the relationship between COVID-19 reported fatalities early in the pandemic and national socio-economic status predating the pandemic,’’ *AIMS Public Health*, vol. 8, no. 3, pp. 439–455, 2021, doi: 10.3934/publichealth.2021034.

[13] D. B. Duong, A. J. King, K. A. Grépin, L. Y. Hsu, J. F. Lim, C. Phillips, T. T. Thai, I. Venkatachalam, F. Vogt, E. L. Y. Yam, S. Bazley, L. D.-J. Chang, R. Flaugh, B. Nagle, J. D. Ponniah, P. Sun, N. K. Trad, and D. M. Berwick, ‘‘Strengthening national capacities for pandemic preparedness: A cross-country analysis of COVID-19 cases and deaths,’’ *Health Policy Planning*, vol. 37, no. 1, pp. 55–64, Jan. 2022, doi: 10.1093/heapol/czab122.

[14] A. Tiwari, A. V. Dadhania, V. A. B. Ragunathrao, and E. R. A. Oliveira, ‘‘Using machine learning to develop a novelCOVID-19 vulnerability index (C19VI),’’ *Sci. Total Environ.*, vol. 773, Jun. 2021, Art. no. 145650, doi: 10.1016/j.scitotenv.2021.145650.

[15] B. Sadeghi, R. C. Y. Cheung, and M. Hanbury, ‘‘Using hierarchical clustering analysis to evaluate COVID-19 pandemic preparedness and performance in 180 countries in 2020,’’ *BMJ Open*, vol. 11, no. 11, Nov. 2021, Art. no. e049844, doi: 10.1136/bmjopen-2021-049844.

[16] M. Coccia, ‘‘Preparedness of countries to face COVID-19 pandemic crisis: Strategic positioning and factors supporting effective strategies of prevention of pandemic threats,’’ *Environ. Res.*, vol. 203, Jan. 2022, Art. no. 111678, doi: 10.1016/j.envres.2021.111678.

[17] Y.-H.Ying,W.-L. Lee,Y.-C. Chi, M.-J. Chen, and K. Chang, ‘‘Demographics, socioeconomic context, and the spread of infectious disease: The case of COVID-19,’’ *Int. J. Environ. Res. Public Health*, vol. 19, no. 4, p. 2206, Feb. 2022, doi: 10.3390/ijerph19042206.

[18] F. F. Tavares and G. Betti, ‘‘The pandemic of poverty, vulnerability, and COVID-19: Evidence from a fuzzy multidimensional analysis of deprivations in Brazil,’’ *World Develop.*, vol. 139, Mar. 2021, Art. no. 105307.

[19] S. Alkire, R. Nogales, N. N. Quinn, and N. Suppa, ‘‘Global multidimensional poverty and COVID-19: A decade of progress at risk?’’ *Social Sci. Med.*, vol. 291, Dec. 2021, Art. no. 114457.

[20] S. K. Satapathy, S. Saravanan, S. Mishra, and S. N. Mohanty, ‘‘A comparative analysis of multidimensional COVID-19 poverty determinants: An observational machine learning approach,’’ *New Gener. Comput.*, vol. 41, no. 1, pp. 155–184, Mar. 2023, doi: 10.1007/s00354-023-00203-8.

[21] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, ‘‘Applications of machine learning and artificial intelligence for COVID-19 (SARS-CoV-2) pandemic: A review,’’ *Chaos, Solitons Fractals*, vol. 139, Oct. 2020, Art. no. 110059, doi: 10.1016/j.chaos.2020.110059.

[22] J. Kaliappan, K. Srinivasan, S. M. Qaisar, K. Sundararajan, C.-Y. Chang, and C. Suganthan, ‘‘Performance evaluation of regression models for the prediction of the COVID-19 reproduction rate,’’ *Frontiers Public Health*, vol. 9, Sep. 2021, Art. no. 729795, doi: 10.3389/fpubh.2021.729795.

[23] S. Bala, ‘‘COVID-19 outbreak prediction analysis using machine learning,’’ *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 9, no. 1, pp. 1–7, 2021, doi: 10.22214/ijraset.2021.32690.

[24] C. Zhan, Y. Zheng, H. Zhang, and Q. Wen, ‘‘Random-forest-bagging broad learning system with applications for COVID-19 pandemic,’’ *IEEE Internet Things J.*, vol. 8, no. 21, pp. 15906–15918, Nov. 2021, doi: 10.1109/JIOT.2021.3066575.

[25] S. Ballı, ‘‘Data analysis of COVID-19 pandemic and short-term cumulative case forecasting using machine learning time series methods,’’ *Chaos, Solitons Fractals*, vol. 142, Jan. 2021, Art. no. 110512, doi: 10.1016/j.chaos.2020.110512.

[26] C.-P. Kuo and J. S. Fu, ‘‘Evaluating the impact of mobility on COVID-19 pandemic with machine learning hybrid predictions,’’ *Sci. Total Environ.*, vol. 758, Mar. 2021, Art. no. 144151, doi: 10.1016/j.scitotenv.2020.144151.

[27] A. Zamitalo, Q. Xie, M. Allam, P. Philip, W. Shi, F. Giuste, B. Marteau, M. Murakoso, and M. D. Wang, ‘‘Development of machine learning regression model for COVID-19 drug target prediction,’’ in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Las Vegas, NV, USA, Dec. 2022, pp. 2808–2815, doi: 10.1109/BIBM55620.2022.9995319.

