Received 4 August 2023, accepted 17 August 2023, date of publication 22 August 2023, date of current version 29 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3307495

RESEARCH ARTICLE

Accuracy and Performance of Machine Learning Methodologies: Novel Assessments of Country Pandemic Vulnerability Based on Non-Pandemic Predictors

MARCO M. VLAJNIC, (Member, IEEE), AND STEVEN J. SIMSKE, (Fellow, IEEE)
Department of Systems Engineering, Colorado State University, Fort Collins, CO 80523, USA

Corresponding author: Marco M. Vlajnic (mvlajnic@colostate.edu)

ABSTRACT The devastating effects of the COVID-19 pandemic created a need for sensitive and accurate machine learning methodologies for assessment of predictors of pandemic vulnerability. The performance of machine learning methodologies was assessed to correlate, predict, and rank selected demographic, health, and economic public health parameters, relative to COVID-19 case fatality rates in 26 countries. Random Forest Regressor (RFR) and Extreme Gradient Boosting models (XGBoost), both with distribution lags, a novel K-means-Coefficient of Variance (K-means-COV) sensitivity analysis approach and Ordinary Least Squares Multifactor Regression methodologies were used to evaluate correlation of predictive non-pandemic features, grouped into two novel public health indices, Population Health Index (PHI) and Country Health Index (CHI). A novel scoring model was developed for country level pandemic risk assessment. Multiple analyses demonstrated that XGBoost methodology had higher sensitivity and accuracy across all performance metrics relative to RFR, proving that *cardiovascular death rate* was the most dominant predictive feature for PHI for 46% of countries, and *hospital beds per thousand* people for CHI (46%). 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. All assessed machine learning methodologies performed with high accuracy and demonstrated strong predictive value. Only 42.3% of countries in the PHI and 15.4% in the CHI were identified to have a low pandemic vulnerability risk.

INDEX TERMS Country Health Index, COVID-19, K-means-Coefficient of Variance sensitivity analyses, multifactor regression, pandemic risk scoring model, Population Health Index, proactive pandemic readiness, Public Health Index, Random Forest Regressor, XGBoost Regressor.

I. INTRODUCTION

COVID-19 became the world's number one health problem in a very short timeframe. 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]. The world was not prepared, and the health systems were struggling to manage the high number of infected patients that required advanced care, often in intensive care units and with high fatality rates. Scientists and pharmaceutical manufacturers around the world were accelerating research to develop potent vaccines against COVID-19. In addition, many researchers started analyzing the available data in the effort to understand the disease, guide treatment of patients, forecast the spread of the pandemic and understand critical factors that impact preparedness of countries to respond to the pandemic from the public health perspective. The first vaccines were approved for use in December of 2020, with massive vaccination efforts

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

staring in 2021 around the world. As of March 21, 2023, COVID-19 was already responsible for 6.9 million deaths and a loss of $3.8 trillion in output worldwide, and an estimated loss of $202.6 billion in revenue for America's hospitals and healthcare systems [4], [5], [6], [7]. The real cost of this pandemic, impact on people, their mental health and long-term health complications is yet to be understood. At the same time, the full impact on the economies worldwide is unknown and will be the subject of research and investigation for years to come.

To assure better pandemic readiness, researchers worldwide analyzed available COVID-19 data, recognizing the need to define, predict, and better understand critical country-level factors contributing to the COVID-19 morbidity and mortality. Researchers 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].

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. COVID-19 Vulnerability Index and pandemic severity Impact Assessment were used for identifying and mapping vulnerable counties [14], [15]. At the country level, the Resilience index r and the Preparedness and prevention Index p, were used to measure impact of average mortality, hospital and intensive care unit occupancy, and impact on vaccination [16]. Poverty was also identified as an important factor, and multidimensional poverty indices were used at a global level, and COVID-19 poverty vulnerability index at a country level, showing considerable inequality among regions and ethnic groups and tracing the trends in increasing infection and a higher mortality rate in vulnerable regions [18], [19], [21].

Data used for analyses varied from a single data source to multiple sources supplementing the main databases, from a single data point to longitudinal data collected over time, and from several weeks 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 varied from descriptive statistics to comprehensive advanced data analytics, sophisticated machine learning and artificial intelligence methodologies, to predict and forecast development of the COVID-19 pandemic, and to screen and guide contact tracing and drug development for SARS-CoV-2 virus. Researchers utilized different methodologies: 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. It became abundantly clear that there were many factors that contribute and influence the pandemic risk at a country level. For example, population demographics, sex, age, racial minority, economic and socio-political factors, and the presence of comorbidities such as obesity and cardiovascular disease. While all these factors played a significant role in determining mortality rates, there are significant variations between countries in terms of size, public governance, expenditures in health system, as well as in testing and reporting. These variations continue to create substantial limitations in standardizing assessment approaches [10], [14], [15], [16]. Documented heterogeneity across countries necessitates more sophisticated testing methods and more simplified and standardized models. In an attempt to improve these models, non-pandemic parameters can be used to accurately and proactively predict country pandemic vulnerability. Results of these analyses should stimulate development of appropriate strategies and actions by country public health officials, policymakers, as well as disaster management agencies.

