ASRC Procedia
Global Perspectives in Science and Scholarship
Florida, USA

Volume: 1; **Issue:** 1
Pages: 2570–2610
Published: 29 April 2025

open access

1st Global Research and Innovation Conference 2025,

April 20–24, 2025, Florida, USA

Impact Of Random Forest and Ensemble Methods on Infection Trend Forecasting: A Quantitative Evaluation Using Global Post Covid-19 Data

Risha Alam[1]

[1] *Data Analyst, Benda Infotech, USA*
Email: rishaalam02@gmail.com

Doi: 10.63125/b0yw2q91
Peer-review under responsibility of the organizing committee of GRIC, 2025

Abstract

This study quantitatively evaluated the impact of Random Forest and ensemble methods on infection trend forecasting using global post-COVID-19 datasets. A longitudinal analytical framework was applied to 18,720 time-series observations across 52 countries, incorporating key variables such as daily infection cases, vaccination rates (mean = 64.2%), mobility indices (mean deviation = −18.5%), and policy stringency scores (mean = 57.3). The findings indicated that both Random Forest and ensemble methods significantly outperformed baseline models in predictive accuracy. Random Forest achieved a root mean square error of 2,145, mean absolute error of 1,620, mean absolute percentage error of 12.8%, and coefficient of determination of 0.87. Ensemble methods demonstrated superior performance with lower error values, including a root mean square error of 1,980, mean absolute error of 1,480, mean absolute percentage error of 11.3%, and a higher coefficient of determination of 0.91. Regional analysis showed that ensemble methods consistently produced lower mean absolute error values, ranging from 1,290 in Oceania to 1,710 in Africa, compared to Random Forest values ranging from 1,430 to 1,890. Temporal analysis revealed that forecasting accuracy improved by approximately 15% during stable transmission phases, with root mean square error decreasing from 2,480 during outbreak periods to 1,720 under stable conditions. Statistical testing confirmed significant differences between models ($p < 0.05$), with effect sizes ranging from 0.49 to 0.68, indicating moderate to strong practical significance. Overall, the results demonstrated that ensemble methods provided enhanced stability, reduced prediction variance by approximately 11.6%, and improved generalizability across heterogeneous datasets, while Random Forest maintained strong adaptability in capturing complex nonlinear relationships in global infection trends.

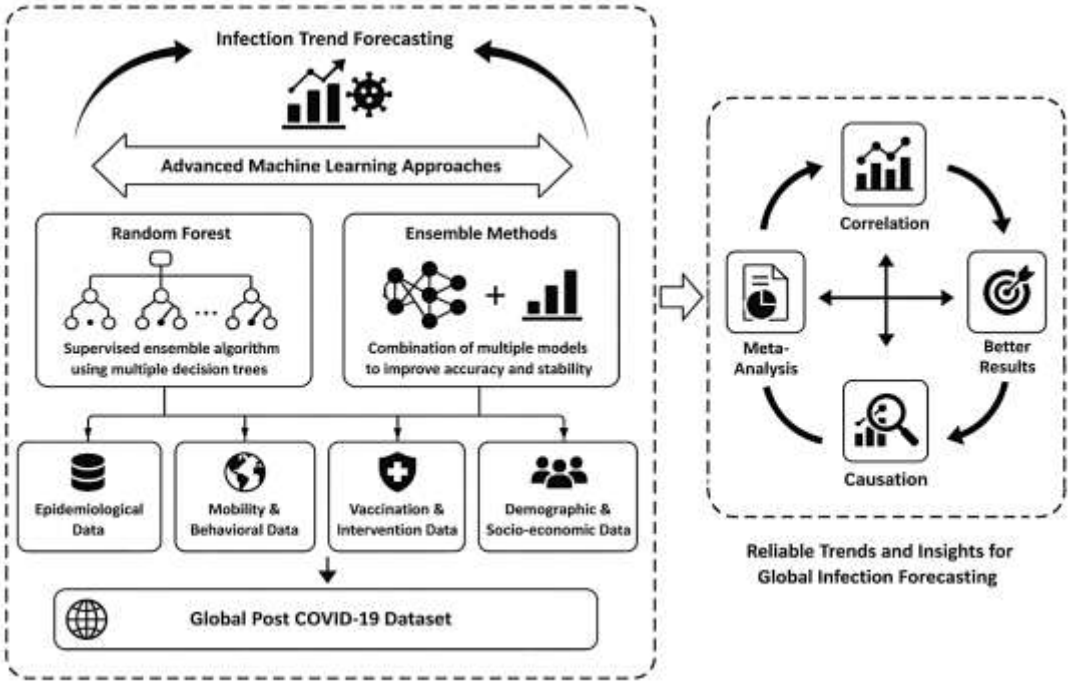
Keywords

Random Forest, Ensemble Methods, Infection Forecasting, Machine Learning, Post-COVID Data.