[28] A. W. Sievering, P. Wohlmuth, N. Geßler, M. A. Gunawardene, K. Herrlinger, B. Bein, D. Arnold, M. Bergmann, L. Nowak, C. Gloeckner, I. Koch, M. Bachmann, C. U. Herborn, and A. Stang, ‘‘Comparison of machine learning methods with logistic regression analysis in creating predictive models for risk of critical in-hospital events in COVID-19 patients on hospital admission,’’ *BMC Med. Informat. Decis. Making*, vol. 22, no. 1, p. 309, Nov. 2022, doi: 10.1186/s12911-022-02057-4.

[29] J.Wu, P. Zhang, L. Zhang,W. Meng, J. Li, C. Tong, Y. Li, J. Cai, Z. Yang, J. Zhu, M. Zhao, H. Huang, X. Xie, and S. Li, ‘‘Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results,’’ *MedRxiv*, Apr. 2020.

[30] A. Loreggia, A. Passarelli, and M. S. Pini, ‘‘The effects of air quality on the spread of the COVID-19 pandemic in Italy: An artificial intelligence approach,’’ 2021, *arXiv:2104.12546*.

[31] Z. Cao, Z. Qiu, F. Tang, S. Liang, Y. Wang, H. Long, C. Chen, B. Zhang, C. Zhang, Y. Wang, K. Tang, J. Tang, J. Chen, C. Yang, Y. Xu, Y. Yang, S. Xiao, D. Tian, G. Jiang, and X. Du, ‘‘Drivers and forecasts of multiple waves of the coronavirus disease 2019 pandemic: A systematic analysis based on an interpretable machine learning framework,’’ *Transboundary Emerg. Diseases*, vol. 69, no. 5, pp. e1584–e1594, Sep. 2022, doi: 10.1111/tbed.14492.

[32] M. Kazemi, N. L. Bragazzi, and J. D. Kong, ‘‘Assessing inequities in COVID-19 vaccine roll-out strategy programs: A cross-country study using a machine learning approach,’’ *Vaccines*, vol. 10, no. 2, p. 194, Jan. 2022, doi: 10.3390/vaccines10020194.

[33] D. McCoy, W. Mgbara, N. Horvitz, W. M. Getz, and A. Hubbard, ‘‘Ensemble machine learning of factors influencing COVID-19 across U.S. counties,’’ *Sci. Rep.*, vol. 11, no. 1, Jun. 2021, Art. no. 11777, doi: 10.1038/s41598-021-90827-x.

[34] S. Katragadda, R. T. Bhupatiraju, V. Raghavan, Z. Ashkar, and R. Gottumukkala, ‘‘Examining the COVID-19 case growth rate due to visitor vs. local mobility in the United States using machine learning,’’ *Sci. Rep.*, vol. 12, no. 1, Jul. 2022, Art. no. 12337, doi: 10.1038/s41598-022-16561-0.

[35] M. Tumbas, S. Markovic, I. Salom, and M. Djordjevic, ‘‘A largescale machine learning study of sociodemographic factors contributing to COVID-19 severity,’’ *Frontiers Big Data*, vol. 6, Mar. 2023, Art. no. 1038283, doi: 10.3389/fdata.2023.1038283.

[36] C. Nicholson, L. Beattie, M. Beattie, T. Razzaghi, and S. Chen, ‘‘A machine learning and clustering-based approach for county-level COVID-19 analysis,’’ *PLoS ONE*, vol. 17, no. 4, Apr. 2022, Art. no. e0267558, doi: 10.1371/journal.pone.0267558.

[37] G. A. Mensah, G. A. Roth, and V. Fuster, ‘‘The global burden of cardiovascular diseases and risk factors: 2020 and beyond,’’ *J. Amer. College Cardiol.*, vol. 74, pp. 2529–2532, Nov. 2019.

[38] C. W. Tsao et al., ‘‘Heart disease and stroke statistics—2023 update: A report from the American Heart Association,’’ *Circulation*, vol. 147, no. 8, pp. e93–e621, Feb. 2023, doi: 10.1161/CIR.0000000000001123.

[39] A. Jafari, A. Rajabi, M. Gholian-Aval, N. Peyman, M. Mahdizadeh, and H. Tehrani, ‘‘National, regional, and global prevalence of cigarette smoking among women/females in the general population: A systemic review and meta-analyses,’’ *Environ. Health Preventive Med.*, vol. 26, p. 5, Jan. 2021, doi: 10.1186/s12199-020-00924-y.

[40] NIDA. (Apr. 26, 2023). *Introduction*. [Online]. Available: https://nida.nih.gov/publications/research-reports/tobacco-nicotine-ecigarettes/introduction

[41] M. Fang, D. Wang, J. Coresh, and E. Selvin, ‘‘Undiagnosed diabetes in U.S. adults: Prevalence and trends,’’ *Diabetes Care*, vol. 45, no. 9, pp. 1994-2002, Sep. 2022, doi: 10.2337/dc22-0242.

[42] L. Liu, ‘‘Biostatistical basis of inference in heart failure study,’’ in *Heart Failure: Epidemiology and Research Methods*. 2018, doi: 10.1016/B978-0-323-48558-6.00004-9.

[43] S. Ren and A. Fan, ‘‘K-means clustering algorithm based on coefficient of variation,’’ in *Proc. 4th Int. Congr. Image Signal Process.*, Shanghai, China, Oct. 2011, pp. 2076–2079, doi: 10.1109/CISP.2011.6100578.