In the effort to improve upon the work done so far, this paper analyzes non-pandemic parameters utilizing four machine learning methodologies, including one novel approach. The performance and accuracy of these methodologies were assessed and compared. These models utilized a comprehensive 3-year longitudinal dataset, to assess correlation and predictive value of selected demographic, health, and economic non-pandemic parameters relative to COVID-19 case fatality rates in 26 countries. The results of these

TABLE 1. Public health indices definitions from the our world in data 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	

analyses created a foundation for the development of a novel country specific pandemic risk scoring model.

II. MATERIALS AND METHODS

A. DATA

This research used a dataset from the Oxford University *Our World in Data* Covid 19 Dataset [8]. This dataset contains data points collected on an ongoing basis from Johns Hopkins University, Center for Systems Science and Engineering COVID-19 data, OXFORD COVID-19 Government Response Tracker, and European Centre for Disease Control, from January 2020 to present. The original dataset contains data from 207 countries and territories from which 26 countries were selected for this research: United States, Canada, Finland, Iceland, Denmark, Belgium, Sweden, United Kingdom, Czechia, Slovakia, Switzerland, Slovenia, Austria, Italy, Ireland, Portugal, France, Netherlands, Luxembourg, Spain, Serbia, Bulgaria, Romania, Latvia, Cyprus, and Estonia. Data for this research paper 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.

The case fatality rate (CFR), an epidemiologic metric defined as the proportion of deaths within an observed population of interest [42], was calculated by dividing the respective values in the total deaths column by the total cases column of the dataset, 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 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.

The pandemic risk scoring model presented in Table 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 paper (Tables S198-S199).

Countries with the same Pandemic Risk score (Table 2) of predictive features were paired together to accommodate paired country analyses. Tables summarizing the distribution of country pairs can be located in the paper supplement (Tables S198, S199, and S200).

B. METHODOLOGIES

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

TABLE 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	$\leq$ 204	1	*Hospital beds per thousand*	$\geq$ 5.712	1
	205-323	2		3.966-5.711	2
	$\geq$ 324	3		$\leq$ 3.965	3
diabetes prevalence (%)	$\leq$ 5.49	1	*human development index**	$\geq$ 0.908	1
	5.50-6.99	2		0.857-0.907	2
	$\geq$ 7	3		$\leq$ 0.856	3
male smokers (%)	$\leq$ 27.7	1	*extreme poverty*	$<$ 0.20	1
	27.8-40.3	2		0.20-0.99	2
	$\geq$ 40.4	3		$>$ 1.00	3
female smokers (%)	$\leq$ 20.6	1	*gdp per capita*	$>$ 50,000	1
	20.7-29.3	2		35,000-50,000	2
	$\geq$ 29.4	3		$<$ 35,000	3
life expectancy (years)	$\geq$ 80.89	1	*population density*	$<$ 100	1
	77.97-80.88	2		100-200	2
	$\leq$ 77.96	3		$>$ 200	3
aged 65 older (%)	$\geq$ 19.82	1	*population*	$<$ 10M	1
	16.62-19.81	2		10-50M	2
	$\leq$ 16.61	3		$>$ 50M	3
median age (years)	$\geq$ 44.5	1			
	40.9-44.4	2			
	$\leq$ 40.8	3			

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. For the second part of the analyses, data was analyzed at the country level, as single and paired analyses, with 70% of the training data representing data from the start of the pandemic in March 2020 until January 2022. 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. Feature engineering, feature selection, and data quality assessments (completeness, reliability, consistency, validity, and no redundancy) were also completed. Data Exploration and Visualization was conducted utilizing oversampling with SMOTE (Synthetic Minority Oversampling Technique) [36], along with the development of correlation matrix of different variables in the dataset, and exploration of the dataset using graphics and visualization.

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. The second set of methodologies included a novel K-means-Coefficient of Variance sensitivity analysis approach validated by Ordinary Least Squares Multifactor Regression (OLS MFR) model. Research models in this paper 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.

1) 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]. To improve the performance of the RFR and XGBoost models, a distribution lag was applied to the derived case fatality rate variable. Distribution lags play important roles in explaining the short-run dynamic and long-run cumulative effects of features on a response variable [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 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 $\boldsymbol{R}^2$ 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), $\boldsymbol{R}^2$ 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. Paired country analyses were conducted to assess if countries that were paired based on the same Pandemic Risk score have the same or similar top three predictive features to the single country analyses of each country in the pair. 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.

2) K-MEANS-COEFFICIENT OF VARIANCE SENSITIVITY ANALYSIS AND ORDINARY LEAST SQUARES MULTIFACTOR REGRESSION

A novel model approach for K-means-Coefficient of Variance sensitivity analysis was introduced to evaluate predictive features and determine their final ranking order relative to the COVID-19 case fatality rate. In the past, COV methodology was used to improve K-means clustering accuracy by introducing a variation coefficient weight vector to decrease the effects of irrelevant features [43]. Using K-means-COV methodology for this research introduced several advantages. K-means clustering approach can be easily customized and adjusted to new instances and examples 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]. Applying K-means-COV sensitivity analysis provides an in-depth understanding of the relationship between independent (input) and dependent (output) variables. This methodology tests and assesses the robustness of the results and validates the prediction results of more traditional and standard machine learning models [62].

In this paper, 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 applying equation (1) below:

$$COV = \frac{\sigma}{\mu} \tag{1}$$

where, σ is the standard deviation from the mean of the country-to-country difference and μ is the mean of the country-to-country difference. The ratio of those two values, as applied in (1) 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 (OLS) Multifactor Regression. OLS Multifactor Regression 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 was validated with OLS MFR model. Additional validation was performed by conducting RFR and XGBoost analyses on remaining features, and the final ranking order of predictive features across different methodologies was compared.

