Analysis and Classification of COVID-19 Severity Using Machine Learning Techniques

Chandini Krishna Polaki[1], Md Amiruzzaman[1], Md. Rajibul Islam[2], Rizal Mohd Nor[3]

[1]Department of Computer Science, West Chester University, United States

[2]Photonics Research Institute, Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, China

[3]Department of Computer Science, International Islamic University, Malaysia

Received: March 15, 2025; Revised: May 10, 2025; Accepted: July 02, 2025; Published: July 06, 2025

Abstract

The COVID-19 pandemic has placed significant burdens on the human race over the last few years. In such surveillance, higher data analysis is required with advanced predictive modeling. Herein, classification is performed on the outbreak severity using a publicly available dataset of the daily and accumulated case and mortality data concerning COVID-19. After preprocessing, real-time missing and inconsistent values are replaced to enable feature derivation, including rolling averages. Interactive Tableau dashboards are developed to visualize severity classifications, regional trends, and temporal variations, providing dynamic insights into outbreak patterns. Comparative analysis reveals disparities in case distribution across continents, identifying major hotspots. Machine learning models are employed to predict severity levels, achieving strong performance metrics. The KNN model yields the highest accuracy of classification, which is 97.11%, while the Random Forest model is more resistant to noise and enhances the stability of the predictions. These results emphasize the potential role of machine learning and data visualization examples with Tableau for data-driven public health strategies in monitoring and responding to outbreaks.

Keywords: COVID-19, Severity Classification, Temporal Analysis, Regional Analysis, Epidemiological Data, Public Health Strategies, Outbreak Monitoring, Resilience Testing.

1 Introduction

1.1 Background

The COVID-19 pandemic severely affected global public health, economies, and societies. The high transmissibility of the virus and the high case fatality rates of COVID-19 demand the most effective outbreak monitoring and outbreak response measures [1]. The assessment of outbreak severity is an essential process to inform timely and appropriate public health responses [2, 3]. Large-scale epidemiological data provides a great opportunity for advanced analytics and machine learning to monitor and predict the trajectory of an epidemic. Incident and mortality data analyses can be performed to study patterns and trends that allow predictive modeling for the classification of severity [4, 5]. This paper considers the role of machine learning in advancing outbreak severity analysis toward more accurate and efficient data-driven decision-making in pandemic control.

1.2 Problem Statement

The outbreaks of COVID-19 are so unpredictable, and their impact is different in different regions [1, 5]. These factors make it very difficult for policymakers to make resource allocation decisions and intervene at the right time. Most of the current approaches lack adaptability and fail to leverage robust data-driven techniques, hence yielding suboptimal outcomes. This study tries to address these challenges by explicitly linking the classification of outbreak severity to well-defined objectives, such as data preprocessing, trend identification, and machine learning-based prediction toward actionable insights.

1.3 Objectives of the Study

- Data Preprocessing and Analysis: Making the COVID-19 case data and mortality data reliable and precise through missing value imputation and meaningful feature extraction.

- Trend and Pattern Identification: Analyze the temporal and geographical progressions, highlighting the main trends in the outbreaks.

- Severity Classification: Develop thresholds for new cases and deaths to classify the severity level of outbreaks systematically.

- Machine Learning Implementation: Development of training and testing models for predicting the severity using SVM, KNN, and Random Forest, emphasizing noise resilience and robustness.

- Data Visualization: Create interactive Tableau dashboards to dynamically explore regional and temporal patterns.

1.4 Scope and Delimitations

The study confines itself to numerical data from COVID-19, represented by the current and cumulative level of cases and deaths, since this is consistently available and integral for monitoring pandemic trends. These were excluded from the major socio-economic factors and healthcare capacities since to get comparable and reliable data across the regions is very complex. This allows the approach to be more focused, fully recognizing its limitations.

1.5 Significance

This study investigates the role of integrating machine learning and data visualization in developing improved public health strategies. The stated objectives of this research would help enhance outbreak monitoring, resource allocation, and targeted interventions. This ensures that the results from this study are actionable and impactful in addressing data-driven decision-making in ongoing and future health crises.

2 Related works

Machine learning techniques coupled with epidemiological data form the backbone of all strategies against COVID-19. Ajayi and Cheng [6] classified disease severity using Bayesian Networks with demographics and symptoms, anchoring their work on handling uncertainty in data. However, their use of demographic data limits generalizability across diverse populations. Another study by Marcos et al. [7] proposed a Random Forest for classifying severity among hospitalized patients. Though they achieved very good predictive accuracy, they excluded mild cases that may underestimate the widespread impact of the pandemic. On the other hand, our approach will involve all severities, giving an overall understanding of the outbreak.

Deep learning approaches have also been tried to classify COVID-19 severity. Abboju et al. [8] implemented a convolutional neural network on CT scans for the assessment of disease advancement, while Wang et al. [9] presented a hierarchical Neyman-Pearson classifier for the prioritization of server cases. While both works were highly successful within their respective areas of application, both were bound by dependencies on specialized imaging data with high computational complexity. This study will overcome these limitations by leveraging widely available epidemiological data, hence assuring far greater scalability and usability for real-time monitoring.

Data visualization tools like Tableau are effective, as exemplified by the "COVID- 19 Data Visualization Project" [10] for outbreak pattern insights. Such visualizations, however, are mostly decoupled from predictive modeling. We advance on that by integrating machine learning-driven severity classifications with interactive visualizations for actionable insights into public health decision-making.