[44] Centers for Disease Control and Prevention. (2023). *COVID-19 Deaths by Age Distribution*. [Online]. Available: https://data.cdc.gov/widgets/9bhghcku?mobile_redirect=true

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, ‘‘Scikit-learn: Machine Learning in Python,’’ *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[46] V. Singh, M. Pencina, A. J. Einstein, J. X. Liang, D. S. Berman, and P. Slomka, ‘‘Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging,’’ *Sci. Rep.*, vol. 11, no. 1, Jul. 2021, Art. no. 14490, doi: 10.1038/s41598-021-93651-5.

[47] A. R. Khan, K. T. Hasan, S. Abedin, and S. Khan, ‘‘Distributed lag inspired machine learning for predicting vaccine-induced changes in COVID-19 hospitalization and intensive care unit admission,’’ *Sci. Rep.*, vol. 12, no. 1, Nov. 2022, Art. no. 18748, doi: 10.1038/s41598-022-21969-9.

[48] Q. Pan, F. Harrou, and Y. Sun, ‘‘A comparison of machine learning methods for ozone pollution prediction,’’ *J. Big Data*, vol. 10, no. 1, p. 63, May 2023, doi: 10.1186/s40537-023-00748-x.

[49] M. Aria, C. Cuccurullo, and A. Gnasso, ‘‘A comparison among interpretative proposals for random forests,’’ *Mach. Learn. Appl.*, vol. 6, Dec. 2021, Art. no. 100094, doi: 10.1016/j.mlwa.2021.100094.

[50] M. Schonlau and R. Y. Zou, ‘‘The random forest algorithm for statistical learning,’’ *Stata J., Promoting Commun. Statist. Stata*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.

[51] A. Mansoori, M. Zeinalnezhad, and L. Nazarimanesh, ‘‘Optimization of tree-based machine learning models to predict the length of hospital stay using genetic algorithm,’’ *J. Healthcare Eng.*, vol. 2023, pp. 1–14, Feb. 2023, doi: 10.1155/2023/9673395.

[52] J. Montomoli et al., ‘‘Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients,’’ *J. Intensive Med.*, vol. 1, no. 2, pp. 110–116, Oct. 2021, doi: 10.1016/j.jointm.2021.09.002.

[53] N. M. Shahani, X. Zheng, C. Liu, F. U. Hassan, and P. Li, ‘‘Developing an XGBoost regression model for predicting Young’s modulus of intact sedimentary rocks for the stability of surface and subsurface structures,’’ *Frontiers Earth Sci.*, vol. 9, Oct. 2021, Art. no. 761990, doi: 10.3389/feart.2021.761990.

[54] B. Sekeroglu, Y. K. Ever, K. Dimililer, and F. Al-Turjman, ‘‘Comparative evaluation and comprehensive analysis of machine learning models for regression problems,’’ *Data Intell.*, vol. 4, no. 3, pp. 620–652, Jul. 2022, doi: 10.1162/dint_a_00155.

[55] N. Lin, Y. Chen, H. Liu, and H. Liu, ‘‘A comparative study of machine learning models with hyperparameter optimization algorithm for mapping mineral prospectivity,’’ *Minerals*, vol. 11, no. 2, p. 159, Feb. 2021, doi: 10.3390/min11020159.

[56] N. Altman and M. Krzywinski, ‘‘Ensemble methods: Bagging and random forests,’’ *Nature Methods*, vol. 14, no. 10, pp. 933–934, Oct. 2017, doi: 10.1038/nmeth.4438.

[57] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, ‘‘K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data,’’ *Inf. Sci.*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[58] M. Ahmed, R. Seraj, and S. M. S. Islam, ‘‘The k-means algorithm: A comprehensive survey and performance evaluation,’’ *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020, doi: 10.3390/electronics9081295.

[59] M. Zubair, M. A. Iqbal, A. Shil, M. J. M. Chowdhury, M. A. Moni, and I. H. Sarker, ‘‘An improved K-means clustering algorithm towards an efficient data-driven modeling,’’ *Ann. Data Sci.*, pp. 1–20, Jun. 2022, doi: 10.1007/s40745-022-00428-2.

[60] A. Ahmad and L. Dey, ‘‘A k-mean clustering algorithm for mixed numeric and categorical data,’’ *Data Knowl. Eng.*, vol. 63, no. 2, pp. 503–527, Nov. 2007, doi: 10.1016/j.datak.2007.03.016.

[61] Z. Jalilibal, A. Amiri, P. Castagliola, and M. B. C. Khoo, ‘‘Monitoring the coefficient of variation: A literature review,’’ *Comput. Ind. Eng.*, vol. 161, Nov. 2021, Art. no. 107600.

[62] H. Xiao and Y. Duan, ‘‘Sensitivity analysis of correlated inputs: Application to a riveting process model,’’ *Appl. Math. Model.*, vol. 40, nos. 13–14, pp. 6622–6638, Jul. 2016, doi: 10.1016/j.apm.2016.02.008.

[63] J. Qi, Y. Yu, L. Wang, J. Liu, and Y. Wang, ‘‘An effective and efficient hierarchical K-means clustering algorithm,’’ *Int. J. Distrib. Sensor Netw.*, vol. 13, no. 8, Aug. 2017, Art. no. 155014771772862, doi: 10.1177/1550147717728627.