3) PANDEMIC RISK SCORING MODEL

The Pandemic Risk score 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 B.1. Distribution of countries based on their total PHI and CHI risk scores is shown in the results section.

III. 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.

A. RANDOM FOREST REGRESSOR AND XGBOOST REGRESSOR RESULTS

The first part of analyses, utilizing RFR and XGBoost methodologies, ranked all 13 predictive features across the dataset of 26 countries (Table 3).

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 3 and Figure 1(a) and (b).

The accuracy and performance of both methods in the first part of analyses 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 1(a)-(b)].

The second part of RFR and XGBoost analyses analyzed the ranking order of predictive features per country

TABLE 3. 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	

(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 B.1).

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 4, with detailed information presented in the supplement of this paper.

The single country analyses indicated that *cardiovasc death rate*, *aged 65 older*, and *diabetes prevalence* are the most common top predictive features for PHI utilizing RFR analysis. Similarly, *cardiovasc 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 *hospitalbeds per thousand, human development index,* and population. Both methodologies performed with high accuracy, with XGBoost performing better on all five metrics. In addition, this paper looked at the distribution of dominant predictive features, correlating the most with the case fatality rate across countries [Table 5]. The single country analyses performed with XGBoost were utilized for this assessment.

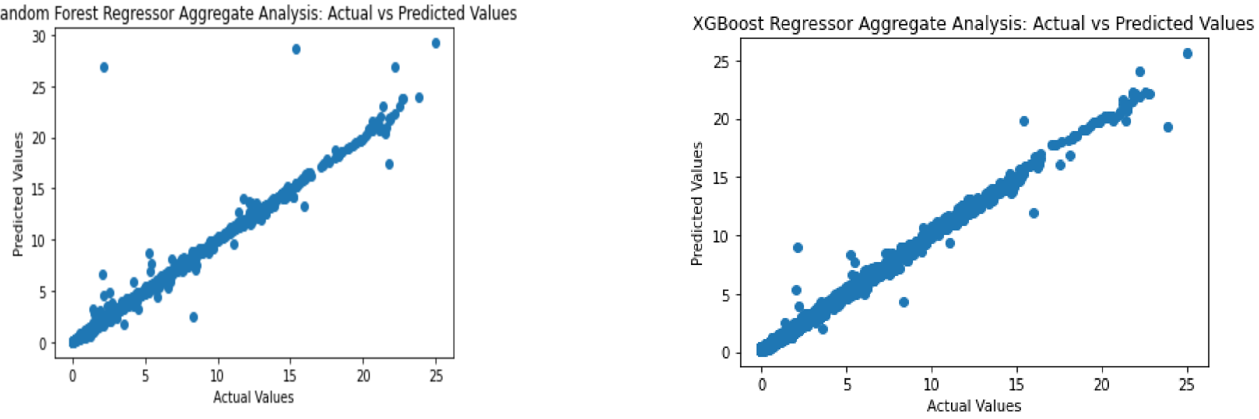
Results for CHI with XGBoost analysis were similar, with *hospital beds per thousand people, population*, and *human development index.* The accuracy of performance of the two methodologies was assessed and presented in Table 4 and Figure 2(a)-(d).

In summary, the *cardiovasc 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 6, with detailed tables presented in the supplement of this paper. The paired country analyses indicated that *diabetes prevalence, cardiovasc death rate*, and *female smokers* are the top three predictive features for the PHI utilizing RFR methodology, similar to the XGBoost PHI results. 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. Accuracy and performance of both models for the second part of analyses was performed and documented in Table 6 and Figure 3(a)-(d).

The sensitivity of the two models was assessed based on the distribution of the feature importance values, indicating that XGBoost is a more variable model with a wider 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 *cardiovasc death rate* as the dominant predictive feature for PHI (Table 7). Three country pairs were selected to represent the distribution across low, medium, and high *cardiovasc death rate* ranges.

Single versus paired country analyses showed similarities in the ranking order of features across the two methodologies. XGBoost PHI and CHI analyses indicate that two out of three features from single country analysis were included in the ranking order of features in the paired analysis, confirming better accuracy of the XGBoost model over RFR.



(a) Random Forest Regressor

(b) XGBoost Regressor

FIGURE 1. Comparison of actual vs predicted values for aggregate analyses.

TABLE 4. Summary of RFR and XGBoost single country analyses results.