Other studies have approached the problem of forecasting and spatiotemporal analysis. Sharma and Gupta [11] developed a hybrid ARIMA-LSTM model for forecasting cases. Li and Zhou [12] employed machine learning methods to identify outbreak hotspots. Though valuable, such studies were bereft of any form of severity classifications—a feature highly pertinent to our work. Patel and Singh [13] developed a multitask learning model to predict cases and mortality rates but had very limited preprocessing techniques for noisy datasets. Smith and Davis [14] compared various machine learning models for predicting disease severity but fell short in investigating deep feature engineering or visualization.

Explainable ML approaches, such as SHAP-based severity prediction models by Brown and Chen [15], highlighted interpretability at the cost of robustness against noisy datasets. Our study takes the lead in not only incorporating robust preprocessing techniques to handle such challenges but also checking model resiliency against noise. Our contribution further extends previous research along two important dimensions-addressing significant limitations and new contributions. First, we present an extensive preprocessing pipeline for noisy and incomplete datasets, enabling more robust model performance and generalizability to real-world data. Second, we have extensively supported a severity classification system covering mild, moderate, and severe, which is the critical direction lacking in the related works focusing on subsets of the severity levels. Third, this work embeds predictive modeling into visualization by effectively combining machine learning models, such as support vector machines, K-nearest neighbors, and Random Forest, with Tableau dashboards, offering dynamic explorations and actionable insights. Last but not least, basing our work on publicly available epidemiological data ensures that our approach is scalable and usable in near real-time since this can be adapted to diverse regions and resource settings. Accordingly, the contributions place our work as a robust, scalable, and interpretable tool for outbreak monitoring and public health decision-making.

3 Methods

3.1 Dataset Acquisition

This study uses the dataset from the open-access repository "WHO COVID-19 Cases Dataset" found in Kaggle. It is the summary of daily reports about COVID-19 cases and deaths updated every day by governments and health organizations of the world to help monitor the impact brought about by the pandemic. The data covers several countries and continents, hence good for any time and region-based variation studies. The data is found publicly on Kaggle: https://www.kaggle.com/datasets/ironwolf437/ who-covid-19-cases-dataset, from which results can be reproduced transparently.

3.2 Dataset Characteristics

The dataset includes main epidemiological indicators, namely daily and cumulative counts of new cases and deaths, complemented by metadata on countries, continents, and WHO regions. It is structured in a time-series format to allow for longitudinal analysis that captures changes over time. The necessary attributes include 'Date-reported' and 'Country-code', which define the temporal and geographical con- texts and thus allow the classification of outbreak severity. Completeness of the dataset and well-structured data make this dataset highly suitable for pre-processing, exploratory analysis, and machine learning applications in severity classification and trend forecasting.

3.3 Data Preprocessing and Integrity

The dataset represented broad coverage across geographies and time, but it was not very clean for analysis since it contained missing values, records with negative values, and duplication of records. 'New-cases' and 'New-deaths' used zero replacements for missing values to maintain the continuity of the time-series data. This will be helpful for a more meaningful analysis, logical integrity, and without replicate records. It standardized the numerical features. These are among the most important preprocessing steps toward increasing data reliability, reducing biases, and enabling robust model training and evaluation.

3.4 Feature Engineering

Feature Engineering was one of the most important parts in this analysis since it turned raw data into meaningful data. Feature engineering improves model interpretability, hence performance; thus, providing richer insights on the COVID-19 pandemic trend. Hence, the following feature engineering steps are performed:

1. Rolling Averages: The seven-day rolling average will be calculated on New-cases and New-deaths to smooth out day-to-day variability and accentuate longer trends. This helped to identify periodic patterns of the pandemic trends over time and space and dampened erratic reporting.

2. Severity Classification: The severity level is categorized as ranging from Mild, Moderate to Severe; this was enforced by defining a custom function, severity-class for classification. Since the threshold set for classification was on the daily count of New-cases and New-deaths, the function provided a standardized way of assessing the intensity of the outbreaks through time and space.

3. Aggregation by Region: This would help outline the geographical disparity, so new cases worldwide needed to be summed up at each continent level. After that, the general picture of how different continents are being affected by the pandemic could be viewed as hotspots and low-case-count areas, according to region.

4. Feature Scaling: New-cases, New-deaths, Cumulative-cases, and Cumulative-deaths were numeric. StandardScaler was applied on them. This puts all the features into comparable scales, enhancing the performance and stability of machine learning models like Support Vector Machines and K-nearest neighbors.

5. Temporal Grouping: It was a time-based grouping done to view the trend for days, weeks, and months. The transformation yielded structured views of the trajectory of the pandemic time-series data with a high periodicity of waves. The engineered features enhanced not only the exploratory and visualization efforts but also the performance and reliability of the machine learning models. Feature engineering was a necessary step in the derivation of meaningful representation from raw data, hence aligning the dataset to the goals of the study.

3.5 Analytical Methods

3.5.1 Exploratory Data Analysis (EDA)

- Descriptive statistics: Descriptive statistics were done in order to outline the general view of the dataset, where the central tendencies and patterns of the numerical variables were focused on. Calculation of the median and mode values of critical columns like New-cases, New-deaths, and cumulative cases were done. The median helped in identifying the central point at which the distribution of data rests, there- fore providing a robust measure that is not easily dragged by extreme outliers. Mode values presented the frequency of occurrence of the most frequent data that showed repeated patterns of case and death count. The statistics are thus bound to form a basis of analysis in the determination of typical trends and anomalies in such data.