[64] C. N. P. G. Arachchige, L. A. Prendergast, and R. G. Staudte, ‘‘Robust analogs to the coefficient of variation,’’ *J. Appl. Statist.*, vol. 49, no. 2, pp. 268–290, Jan. 2022, doi: 10.1080/02664763.2020.1808599.

[65] S. Rambotti and R. L. Breiger, ‘‘Extreme and inconsistent: A case oriented regression analysis of health, inequality, and poverty,’’ *Socius*, vol. 6, Feb. 2020, Art. no. 2378023120906064, doi: 10.1177/2378023120906064.

[66] A. Cheshmehzangi, Y. Li, H. Li, S. Zhang, X. Huang, X. Chen, Z. Su, M. Sedrez, and A. Dawodu, ‘‘A hierarchical study for urban statistical indicators on the prevalence of COVID-19 in Chinese city clusters based on multiple linear regression (MLR) and polynomial best subset regression (PBSR) analysis,’’ *Sci. Rep.*, vol. 12, no. 1, Feb. 2022, Art. no. 1964, doi: 10.1038/s41598-022-05859-8.

[67] B. Mahaboob, B. Venkateswarlu, C. Narayana, J. R. Sankar, and P. Balasiddamuni, ‘‘A treatise on ordinary least squares estimation of parameters of linear model,’’ *Int. J. Eng. Technol.*, vol. 7, no. 4.10, p. 518, Oct. 2018, doi: 10.14419/ijet.v7i4.10.21216.

[68] W. Cheng, J. M. G. Taylor, P. S. Vokonas, S. K. Park, and B. Mukherjee, ‘‘Improving estimation and prediction in linear regression incorporating external information from an established reduced model,’’ *Statist. Med.*, vol. 37, no. 9, pp. 1515–1530, Apr. 2018, doi: 10.1002/sim.7600.

[69] P. Mishra, C. M. Pandey, U. Singh, A. Keshri, and M. Sabaretnam, ‘‘Selection of appropriate statistical methods for data analysis,’’ *Ann. Cardiac Anaesthesia*, vol. 22, no. 3, pp. 297–301, 2019, doi: 10.4103/aca.ACA_248_18.

[70] T. K. Kim, ‘‘Understanding one-way ANOVA using conceptual figures,’’ *Korean J. Anesthesiol.*, vol. 70, no. 1, pp. 22–26, 2017, doi: 10.4097/kjae.2017.70.1.22.

[71] I. H. Sarker, ‘‘Machine learning: Algorithms, real-world applications and research directions,’’ *Social Netw. Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[72] T. O. Hodson, ‘‘Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not,’’ *Geosci. Model Develop.*, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, doi: 10.5194/gmd-15-5481-2022.

[73] Muttappallymyalil et al (2022) Vaccination Rate and Incidence of COVID-19 and Case Fatality Rate (CFR): A Correlational Study Using Data From 2019 to 2021. 22 Aug 20. doi: 10.7759/cureus.28210

[74] World Health Organization. (2023). *WHO coronavirus (COVID-19) dashboard*. World Health Organization. Retrieved April 27, 2023, from https://covid19.who.int/

[75] Hospital for Special Surgery. (2021, November 5). *HSS News*. HSS Study Identifies Risk Factor for "Long-Haul" COVID-19 in People with Rheumatic Diseases. Retrieved April 27, 2023, from https://news.hss.edu/hss-study-identifies-risk-factors-for-long-haul-covid-19-in-people-with rheumatic-diseases/

[76] American Hospital Association. (2021). *American Hospital Association homepage: AHA*. American Hospital Association. Retrieved April 27, 2023, from https://www.aha.org/.

[77] "Score Data Collection Tool - Who." *World Health Organization*, World Health Organization, from www.who.int/data/data-collection-tools/score# Accessed 8 Oct. 2023.

[78] Magazzino C, Mele M and Schneider N (2020) The relationship between air pollution and COVID-19-related deaths: an application to three French cities. Applied Energy 279, 115835.

[79] Huang J et al. (2021) The oscillation-outbreaks characteristic of the COVID-19 pandemic. National Science Review 8. https://doi.org/10.1093/nsr/nwab100

[80] Coccia M. (2021). Effects of the spread of COVID-19 on public health of polluted cities: results of the first wave for explaining the dejà vu in the second wave of COVID-19 pandemic and epidemics of future vital agents. *Environmental science and pollution research international*, *28*(15), 19147–19154. https://doi.org/10.1007/s11356-020-11662-7

[81] Coccia, M. Pandemic Prevention: Lessons from COVID-19. *Encyclopedia* 2021, *1*, 433-444. https://doi.org/10.3390/encyclopedia1020036

[82] Coccia M. (2021). The impact of first and second wave of the COVID-19 pandemic in society: comparative analysis to support control measures to cope with negative effects of future infectious diseases. *Environmental research*, *197*, 111099. https://doi.org/10.1016/j.envres.2021.111099

[83] Coccia M. (2022). COVID-19 pandemic over 2020 (with lockdowns) and 2021 (with vaccinations): similar effects for seasonality and environmental factors. *Environmental research*, *208*, 112711. https://doi.org/10.1016/j.envres.2022.112711

[84] Coccia M. (2022). Optimal levels of vaccination to reduce COVID-19 infected individuals and deaths: A global analysis. *Environmental research*, *204*(Pt C), 112314. https://doi.org/10.1016/j.envres.2021.112314

[85] Coccia M. (2022). Preparedness of countries to face COVID-19 pandemic crisis: Strategic positioning and factors supporting effective strategies of prevention of pandemic threats. *Environmental research*, *203*, 111678. https://doi.org/10.1016/j.envres.2021.111678

[86] Aldila D et al. (2021) Impact of early detection and vaccination strategy in COVID-19 eradication program in Jakarta, Indonesia. BMC Research Notes 14, 132.