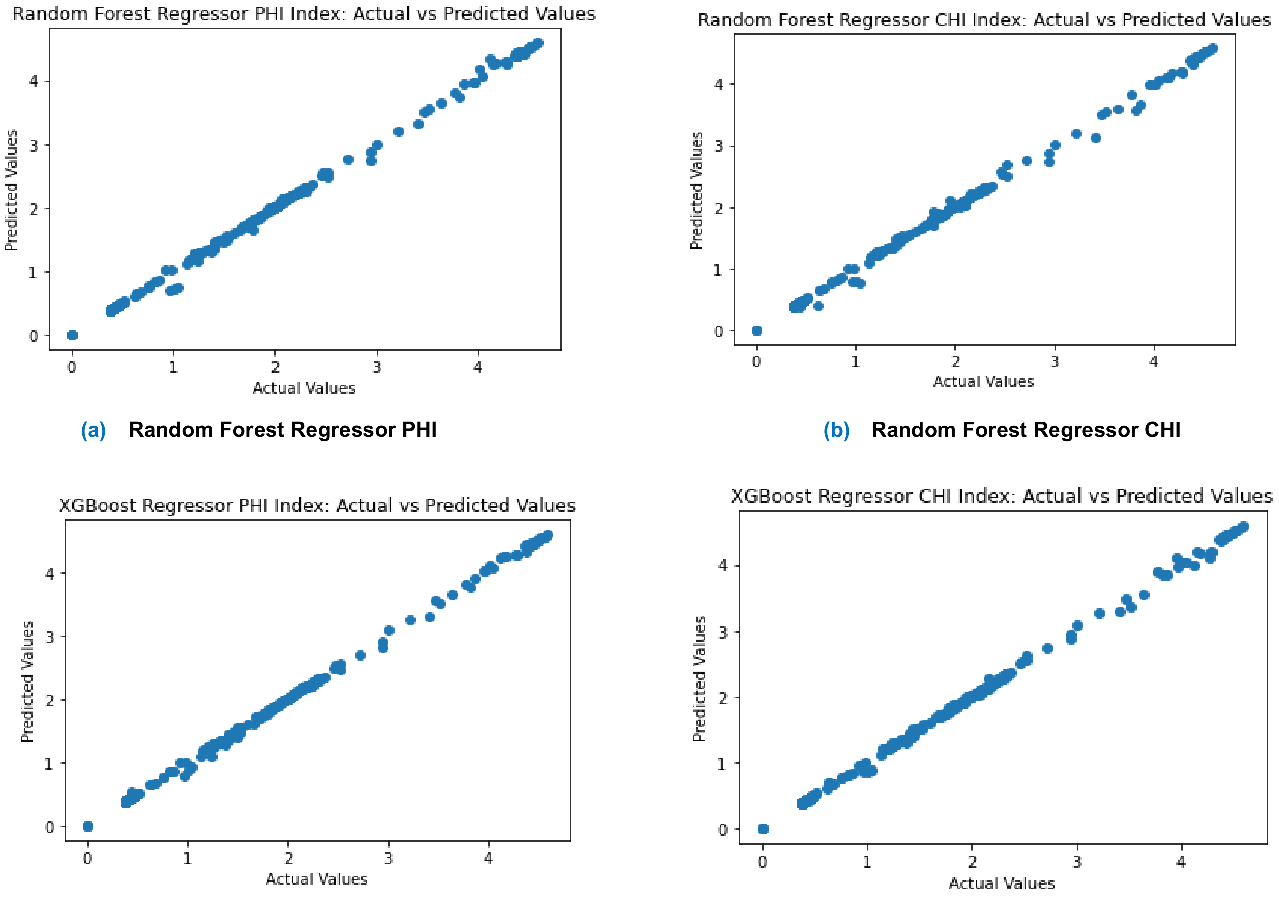
Random Forest Regressor Model				XGBoost Regressor Model			
PHI	**Importance range**	**CHI**	**Importance range**	**PHI**	**Importance range**	**CHI**	**Importance range**
cardiovasc death rate	0.0174- 0.9726	*hospital beds per thousand*	0.0020-0.9750	*cardiovasc death rate*	0.0166-0.9133	*hospital beds 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	

Single versus paired analysis was also performed for *hospital beds per thousand* people feature, as the dominant predictive feature for CHI, providing similar results (Table S203 in the supplement).

B. K-MEANS-COEFFICIENT OF VARIANCE SENSITIVITY AND OLS 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 8). 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 [Figure 4(a)], 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



(a) Random Forest Regressor PHI

(b) Random Forest Regressor CHI

(c) XGBoost Regressor PHI

(d) XGBoost Regressor CHI

FIGURE 2. **Comparison of actual vs predicted values for single country analyses.**

of six iterations. The feature importance scatter plots for both iterations are presented in Figure 4(a)-(b). 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.

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 9). The feature importance scatter plots for both iterations are presented in Figures 5(a)-(b) and 6. The final importance rank for the second part of analyses (per PHI and CHI) was defined based on the line graph of the last iteration of COV versus OLS MFR Feature Importance, demonstrating a linear relationship between the remaining features.

Both sets of analyses with 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, validating the novel K-means-COV sensitivity methodology approach. 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 10).

C. PANDEMIC RISK SCORING MODEL RESULTS

The Pandemic Risk Scoring Model was developed based on the feature ranges (Table 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 Figure 7(a)-(b) and Table 11.

As shown in the Figure 7(a) (PHI), 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, in Figure 7(b), shows that 69.2% of countries have medium pandemic risk, while 15.4% of countries have low and

TABLE 5. Distribution of predictive features across countries.

Distribution of dominant predictive features across countries					
PHI			**CHI**		
feature	**countries (total, %)**	**countries**	**feature**	**countries (total, %)**	**countries**
cardiovasc death rate	12 (46%)	Bulgaria, Czechia, France, Finland, Irlanda, 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			

high risk. The United States is classified as a medium risk country, for both indices, together with most of the countries in Europe.

IV. DISCUSSION AND LIMITATIONS

During the COVID-19 pandemic, significant research was conducted to enhance understanding and improve the status of the pandemic. This type of research was driven by academic and research sites. The devastating effect of the COVID-19 pandemic created a need for the development of more sophisticated machine learning testing models and simplified standardized tools that can increase usability and interactivity at the country level. Models and tools that use commonly collected non-pandemic parameters allow countries to participate in 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.