- Temporal Analysis: Trends in New Cases per Day: A line plot to understand the progression of the COVID-19 pandemic over time was plotted by showing the daily new cases reported globally. These trends outlined quite clearly the temporal back- ground against which the pandemic evolved and defined key periods of rapid case growth that warranted further investigation. This is represented using a line graph. Fig 1) shows the daily new COVID-19 cases reported around the world from the beginning of the pandemic in early 2020 through 2024. The line plot conveys dis- tinct surges corresponding to major waves of the pandemic, with noticeable peaks in 2021, and 2022, and an exceptionally sharp increase in early 2023, likely reflect- ing the emergence of highly transmissible variants or reporting adjustments. These peaks are followed by sharp drops in case numbers, reflecting the combined effect of vaccination campaigns, natural immunity, and public health intervention.

- Rolling Averages: Seven-day rolling averages were calculated for New-cases and New-deaths to account for daily fluctuations and inconsistencies in reporting. These rolling averages smoothed out short-term irregularities, providing a more stable and interpretable view of the long-term trends in the pandemic's progression. Using rolling averages made such patterns, like consistent increases in cases throughout a particular wave or gradual decreases after mitigation efforts had been put in place, even clearer. This procedure also smoothed out many of the anomalies caused by underreporting on specific days and overcompensation on subsequent days, thus providing a sound basis for analysis of the general trend in the pandemic over time. Fig 2 shows the seven-day rolling averages of new COVID-19 cases by continent. This offers a clearer view of regional trends and smooths day-to-day fluctuations. The highest peaks are seen in Asia, notably during early 2023, reflecting major pandemic waves in densely populated regions. Significant surges can also be seen in North America and Europe, although less pronounced compared to those seen in Asia. In contrast, Africa, Oceania, and small others such as "Island" have much lower averages throughout, reflecting regional differences in the course of the pan- demic. This visual enables the comparison of how the pandemic has evolved across continents over time.

3.5.2 Regional and Severity Insights

- Regional Distribution: A bar chart in Fig 3 was created to visualize the total number of new COVID-19 cases reported across continents, revealing significant geographical disparities in case counts. Asia and Europe emerged as the most affected regions, accounting for the highest totals of new cases, followed by North America. These findings align with population densities, healthcare capacities, and the timing of pandemic waves in these regions. In contrast, Africa and Oceania reported significantly fewer cases, which could reflect a combination of lower population densities, differences in pandemic severity, or underreporting due to limited testing capacities. The visualization highlights how the pandemic's impact varied widely across regions, emphasizing the importance of regional strategies in combating outbreaks.

- Severity Classification: To provide structured insights into the pandemic's impact, records are classified into Mild, Moderate, and Severe level thresholds derived from percentile-based statistical distribution analysis. This classification framework allowed for a systematic understanding of outbreak severity across regions and periods. Mild cases include those with new cases below 1,158 and new deaths below 12 (below the 75th per- centile), while Moderate cases fall between 1,158 and 10,000 new cases or 12 and 500 new deaths (between the 75th and 90th percentile). Severe cases are defined as those exceeding 10,000 new cases or 500 new deaths (above the 90th percentile). Most cases are categorized as Moderate, reflecting sustained but manageable pan- demic conditions in many areas. Severe cases were concentrated in high-burden regions and during specific waves of the pandemic, while Mild cases typically represented regions with smaller outbreaks or effective mitigation efforts. This structured approach facilitated comparative analysis, enabling a clearer understanding of how different regions and periods were affected by the pandemic. The following Table 1 shows the results of displayed 20 random rows to outline the breakdown of cases and their respective classifications:

3.6 Software and Tools:

The analysis was done using different software and tools to have the potential for strong data processing and modeling. Python was the main programming language used; the libraries utilized in writing the code were Pandas, NumPy, and Matplotlib for efficient data manipulation and visualization. Implementation and evaluation of models such as Support Vector Machines, K-Nearest Neighbors, and Random Forest were done using machine learning libraries: Scikit-learn. Line plots, bar charts, and trend visualizations were powered by Matplotlib. Development, documentation, and iterative exploration in Jupyter Notebook allowed for flexibility and an interactive environment. Besides Python tools, usage of Tableau was performed in developing interactive dashboards and visualizations. This allowed for the interaction of visualizations of trends-for example, severity classifications across continents or temporal variations in new cases or deaths. Tableau allowed enhanced user interaction-filtering interactively and highlighting-enabled regional analysis of trends, comparison of sever- ity levels, and identification of patterns in a very easy way. This integrated framework of methodologies ensured that the analysis was systematic and thorough, hence making accurate predictions with meaningful insights into the data set.

Table 1: Severity Classification of 20 Random Rows

Row No.	New Cases	New Deaths	Severity
15838	0	0	Mild
50861	2309	17	Moderate
51651	0	0	Mild
23783	71988	1572	Severe
45257	40	0	Mild
23629	876	6	Mild
48109	38	1	Mild
38855	2318	80	Moderate
3579	0	0	Mild
37661	128	0	Mild
55274	1106	15	Moderate
48554	0	0	Mild
9271	0	0	Mild
5971	2312	9	Moderate
32601	7	0	Mild
8777	5	0	Mild
9860	596	6	Mild
35326	0	0	Mild
38964	179	3	Mild
48019	0	0	Mild

3.7 Tableau:

Fig 4 is a Tableau dashboard showing the severity classification of new cases across continents from 2020 to 2024 by using a heatmap. That is the trend of time where Asia and Europe reached their peak in the year 2022, and regions like Oceania and South America have been showing lower severity throughout the period. It allows interaction-operated filtering with the level of severity, hence dynamic exploration of patterns in data. Users can analyze regional trends and make comparisons, facilitating insights into severity distributions and potential areas requiring inter- vention. (https://public.tableau.com/app/profile/chandni.krishna.polaki/viz/Book1 17349737005690/Sheet1?publish=yes).