[87] Fontanet and Cauchemez (2020) COVID-19 herd immunity: where are we? Nat Rev Immunol 2020 Oct;20(10):583-584.doi: 10.1038/s41577-020-00451-5.

[88] Centers for Disease Control and Prevention. (2023, December 7). *Use of updated COVID-19 vaccines 2023–2024 formula for persons aged ≥6 months: Recommendations of the Advisory Committee on Immunization Practices - United States, September 2023*. Centers for Disease Control and Prevention. https://www.cdc.gov/mmwr/volumes/72/wr/mm7242e1.htm

[89] Magazzino, C., Mele, M., & Coccia, M. (2022). A machine learning algorithm to analyse the effects of vaccination on COVID-19 mortality. *Epidemiology and infection*, *150*, e168. https://doi.org/10.1017/S0950268822001418

[90] R. Jeffrey Melton, Robert C. Sinclair (2021) COVID-19 Infection Rates Are Related to Population Rates of Vaccination: A Response to Subramanian and Kumar. https://www.researchgate.net/publication/355929758

[91] Chen Y. T. (2023). Effect of vaccination patterns and vaccination rates on the spread and mortality of the COVID-19 pandemic. *Health policy and technology*, *12*(1), 100699. https://doi.org/10.1016/j.hlpt.2022.100699

[92] E. Mathieu, H. Ritchie, L. Rodés-Guirao, C. Appel, C. Giattino, J. Hasell, B. Macdonald, S. Dattani, D. Beltekian, E. Ortiz-Ospina, and M. Roser. (2020). *Coronavirus Pandemic (COVID-19)*. [Online]. Available: https://ourworldindata.org/coronavirus

[93] Kwok, S. W. H., Vadde, S. K., & Wang, G. (2021). Tweet Topics and Sentiments Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Analysis. *Journal of medical Internet research*, *23*(5), e26953. https://doi.org/10.2196/26953

[94] Lincoln, T. M., Schlier, B., Strakeljahn, F., Gaudiano, B. A., So, S. H., Kingston, J., Morris, E. M. J., & Ellett, L. (2022). Taking a machine learning approach to optimize prediction of vaccine hesitancy in high income countries. *Scientific reports*, *12*(1), 2055. https://doi.org/10.1038/s41598-022-05915-3

[95] Liew, T. M., & Lee, C. S. (2021). Examining the Utility of Social Media in COVID-19 Vaccination: Unsupervised Learning of 672,133 Twitter Posts. *JMIR public health and surveillance*, *7*(11), e29789. https://doi.org/10.2196/29789

[96] Kir et al (2022) Augmented Artificial Neural Network Model for the COVID-19 Mortality Prediction: Preliminary Analysis of Vaccination in Turkey. DOI: 10.35377/saucis.05.01.999373

[97] Guha, S., Kodipalli, A. (2023). Deep Learning Sequence Models for Forecasting COVID-19 Spread and Vaccinations. In: Tistarelli, M., Dubey, S.R., Singh, S.K., Jiang, X. (eds) Computer Vision and Machine Intelligence. Lecture Notes in Networks and Systems, vol 586. Springer, Singapore. https://doi.org/10.1007/978-981-19-7867-8_29

[98] Noroozi-Ghaleini, E., & Shaibani, M. J. (2023). Investigating the effect of vaccinated population on the COVID-19 prediction using FA and ABC-based feed-forward neural networks. *Heliyon*, *9*(2), e13672. https://doi.org/10.1016/j.heliyon.2023.e13672

[99] Rashed, E. A., Kodera, S., & Hirata, A. (2022). COVID-19 forecasting using new viral variants and vaccination effectiveness models. *Computers in biology and medicine*, *149*, 105986. https://doi.org/10.1016/j.compbiomed.2022.105986

[100] Xue, J., Chen, J., Hu, R., Chen, C., Zheng, C., Su, Y., & Zhu, T. (2020). Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. *Journal of medical Internet research*, *22*(11), e20550. https://doi.org/10.2196/20550

[101] Cresswell, K., Tahir, A., Sheikh, Z., Hussain, Z., Domínguez Hernández, A., Harrison, E., Williams, R., Sheikh, A., & Hussain, A. (2021). Understanding Public Perceptions of COVID-19 Contact Tracing Apps: Artificial Intelligence-Enabled Social Media Analysis. *Journal of medical Internet research*, *23*(5), e26618. https://doi.org/10.2196/26618

[102] Hussain, A., Tahir, A., Hussain, Z., Sheikh, Z., Gogate, M., Dashtipour, K., Ali, A., & Sheikh, A. (2021). Artificial Intelligence-Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States: Observational Study. *Journal of medical Internet research*, *23*(4), e26627. https://doi.org/10.2196/26627

[103] Cheng, C., Jiang, W. M., Fan, B., Cheng, Y. C., Hsu, Y. T., Wu, H. Y., Chang, H. H., & Tsou, H. H. (2023). Real-time forecasting of COVID-19 spread according to protective behavior and vaccination: autoregressive integrated moving average models. *BMC public health*, *23*(1), 1500. https://doi.org/10.1186/s12889-023-16419-8