This paper 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. These analyses were done to rank demographic, health, and economic parameters (predictive features) for 26 countries relative to their importance and correlation with the COVID-19 case fatality rates and grouped into two novel public health indices, Population Health Index and Country Health Index. Grouping of variables allowed for the interpretation of the results in the appropriate context of public health and for a 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 RFR and XGBoost models were selected as the most relevant models for feature importance evaluation, frequently utilized by researchers [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. 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.

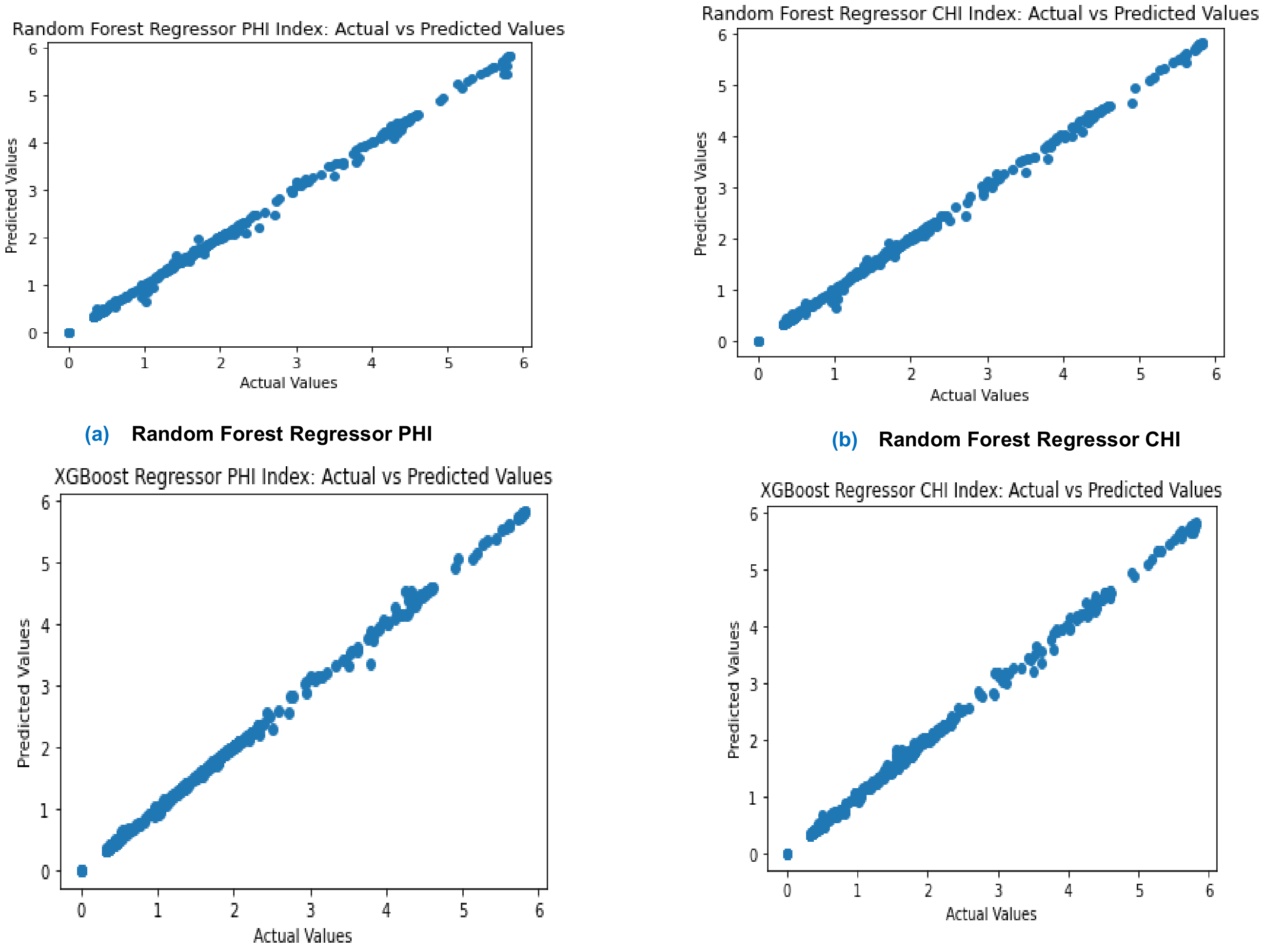
The K-means-Coefficient of Variance sensitivity analysis was developed, assessed, and validated with OLS MFR methodology. Additional validation was conducted with RFR and XGBoost methodologies showing similar results in the final ranking order of predictive features. The novel approach

TABLE 6. Summary of RFR and XGBoost paired country analyses 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
cardiovasc death rate	0.0008-0.9692	*extreme poverty*	0.0135-0.9739	*cardiovasc 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	

of K-means-COV sensitivity methodology brings additional value to the field of Systems Engineering. 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 methodology used in this research paper 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



(a) Random Forest Regressor PHI

(b) Random Forest Regressor CHI

(c) Random Forest Regressor PHI

(d) Random Forest Regressor CHI

FIGURE 3. Comparison of actual vs predicted values for paired country analyses.

rankings while not providing complete visibility into the coefficients, as in linear regression algorithms [49]. XGBoost Regressor methodology can also be vulnerable to overfitting with multiple decision trees or when trained on a smaller dataset. 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 research paper, both methods were enhanced with the distribution lag, leading to more robust predictions and help in preventing overfitting [47]. K-means-COV sensitivity methodology requires the specification of the number of clusters (K) and is sensitive towards outliers [57]. COV methodology provides a measure of relative variability but does not give insights into the nature or causes of the variability, making the interpretation of results more difficult and in need of further context and domain knowledge [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 paper, the data was appropriately scaled and normalized using the StandardScaler() method in Python before the OLS Multifactor Regression methodology was ran [45].