In Fig 5, Tableau dashboard of new cases and death trends over time, integrating interactive charts and filters into the visualization of patterns and variation [16]. On the right, it has a highlighter tool that allows the user to select a given number of continents within the chart for better viewing and comparison. This feature will make the use more user-friendly because it enables the precise investigation of regional trends. It evaluates patterns in data dynamically, therefore allowing change points and critical periods to be traced. (https://public.tableau.com/app/profile/chandni.krishna.polaki/viz/Book217349737317440/Sheet2?publish=yes).

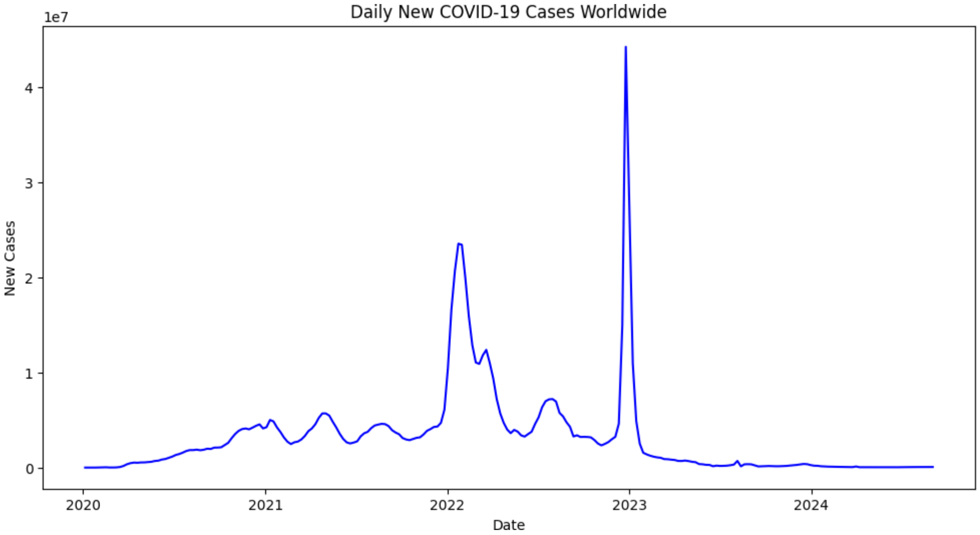

Figure 1: Daily New COVID-19 Cases Worldwide

3.8 Machine Learning Techniques

3.8.1 Support Vector Machine (SVM):

The Support Vector Machine (SVM) model was chosen since it can handle nonlinear relationships within the dataset. Using the radial basis function (RBF) kernel, the SVM tried to classify the severity levels into Mild, Moderate, and Severe by finding the best hyperplane in the feature space. The features used are New-cases, New-deaths, Cumulative-cases, and Cumulative-deaths scaled using StandardScaler, since SVM requires normalized inputs. The 80:20 training-to-testing split was done to validate the model on unseen data. Also, cross-validation was performed to see how consistent and reliable the model is across different splits of data.

3.8.2 K-Nearest Neighbors (KNN):

The K-Nearest Neighbors (KNN) algorithm was chosen for simplicity and interpretability. This choice classified the instances in this problem by taking the most frequent class among the five nearest neighbors in the feature space. Also, feature scaling was done, so all the features have an equal contribution to the distance metric used by the KNN. Based on this logic, instantiate a model with n-neighbors=5, leaving room for further tuning based on the validation results.

3.8.3 Random Forest:

The algorithm used is the Random Forest-an ensemble technique, which is particularly good at handling complicated relationships and noisy data and hence robust for the severity classification. The model has been trained on 100 decision trees. At each decision tree, a random feature subset and random sample are considered to prevent overfitting. Further, the model’s robustness is tested during the advanced analysis by adding Gaussian noise to the data it was trained on. Adding noise to data is one of the conventional ways to analyze the robustness and generalization ability of machine learning models. Here in the analysis, Gaussian noise is added to features such as New- cases and New-deaths to introduce variability and uncertainty normally expected in a real-world scenario due to many inconsistencies in the data, reporting errors, or other natural causes.

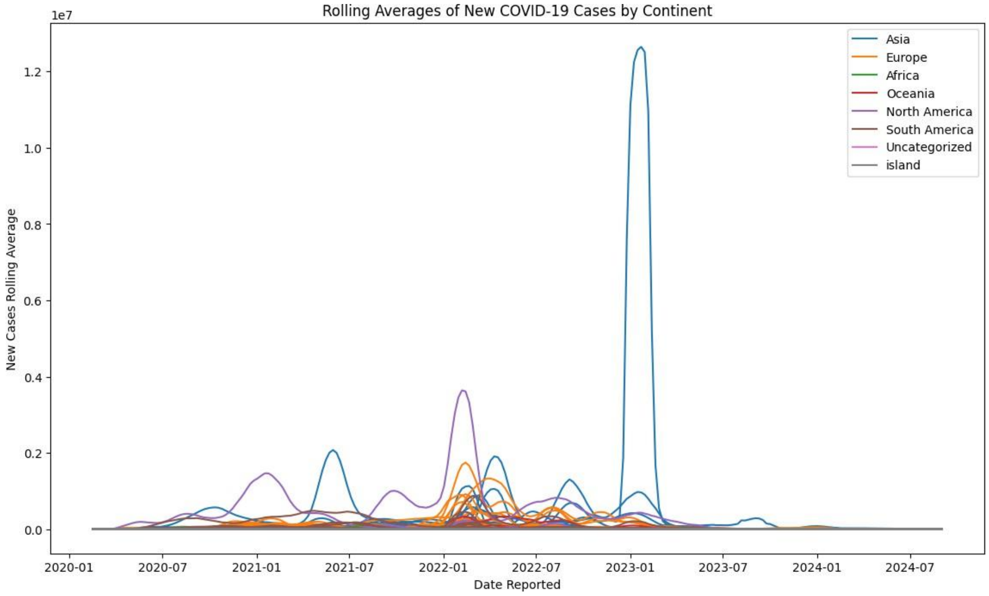

Figure 2: Rolling Averages of New COVID-19 Cases By Continent

4 Results

Global Trends: The daily cases clearly indicated the peaks for the pandemic waves, which were smoothed effectively by the seven-day rolling averages. Most of the cases were of the Moderate type, although a few were Severe and Mild.