[104] Dhamodharavadhani, S., & Rathipriya, R. (2023). Vaccine rate forecast for COVID-19 in Africa using hybrid forecasting models. *African health sciences*, *23*(1), 93–103. https://doi.org/10.4314/ahs.v23i1.11

[105] Kumar, R., Gupta, M., Agarwal, A., Mukherjee, A., & Islam, S. M. N. (2023). Epidemic efficacy of Covid-19 vaccination against Omicron: An innovative approach using enhanced residual recurrent neural network. *PloS one*, *18*(3), e0280026. https://doi.org/10.1371/journal.pone.0280026

[106] Liu, Longjian. (2018). Biostatistical Basis of Inference in Heart Failure Study. 10.1016/B978-0-323-48558-6.00004-9.

[107] M. M. Vlajnic and S. J. Simske, "Accuracy and Performance of Machine Learning Methodologies: Novel Assessments of Country Pandemic Vulnerability Based on Non-Pandemic Predictors," in IEEE Access, vol. 11, pp. 90575-90594, 2023, doi: 10.1109/ACCESS.2023.3307495.

[108] World Health Organization. (2023, August 8). *Status of covid-19 vaccines within who EUL/PQ evaluation process*. Status of COVID-19 Vaccines within WHO EUL/PQ evaluation process. https://extranet.who.int/prequal/sites/default/files/document_files/Status_COVID_VAX_08August2023.

[109] Thomas, S. J., Moreira, E. D., Jr, Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Pérez Marc, G., Polack, F. P., Zerbini, C., Bailey, R., Swanson, K. A., Xu, X., Roychoudhury, S., Koury, K., Bouguermouh, S., Kalina, W. V., Cooper, D., Frenck, R. W., Jr, Hammitt, L. L., … C4591001 Clinical Trial Group (2021). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine through 6 Months. *The New England journal of medicine*, *385*(19), 1761–1773. https://doi.org/10.1056/NEJMoa2110345

[110] Katella, K. (2023, October 5). *Comparing the COVID-19 vaccines: How are they different?*. Yale Medicine. https://www.yalemedicine.org/news/covid-19-vaccine-comparison

[111] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, ‘‘Scikit-learn: Machine Learning in Python,’’ *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[112] Nicholson, C., Beattie, L., Beattie, M., Razzaghi, T., & Chen, S. (2022). A machine learning and clustering-based approach for county-level COVID-19 analysis. *PloS one*, *17*(4), e0267558. https://doi.org/10.1371/journal.pone.0267558

[113] Rguibi, M. A., Moussa, N., Madani, A., Aaroud, A., & Zine-Dine, K. (2022). Forecasting Covid-19 Transmission with ARIMA and LSTM Techniques in Morocco. *SN computer science*, *3*(2), 133. https://doi.org/10.1007/s42979-022-01019-x

[114] Facebook Open Source. (2023). *Prophet: Forecasting at scale*. Prophet. https://facebook.github.io/prophet/

[115] Pyo, J., Pachepsky, Y., Kim, S., Abbas, A., Kim, M., Kwon, Y. S., Ligaray, M., & Cho, K. H. (2023). Long short-term memory models of water quality in inland water environments. *Water research X*, *21*, 100207. https://doi.org/10.1016/j.wroa.2023.100207

[116] G. Li, N. Yang, A Hybrid SARIMA-LSTM Model for Air Temperature Forecasting. *Adv. Theory Simul.* 2023, 6, 2200502. https://doi.org/10.1002/adts.202200502

[117] Jin, Y.; Wang, R.; Zhuang, X.; Wang, K.; Wang, H.; Wang, C.; Wang, X. Prediction of COVID-19 Data Using an ARIMA-LSTM Hybrid Forecast Model. *Mathematics* 2022, *10*, 4001. https://doi.org/10.3390/math10214001

[118] Y. -C. Jin, Q. Cao, K. -N. Wang, Y. Zhou, Y. -P. Cao and X. -Y. Wang, "Prediction of COVID-19 Data Using Improved ARIMA LSTM Hybrid Forecast Models," in *IEEE Access*, vol. 11, pp. 67956-67967, 2023, doi: 10.1109/ACCESS.2023.3291999.

[119] Hema Priya, N., Adithya Harish, S.M., Ravi Subramanian, N., Surendiran, B. (2022). Covid-19: Comparison of Time Series Forecasting Models and Hybrid ARIMA-ANN. In: Rathore, V.S., Sharma, S.C., Tavares, J.M.R., Moreira, C., Surendiran, B. (eds) Rising Threats in Expert Applications and Solutions. Lecture Notes in Networks and Systems, vol 434. Springer, Singapore. https://doi.org/10.1007/978-981-19-1122-4_59

[120] de Araújo Morais, L. R., & da Silva Gomes, G. S. (2022). Forecasting daily Covid-19 cases in the world with a hybrid ARIMA and neural network model. *Applied soft computing*, *126*, 109315. https://doi.org/10.1016/j.asoc.2022.109315

[121] Wan Mohamad Nawi, W. I. A., K Abdul Hamid, A. A., Lola, M. S., Zakaria, S., Aruchunan, E., Gobithaasan, R. U., Zainuddin, N. H., Mustafa, W. A., Abdullah, M. L., Mokhtar, N. A., & Abdullah, M. T. (2023). Developing forecasting model for future pandemic applications based on COVID-19 data 2020-2022. *PloS one*, *18*(5), e0285407. https://doi.org/10.1371/journal.pone.0285407