The XGBoost Regressor model performed better and with higher accuracy than RFR. The XGBoost single country analysis identified *cardiovasc 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 *cardiovasc 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.

TABLE 7. Single vs paired country analyses based on *cardiovasc death rate*.

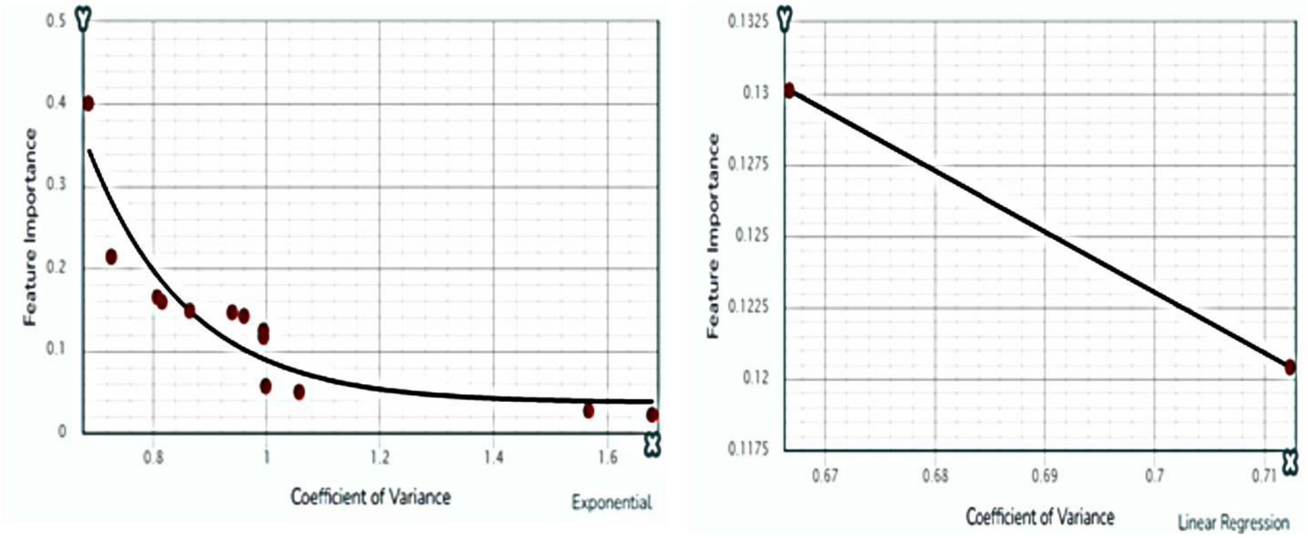
Single vs paired country analyses; feature: *cardiovasc death rate* per 100,000 people					
		RFR		**XGBoost**	
Single/paired country	**Actual *cardiovasc death rate***	**PHI**	**CHI**	**PHI**	**CHI**
United Kingdom	122.137	*aged 65 older, diabetes prevalence, median age*	*hospital beds per thousand, 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*

These results support the hypothesis of *cardiovasc death rate* as a well-known 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, this vulnerability applies to countries with a higher diabetes prevalence [41]. Countries with a higher life expectancy usually have a population that is older and therefore more frail and 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 and XGBoost, 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 a known risk factor for the health of individuals, as well as the overall population, and is traditionally

TABLE 8. 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**		**COV Vs Feature Importance Values**	
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
hospital beds per thousand	0.6874	*cardiovasc 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
cardiovasc 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



(a) First Iteration

(b) Last Iteration

FIGURE 4. Coefficient of variance (COV) versus ordinary least squares (OLS) multifactor regression feature importance scatter plot.

observed with a higher percentage of male smokers than female smokers. It is not surprising that the percentage of female smokers is now being identified 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, is a steady 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*, both being clear indicators of the economic