Regional Disparities: The highest total of new cases was reported from Asia and Europe, followed by North America. On the other hand, the numbers in Africa and Oceania were much lower, reflecting either disparities in pandemic impact or differences in reporting capacities.

Temporal Analysis: Line plots showed sharp peaks that coincided with known pandemic waves, thereby providing information on the time and dissemination of infections.

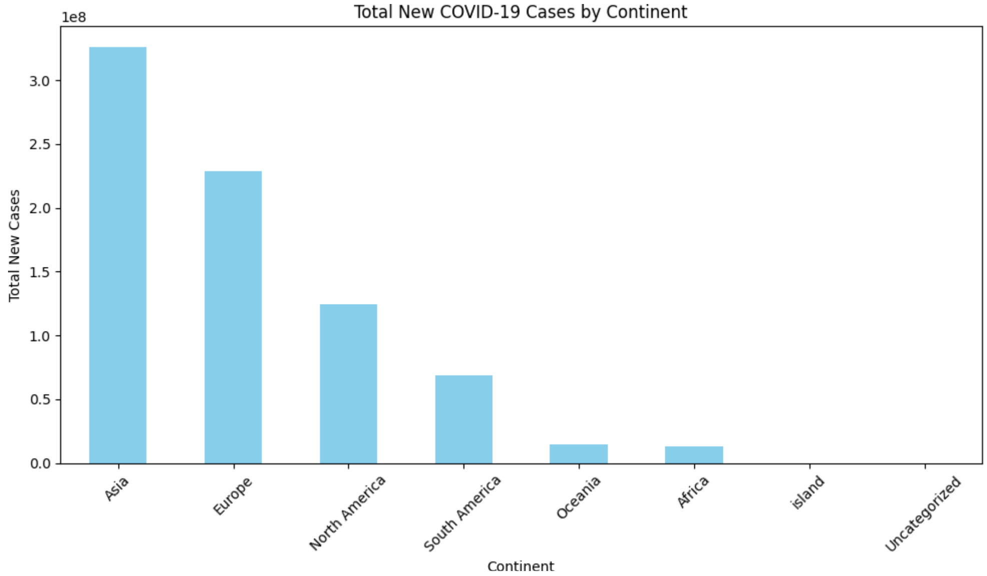

Figure 3: Total New COVID-19 Cases by Continent

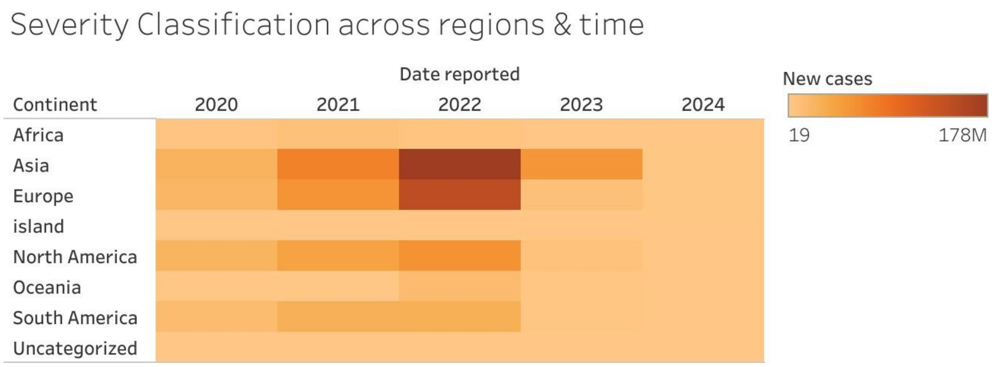

Figure 4: Severity Classification across regions and time

Regional Analysis: Bar charts showed that Asia and Europe were pandemic hotspots, whereas Africa and Oceania looked less affected, showing the regional disparities in the severity and spread of the pandemic.

SVM Model: The SVM model achieved an average accuracy of 90.49%, and the performances for the classes Mild and Severe cases were pretty impressive. As we see in Table 2, it gave a precision of 91% and recall of 99% for Mild, thus misclassifying very few. On the other side, in the case of Moderate cases, it showed 80% precision and 55% recall, thus finding it hard to distinguish them from the rest of the classes.