[122] Borges, D., & Nascimento, M. C. V. (2022). COVID-19 ICU demand forecasting: A two-stage Prophet-LSTM approach. *Applied soft computing*, *125*, 109181. https://doi.org/10.1016/j.asoc.2022.109181

[123] Long, B.; Tan, F.; Newman, M. Forecasting the Monkeypox Outbreak Using ARIMA, Prophet, Neural Prophet, and LSTM Models in the United States. *Forecasting* 2023, *5*, 127-137. https://doi.org/10.3390/forecast5010005

[124] Shastri, S., Singh, K., Kumar, S., Kour, P., & Mansotra, V. (2020). Time series forecasting of Covid-19 using deep learning models: India USA comparative case study. *Chaos, solitons, and fractals*, *140*, 110227. https://doi.org/10.1016/j.chaos.2020.110227

[125] Devaraj, J., Madurai Elavarasan, R., Pugazhendhi, R., Shafiullah, G. M., Ganesan, S., Jeysree, A. K., Khan, I. A., & Hossain, E. (2021). Forecasting of COVID-19 cases using deep learning models: Is it reliable and practically significant?. *Results in physics*, *21*, 103817. https://doi.org/10.1016/j.rinp.2021.103817

[126] Z. Li, Y. Wang, Y. Wang, Y. Zheng and H. Su, "Covid-19 Epidemic Trend Prediction Based on CNN-StackBiLSTM," *2022 IEEE 11th Data Driven Control and Learning*

Systems Conference (DDCLS), Chengdu, China, 2022, pp. 970-975, doi: 10.1109/DDCLS55054.2022.9858588.

[127] R. R. Maaliw, Z. P. Mabunga and F. T. Villa, "Time-Series Forecasting of COVID-19 Cases Using Stacked Long Short-Term Memory Networks," *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Zallaq, Bahrain, 2021, pp. 435-441, doi: 10.1109/3ICT53449.2021.9581688.

[128] Ali, F., Ullah, F., Khan, J. I., Khan, J., Sardar, A. W., & Lee, S. (2023). COVID-19 spread control policies based early dynamics forecasting using deep learning algorithm. *Chaos, solitons, and fractals*, *167*, 112984. https://doi.org/10.1016/j.chaos.2022.112984

[129] Sah, Sweeti & B, Surendiran & Dhanalakshmi, R. & Mohanty, Sachi & Alenezi, Fayadh & Polat, Kemal. (2022). Forecasting COVID-19 Pandemic Using Prophet, ARIMA, and Hybrid Stacked LSTM-GRU Models in India. Computational and Mathematical Methods in Medicine. 2022. 1-19. 10.1155/2022/1556025.

[130] Liu, S., Wan, Y., Yang, W., Tan, A., Jian, J., & Lei, X. (2022). A Hybrid Model for Coronavirus Disease 2019 Forecasting Based on Ensemble Empirical Mode Decomposition and Deep Learning. *International journal of environmental research and public health*, *20*(1), 617. https://doi.org/10.3390/ijerph20010617

[131] da Silva, T. T., Francisquini, R., & Nascimento, M. C. V. (2021). Meteorological and human mobility data on predicting COVID-19 cases by a novel hybrid decomposition method with anomaly detection analysis: A case study in the capitals of Brazil. *Expert systems with applications*, *182*, 115190. https://doi.org/10.1016/j.eswa.2021.115190

[132] Lesouple, Julien & Baudoin, Cédric & Spigai, M. & Tourneret, Jean-Yves. (2021). Generalized Isolation Forest for Anomaly Detection. Pattern Recognition Letters. 149. 10.1016/j.patrec.2021.05.022.

[133] H. Xu, G. Pang, Y. Wang and Y. Wang, "Deep Isolation Forest for Anomaly Detection," in IEEE Transactions on Knowledge and Data Engineering, vol. 35, no. 12, pp. 12591-12604, 1 Dec. 2023, doi: 10.1109/TKDE.2023.3270293.

[134] Cai G, Wu Z, Peng L. Forecasting volatility with outliers in Realized GARCH models. *Journal of Forecasting*. 2021; 40: 667–685. https://doi.org/10.1002/for.2736

[135] Assistant Secretary for Public Affairs (ASPA). (2023, December 15). *Covid-19 vaccines*. HHS.gov. https://www.hhs.gov/coronavirus/covid-19-vaccines/index.html

[136] OCHA. (2021, April 20). *Unequal vaccine distribution self-defeating, World Health Organization chief tells Economic and Social Council's Special ministerial meeting – world*. ReliefWeb. https://reliefweb.int/report/world/unequal-vaccine-distribution-self-defeating-world-health-organization-chief-tells

[137] Lane, S., MacDonald, N. E., Marti, M., & Dumolard, L. (2018). Vaccine hesitancy around the globe: Analysis of three years of WHO/UNICEF Joint Reporting Form data-2015-2017. *Vaccine*, *36*(26), 3861–3867. https://doi.org/10.1016/j.vaccine.2018.03.063

[138] Larson H. J. (2018). The state of vaccine confidence. *Lancet (London, England)*, *392*(10161), 2244–2246. https://doi.org/10.1016/S0140-6736(18)32608-4