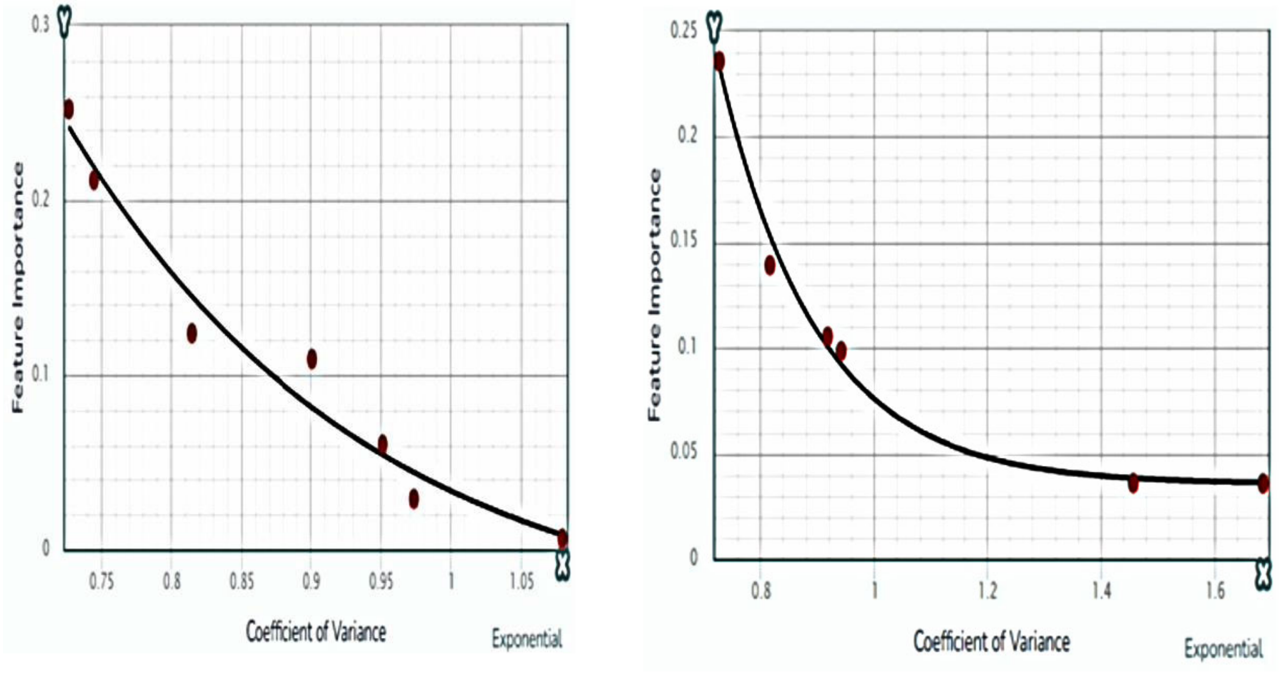
TABLE 9. Summary of results across K-means-COV and OLS MFR with comparison, Second part of analyses, first and last iteration.

First Iteration					
K-means clustering with COV Sensitivity Analysis		**OLS Multifactor Regression**		**COV Vs Feature Importance Values**	
PHI					
Features	**Coefficient of Variance**	**Features**	**Importance Values**	**Coefficient of Variance**	**Feature Importance Values**
female smokers	0.7271	*cardiovasc 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
cardiovasc 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**		**COV Vs Feature Importance Values (Fifth Iteration)**	
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

prosperity of the country, and also an indirect marker of investments in the health system infrastructure and pandemic readiness measures.

Looking specifically at the United States and the results from the XGBoost single country analysis, the research identified *aged 65 older, median age,* and *life expectancy* as 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 allows countries to be classified into low, medium, or high-risk categories. In general, countries with a lower to medium overall PRS PHI are countries that will have a better overall pandemic response, representing countries with a younger and healthier



(a) PHI

(b) CHI

FIGURE 5. Coefficient of variance (COV) versus ordinary least squares (OLS) multifactor regression feature importance scatter plot first iteration.