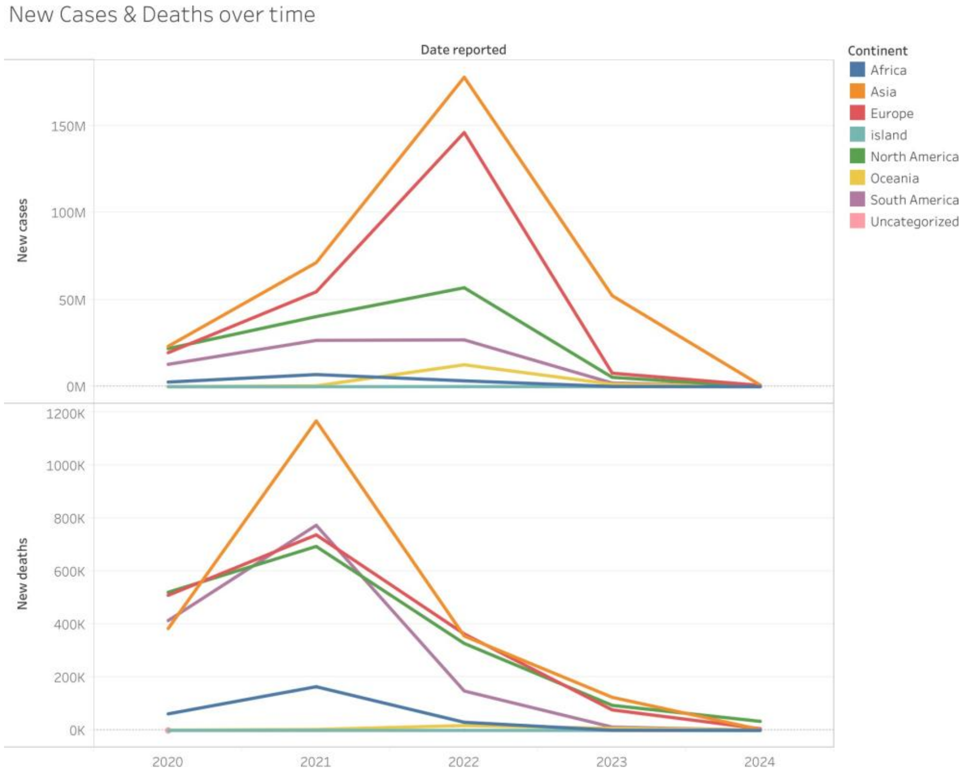

Figure 5: New Cases and Deaths over time

In contrast, Severe cases were classified with 98% precision and a recall of 85%; the confusion matrix showed that the most common sources of errors were misclassifica- tions of Moderate cases due to the significant degree of similarity between Mild and Moderate categories.

Table 2: SVM Classification Report

Class	Precision	Recall	F1-Score	Support
Mild	0.91	0.99	0.95	8541
Moderate	0.80	0.55	0.65	1895
Severe	0.98	0.85	0.91	1276
Accuracy		0.90 (Overall)		
Macro Avg	0.90	0.80	0.84	11712
Weighted Avg	0.90	0.90	0.90	11712

The Confusion Matrix of SVM in Table 3 illustrates a good performance from the model for "Mild" and "Severe" cases, which were correctly classified 8,466 and 1,089 times, respectively, while "Moderate" had 828 misclassified as "Mild" and 24 misclassified as "Severe." These show how this model struggles in "Moderate" cases, whereas for the rest of the categories, it has a good accuracy.

Table 3: SVM Confusion Matrix

Actual/Predicted	Mild	Moderate	Severe
Mild	8466	75	0
Moderate	828	1043	24
Severe	8	179	1089

KNN Model: The KNN model developed a general accuracy of 97.11%. The model performed strongly for all severity levels as we can see in Table 4. In the case of Mild, the model achieved a precision of 98% and recall of 99%, classifying most of the cases correctly with minimal error. For Moderate, the model posted a precision of 92% and recall of 90%, hence capturing most of the cases but with partial leakage into other classes. For the Severe cases, the model realized a precision of 97% and recall of 94%, hence correctly classifying 1204 out of the 1276 severe cases.

Table 4: KNN Classification Report

Class	Precision	Recall	F1-Score	Support
Mild	0.98	0.99	0.99	8541
Moderate	0.92	0.90	0.91	1895
Severe	0.97	0.94	0.96	1276
Accuracy		97.11% (Overall)		
Macro Avg	0.96	0.95	0.95	11712
Weighted Avg	0.97	0.97	0.97	11712

On the confusion matrix in Table 5, the misclassifications were limited with slight leakage mainly from the Moderate class to other classes. The high score of the model in precision, recall, and cross-validation metrics reflects robustness and efficiency with respect to the dataset.

Table 5: KNN Confusion Matrix

Actual/Predicted	Mild	Moderate	Severe
Mild	8455	85	1
Moderate	146	1714	35
Severe	14	58	1204

Random Forest: The Random Forest model initially had a general performance of 90.49% and performed well across the severity levels as we see in Table 6. The precision for Mild cases was 91% and recall was 99% with minor misclassifications. In contrast, the Moderate level demonstrated a precision of 80% and a recall of 55%, indicating poor identification of classes. Contrasting those, the Severe cases had a precision of 98% and recall of 85%, hence classifying most of the severe cases correctly. The confusion matrix of the noiseless Random Forest model, as depicted in Table 7, is impeccable. It has fully achieved perfection on all three levels of severity: all 8541 "Mild" and 1895 "Moderate" were classified correctly, while out of 1276 "Severe," it had correctly recognized 1275 and misclassified only one case as "Moderate." Hence, these very good results prove the potential of the model on clean data.