[139] Logan, J., Nederhoff, D., Koch, B., Griffith, B., Wolfson, J., Awan, F. A., & Basta, N. E. (2018). 'What have you HEARD about the HERD?' Does education about local influenza vaccination coverage and herd immunity affect willingness to vaccinate?. *Vaccine*, *36*(28), 4118–4125. https://doi.org/10.1016/j.vaccine.2018.05.037

[140] Kim, D., Keskinocak, P., Pekgün, P., & Yildirim, İ. (2022). The balancing role of distribution speed against varying efficacy levels of COVID-19 vaccines under variants. *Scientific reports*, *12*(1), 7493. https://doi.org/10.1038/s41598-022-11060-8

[141] Paul, E., Steptoe, A., & Fancourt, D. (2021). Attitudes towards vaccines and intention to vaccinate against COVID-19: Implications for public health communications. *The Lancet regional health. Europe*, *1*, 100012. https://doi.org/10.1016/j.lanepe.2020.100012

[142] Brumfiel, G. (2021, April 7). *Vaccine refusal may put herd immunity at risk, researchers warn*. NPR. https://www.npr.org/sections/health-shots/2021/04/07/984697573/vaccine-refusal-may-put-herd-immunity-at-risk-researchers-warn.

[143] Coccia M. (2022). Improving preparedness for next pandemics: Max level of COVID-19 vaccinations without social impositions to design effective health policy and avoid flawed democracies. *Environmental research*, *213*, 113566. https://doi.org/10.1016/j.envres.2022.113566

[144] Akamatsu, T., Nagae, T., Osawa, M., Satsukawa, K., Sakai, T., & Mizutani, D. (2021). Model-based analysis on social acceptability and feasibility of a focused protection strategy against the COVID-19 pandemic. *Scientific reports*, *11*(1), 2003. https://doi.org/10.1038/s41598-021-81630-9

[145] Lalmuanawma, S., Hussain, J., & Chhakchhuak, L. (2020). Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. *Chaos, solitons, and fractals*, *139*, 110059. https://doi.org/10.1016/j.chaos.2020.110059

[146] Peiffer-Smadja, N., Rawson, T. M., Ahmad, R., Buchard, A., Georgiou, P., Lescure, F. X., Birgand, G., & Holmes, A. H. (2020). Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases*, *26*(5), 584–595. https://doi.org/10.1016/j.cmi.2019.09.009

[147] Putrino, A., Raso, M., Magazzino, C., & Galluccio, G. (2020). Coronavirus (COVID-19) in Italy: knowledge, management of patients and clinical experience of Italian dentists during the spread of contagion. *BMC oral health*, *20*(1), 200. https://doi.org/10.1186/s12903-020-01187-3

[148] Vaishya, R., Javaid, M., Khan, I. H., & Haleem, A. (2020). Artificial Intelligence (AI) applications for COVID-19 pandemic. *Diabetes & metabolic syndrome*, *14*(4), 337–339. https://doi.org/10.1016/j.dsx.2020.04.012

[149] Xylogiannopoulos, K. F., Karampelas, P., & Alhajj, R. (2021). COVID-19 pandemic spread against countries' non-pharmaceutical interventions responses: a data-mining driven comparative study. *BMC public health*, *21*(1), 1607. https://doi.org/10.1186/s12889-021-11251-4

[150] Magazzino, C., Mele, M., & Schneider, N. (2021). Assessing a fossil fuels externality with a new neural networks and image optimization algorithm: the case of atmospheric pollutants as confounders to COVID-19 lethality. *Epidemiology and infection*, *150*, e1. https://doi.org/10.1017/S095026882100248X

[151] Romeo, L., & Frontoni, E. (2022). A Unified Hierarchical XGBoost model for classifying priorities for COVID-19 vaccination campaign. *Pattern recognition*, *121*, 108197. https://doi.org/10.1016/j.patcog.2021.108197

[152] Wu, H., Banerjee, R., Venkatachalam, I., Chougale, P. (2022). Impact of Interventional Policies Including Vaccine on COVID-19 Propagation and Socio-economic Factors: Predictive Model Enabling Simulations Using Machine Learning and Big Data. In: Arai, K. (eds) Intelligent Systems and Applications. IntelliSys 2021. Lecture Notes in Networks and Systems, vol 296. Springer, Cham. https://doi.org/10.1007/978-3-030-82199-9_60

[153] *US coronavirus vaccine tracker*. USAFacts. (2024, January 8). https://usafacts.org/visualizations/covid-vaccine-tracker-states/

[154] Office for National Statistics. (2023, March 26). *Coronavirus (COVID-19) latest insights: Vaccines*. Coronavirus (COVID-19) latest insights - Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/articles/coronaviruscovid19latestinsights/vaccines

[155] Hochreiter, S. & Schmidhuber, Jürgen, 1997. Long short-term memory. Neural computation, 9(8), pp.1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[156] Ning, Z., Dai, Z., Zhang, H., Chen, Y., & Yuan, Z. (2023). A clustering method for small scRNA-seq data based on subspace and weighted distance. *PeerJ*, *11*, e14706. https://doi.org/10.7717/peerj.14706

[157] Vlajnic, M. M., & Simske, S. J. (2024). Reaching Pandemic Milestones with Country Primary and Secondary Vaccination Inflection Points: An Assessment of Foundational and Hybrid Forecasting Methodologies. https://journalspress.com/reaching-pandemic-milestones-with-country-primary-and-secondary-vaccination-inflection-points-an-assessment-of-foundational-and-hybrid-forecasting-methodologies/