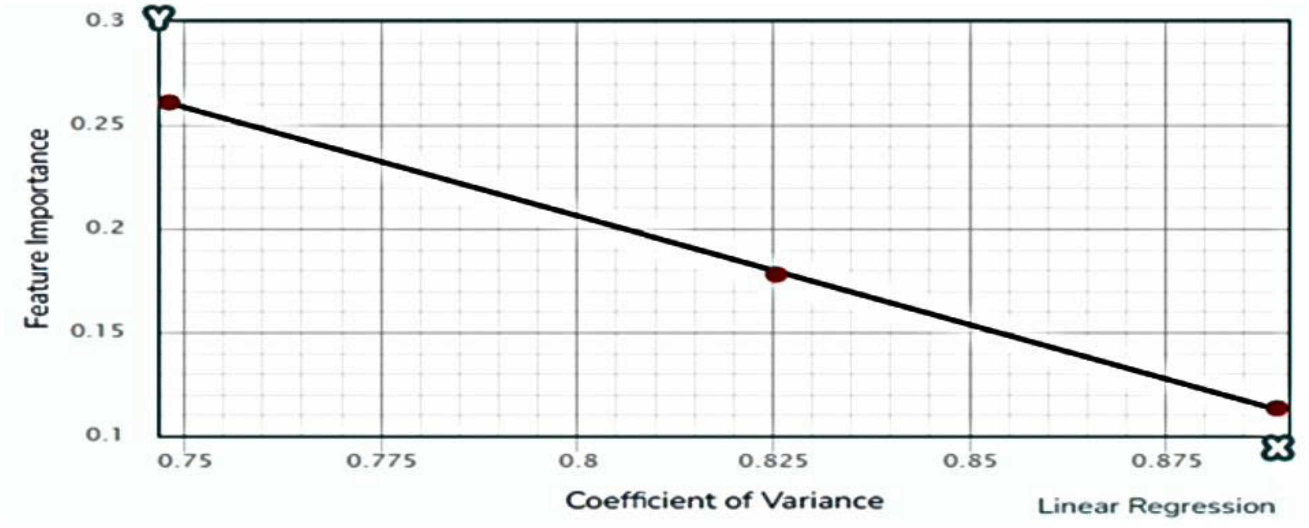


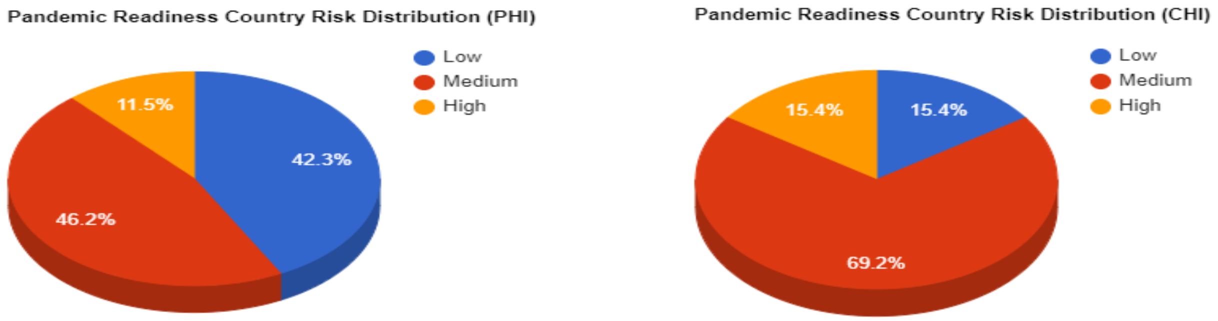
FIGURE 6. Coefficient of variance (COV) versus ordinary least squares (OLS) multifactor regression feature importance line graph last iteration.

population (e.g., Norway, Finland, Sweden, United Kingdom, Denmark, Netherlands, Portugal, United States, Switzerland, etc.). For the CHI, low to medium scores represent countries that have a higher human development index and a stronger health system prepared 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 indicates that 42.3% of the countries have a low pandemic risk for PHI, and only 15.4% for CHI. These findings highlight the need for proactive management of pandemic readiness at a country level, including strategic planning and resourcing. These analyses were conducted based on non-pandemic parameters commonly

TABLE 10. Final ranking order of predictive features across RFR, XGBoost, COV, and 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



(a) PHI

(b) CHI

FIGURE 7. Distribution of countries based on total population health index and country health index pandemic risk scoring model.

TABLE 11. 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

collected by countries. Therefore, they are readily available, and this risk assessment should be applied to any future pandemic or health issue.

V. CONCLUSION AND FUTURE RESEARCH

In conclusion, the research conducted in this paper adds to the overall body of knowledge in machine learning and public health. It 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, the two novel indices, Population Health Index and Country Health Index, 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.

This paper has several limitations that can be utilized to guide further research:

A. ENHANCEMENTS 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) Use of statistical equations on novel public health indices (PHI and CHI) to compare to state-of-art machine learning algorithms [e.g., Linear, Logistic, Multinominal and Ordinal Logistic Regression, Chi-Squared Test, Analysis of Variance (ANOVA), and SVM, KNN, Support Vector Regression, and various deep learning neural network models].

B. ENHANCEMENTS 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.

REFERENCES

[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-haul-covid-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 novel COVID-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 large-scale 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-e-cigarettes/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/9bhg-hcku?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.

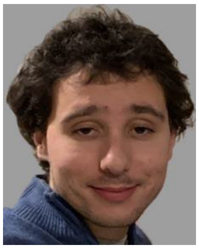
[68] W. Cheng, J. M. G. Taylor, P. S. Vokonas, S. K. Park, and B. Mukherjee, ''Improving estimation and prediction in linear regression incorporating external information from an established reduced model,'' *Statist. Med.*, vol. 37, no. 9, pp. 1515–1530, Apr. 2018, doi: 10.1002/sim.7600.

[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.



MARCO M. VLAJNIC (Member, IEEE) received the B.S. degree in computer science from Rutgers, The State University of New Jersey, New Brunswick, NJ, USA, in 2020, and the M.S. degree in computer science from the Stevens Institute of Technology, Hoboken, NJ, USA, in 2021. He is currently pursuing the Ph.D. degree in systems engineering with Colorado State University, Fort Collins, CO, USA. He is a Systems Engineer and Computer Scientist. Some of the recent projects that he was involved include: time series analysis with machine learning, risk assessment and analysis, development of convolutional and recurrent neural network models, distributed systems and cloud computing, object oriented programming, information systems and database management, and systems modeling language (SysML) architecture development. His research interests include programming, cloud computing, machine and deep learning, artificial intelligence, design and analysis of algorithms, and software development.

STEVEN J. SIMSKE (Fellow, IEEE) received the Ph.D. degree in aerospace engineering and in electrical and computer engineering from the University of Colorado. From 1994 to 2018, he was an Engineer (HP Fellow, since 2011), the Vice President, and the Director of HP Labs. Since 2018, he has been a Professor of systems engineering with Colorado State University (CSU). At CSU, he has a cadre of on-campus students in systems, mechanical, and biomedical engineering, and a larger contingent of online/remote graduate students researching various disciplines. With more than 20 years in the industry, he directed teams to research 3D printing, education, life sciences, sensing, authentication, packaging, analytics, imaging, and manufacturing. He has written four books on analytics, algorithms, and steganography. He is the author of 500 publications and 240 U.S. patents. His research interests include analytics, systems security, sensing, signal and imaging processing, printing and manufacturing, and situationally aware robotics. He is an NAI Fellow, an IS&T Fellow, and its former President (2017–2019). He completed a CSU Faculty Institute for Inclusive Excellence (FIIE) Fellowship, in 2020. He was a CSU Best Teacher Award, in 2022.