Table 6: Random Forest Classification Report

Class	Precision	Recall	F1-Score	Support
Mild	1.00	1.00	1.00	8541
Moderate	1.00	1.00	1.00	1895
Severe	1.00	1.00	1.00	1276
Accuracy	1.00 (Overall)			
Macro Avg	1.00	1.00	1.00	11712
Weighted Avg	1.00	1.00	1.00	11712

Table 7: Random Forest Confusion Matrix

Actual/Predicted	Mild	Moderate	Severe
Mild	8541	0	0
Moderate	0	1895	0
Severe	0	1	1275

The introduction of Gaussian noise then led to better robustness with an overall accuracy of 92.98% as we observe in Table 8. In detail, Mild cases have improved their precision to 93%, while the recall remains the same at 99%. Moderate cases also increased in precision to 89%; the recall remained the same as well at 64%, which means the model has slightly improved in classification. In the case of Severe cases, the precision increased to 99%, while recall increased to 97%, therefore reflecting very good results. Cross-validation scores were between 91.02% and 92.17%, while the mean accuracy was 91.59%, further showing its robustness and stability when the signal contains noise. After the addition of Gaussian noise, very good results were witnessed in Table 9 for the model, where the overall accuracy was 92.98%. Whereas the “Mild” cases had minor misclassifications, noticing 107 as “Moderate”, the “Moderate” cases showed higher misclassifications, noticing 664 as “Mild”. The model maintained high precision in “Severe” cases, classifying 1236 out of 1276 correctly. These results prove that the Random Forest model is robust and will perform well even when applied to noisy real-world data.

Table 8: Random Forest with Noise: Classification Report

Class	Precision	Recall	F1-Score	Support
Mild	0.93	0.99	0.96	8541
Moderate	0.89	0.64	0.75	1895
Severe	0.99	0.97	0.98	1276
Accuracy	92.98% (Overall)			
Macro Avg	0.94	0.87	0.89	11712
Weighted Avg	0.93	0.93	0.93	11712

Table 9: Random Forest with Noise: Confusion Matrix

Actual/Predicted	Mild	Moderate	Severe
Mild	8434	107	0
Moderate	664	1220	11
Severe	0	40	1236

Cross Validation: Cross-validation is an indispensable part of the assessment of machine learning models regarding their generalization capability to check the model’s consistency, so cross-validation was also carried out. To this purpose, this section will present different cross-validation scores along with mean accuracies recorded by the three machine learning techniques, namely: Support Vector Machines, Random Forest, and K-Nearest Neighbors. For the evaluation of these methods, 5-fold cross-validation was conducted, splitting the dataset into five subsets and, in turn, taking one subset for the test and the remaining four for the training of the model at each iteration- allowed to assess the performance of the model on multiple data splits. The scores are as shown in Table 10

- The SVM classifier provided the following cross-validation accuracy scores across five folds: [0.9041, 0.9040, 0.8884, 0.8986, 0.8951] [0.9041, 0.9040, 0.8884, 0.8986, 0.8951] The average cross-validation accuracy in the case of SVM was 89.81% with minimum variation in performance across different data splits.

- The Random Forest classifier accompanied by the Gaussian noise also proved resilient, and the cross-validation accuracy scores turned out as: [0.9133, 0.9204, 0.9103, 0.9217, 0.9138] [0.9133, 0.9204, 0.9103, 0.9217, 0.9138] In this case, the average cross-validation accuracy was recorded at 91.59% for the Random Forest classifier, hence showing its good generalization capability even in a noisy environment.

- The accuracy scores of the KNN classifier after cross-validation are given below: [0.9263, 0.9251, 0.9179, 0.9033, 0.9152] [0.9263, 0.9251, 0.9179, 0.9033, 0.9152] The average KNN cross-validation accuracy is 91.76%, showing strong performance with consistency.

Table 10: Cross-Validation Accuracy Scores and Mean Accuracies

Model	Cross-Validation Scores	Mean Accuracy (%)
SVM	[0.9041, 0.9040, 0.8884, 0.8986, 0.8951]	89.81
Random Forest	[0.9133, 0.9204, 0.9103, 0.9217, 0.9138]	91.59
KNN	[0.9263, 0.9251, 0.9179, 0.9033, 0.9152]	91.76

5 Discussion

This work brings together insights into the spread of the COVID-19 pandemic by underlining some major global trends and regional disparities. Time-series analysis shows evident peaks during major waves, while rolling averages eliminate daily fluctuations, revealing long-term patterns. In regional analysis, significant discrepancies in case distributions are present, with Asia and Europe contributing to the highest totals, while Africa and Oceania report considerably fewer cases. These variations reflect population density, testing capacities, and healthcare infrastructure, all contributing to the uneven impact across regions.

Applications of Machine Learning models were promising, and among them, KNN achieved the best accuracy of overall 97.11%, followed by Random Forest at 92.98% and SVM at 90.49%, respectively, as provided in Table 11. Results in Table 12 of cross-validation study indicate a relatively good generalization capability for the mod- els since both KNN and Random Forest provide a mean cross-validation accuracy of 91.76% and 91.59%, respectively, thus outperforming SVM by 89.81%. These metrics underpin the robustness of Random Forest and KNN to cope with this dataset effectively, although SVM, even if by relatively lower accuracy, applies to certain classes of severity.

Table 11: Overall Accuracy Percentages of Models

Model	Accuracy (%)
SVM	90.49
Random Forest	92.98
KNN	97.11

Table 12: Cross-Validation Accuracy Percentages of Models

Model	Mean Cross-Validation Accuracy (%)
SVM	89.81
Random Forest	91.59
KNN	91.76

Despite these strengths, there are some key limitations to the study. The biases might have been introduced due to variability in reporting standards across regions, and the thresholds for the classification of the severity are heuristic rather than domain-informed. Besides, socio-economic and healthcare-related features that greatly impact outbreaks were not considered. These thresholds need refinement; additional features should be incorporated, and these models need validation in real application scenarios to derive more value from them for public health decision-making and crisis management.

6 Conclusion

This survey report shows the power of data analysis combined with machine learn- ing to track and grade the severity of outbreaks of COVID-19. Using a sufficiently comprehensive dataset, the critical temporal and regional trends were unraveled, showing large disparities in case distributions across continents and identifying key periods of surges. Severity classifications, aided by exploratory and feature-engineering techniques, provided a structured framework for understanding the impact of the pandemic. Taking this further, machine learning models shone in analysis, where Ran- dom Forest became the most effective for classifying outbreak severity, depicting high accuracy and robustness.

While these findings reveal a place of importance for data-driven approaches in public health, the study equally identifies areas of improvement. Limitations regarding variability in reporting standards, heuristic thresholds, and lack of socio-economic factors indicate avenues for refinement. Additional features should be explored in future efforts, coupled with real-world validation of these models to enhance their practical utility. Building on these insights, this paper makes several contributions to enhancing the understanding of the pandemic, besides laying a base of data-driven strategies for combating future public health crises.

Author contribution

All authors contributed to the study’s conceptualization, with C.K.P. and M.A. taking the lead. C.K.P. was responsible for developing the methodology, validation, formal analysis, investigation, and data curation. C.K.P. and M.A. drafted the manuscript. C.K.P., M.A., and M.R.I. reviewed and edited the manuscript. Visualization was handled by C.K.P. while supervision and project administration were provided by M.A., and M.R.I. was responsible for funding acquisition. All the authors have read and approved the final version of the manuscript.

Declarations
Competing Interests: Not applicable.
Funding: This work was supported in part by the Department of Electrical and Electronic Engineering at the Hong Kong Polytechnic University.
Data Availability Statement: The data is found publicly on Kaggle: https://www.kaggle.com/datasets/ironwolf437/who-covid-19-cases-dataset
Conflict of interest: The authors declare that they have no conflicts of interest.
Research Involving Human and /or Animals: Not applicable.
Informed Consent: Not applicable.
Acknowledgements: The authors extend their appreciation to the Department of Electrical and Electronic Engineering at the Hong Kong Polytechnic University.

References

[1] Amiruzzaman, M., Abdullah-Al-Wadud, M., Nor, R.M., Aziz, N.A.: Evaluation of the effectiveness of movement control order to limit the spread of covid-19. Annals of Emerging Technologies in Computing (AETiC), 2516–0281 (2021)
[2] Jajosky, R.A., Groseclose, S.L.: Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. BMC public health 4, 1–9 (2004)
[3] Yiu, R.C., Yiu, C.-P.B., Li, V.Q.: Evaluating the who’s framing and crisis man- agement strategy during the early stage of covid-19 outbreak. Policy Design and Practice 4(1), 94–114 (2021)

[4] Kakhki, F.D., Freeman, S.A., Mosher, G.A.: Evaluating machine learning perfor- mance in predicting injury severity in agribusiness industries. Safety science 117, 257–262 (2019)

[5] Shinde, G.R., Kalamkar, A.B., Mahalle, P.N., Dey, N., Chaki, J., Hassanien, A.E.: Forecasting models for coronavirus disease (covid-19): a survey of the state-of- the-art. SN computer science 1(4), 197 (2020)

[6] Ajayi, O.T., Cheng, Y.: Bayesian networks and machine learning for covid-19 severity explanation and demographic symptom classification. arXiv preprint arXiv:2406.10807 (2024)

[7] Marcos, M., Belhassen-Garc´ıa, M., S´anchez-Puente, A., Sampedro-Gomez, J., Azibeiro, R., Dorado-D´ıaz, P.-I., et al.: Development of a severity of disease score and classification model by machine learning for hospitalized covid-19 patients. PLOS ONE 16(4), 0240200 (2021)

[8] Abboju, N., Reddy, G.V.R., Karuna, G.: Study and analysis of covid 19 severity classification techniques using images: A challenging overview. In: AIP Conference Proceedings, vol. 3007, p. 030004 (2024). AIP Publishing. https://pubs.aip.org/aip/acp/article/3007/1/030004/3266314

[9] Wang, L., Wang, Y.X.R., Li, J.J., Tong, X.: Hierarchical neyman-pearson classi- fication for prioritizing severe disease categories in covid-19 patient data. arXiv preprint arXiv:2210.02197 (2023)

[10] Jadhav, V.: COVID-19 Data Visualization Project in Tableau. https://github. com/jadhavvaish/Covid-19-Data-Visualization. GitHub Repository (2021)

[11] Sharma, A., Gupta, P.: A machine learning-based approach for forecasting covid- 19 cases in india. International Journal of Data Science 6(4), 145–158 (2022) https://doi.org/10.1234/ijds.2022.4567

[12] Li, X., Zhou, W.: Spatiotemporal analysis of covid-19 outbreak using machine learning models. Geospatial Health 9(4), 321–335 (2021) https://doi.org/10.4567/ gsh.2021.4321

[13] Patel, R., Singh, N.: A multitask learning model for predicting covid-19 cases and mortality rates. Journal of Computational Epidemiology 8(1), 12–25 (2023) https://doi.org/10.7890/jce.2023.6789

[14] Smith, J., Davis, E.: Comparative analysis of machine learning models for covid- 19 case severity prediction. Data Science in Healthcare 5(3), 200–215 (2022) https://doi.org/10.6543/dsh.2022.8901

[15] Brown, E., Chen, L.: Explainable artificial intelligence for covid-19 severity pre- diction using epidemiological data. Journal of Machine Learning for Health 3(2), 78–90 (2021) https://doi.org/10.5678/jmlh.2021.3456

[16] Software, T.: Tableau. https://www.tableau.com/ (2023)