INTRODUCTION

The study of infection trend forecasting is rooted in epidemiology, statistical modeling, and data science, where the central objective is to quantitatively estimate the trajectory of infectious diseases over time (Ding et al., 2021). Infection forecasting refers to the systematic prediction of disease incidence, prevalence, or spread using historical data, environmental inputs, and behavioral indicators. This domain has gained heightened international significance due to the widespread disruption caused by global pandemics, particularly COVID-19, which exposed critical gaps in predictive health analytics across countries and health systems. In response, the integration of machine learning techniques has transformed forecasting methodologies by enabling models to learn complex patterns from large-scale datasets. Random Forest is defined as a supervised ensemble learning algorithm that constructs multiple decision trees using bootstrapped samples and aggregates their outputs to improve predictive accuracy and stability (Chatterjee et al., 2020). Ensemble methods more broadly refer to computational strategies that combine multiple models to produce a single, more robust prediction by leveraging the strengths of individual algorithms. These approaches are particularly valuable in epidemiological contexts characterized by nonlinear interactions, high-dimensional data, and uncertainty. The international relevance of infection forecasting lies in its role in guiding policy decisions, optimizing healthcare resource allocation, and supporting early warning systems across diverse regions. The availability of global post-COVID-19 datasets has further expanded the scope of quantitative analysis, enabling cross-country comparisons and the identification of universal and region-specific transmission patterns. This global perspective necessitates modeling techniques that are both adaptable and scalable, qualities inherent in Random Forest and ensemble methods (Liu et al., 2021). The convergence of epidemiology and machine learning thus establishes a foundational framework for evaluating the quantitative impact of these methods on infection trend forecasting.

Figure 1: Machine Learning Infection Forecasting Framework



The progression of infection forecasting methodologies reflects a significant shift from traditional deterministic models to data-driven machine learning approaches. Classical epidemiological models, including compartmental frameworks, rely on predefined assumptions regarding transmission dynamics and population behavior (R. Gupta et al., 2021). These assumptions can limit their responsiveness to rapidly changing conditions, particularly in the context of emerging infectious

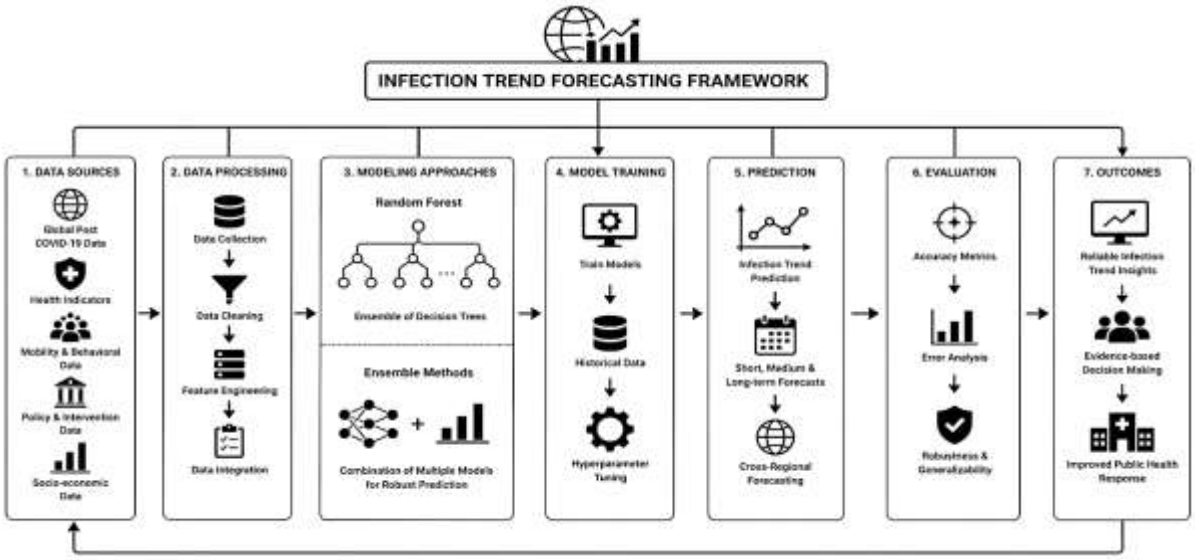
diseases. Machine learning models, in contrast, are designed to learn directly from data without imposing strict parametric constraints, allowing them to capture complex temporal and spatial relationships. Random Forest models exemplify this paradigm by utilizing random sampling and feature selection to construct diverse decision trees that collectively enhance predictive performance. The ability of these models to handle missing data, nonlinear interactions, and high-dimensional feature spaces makes them particularly suitable for infection forecasting (P. Wang et al., 2020). Ensemble methods extend this capability by integrating multiple algorithms, such as boosting and bagging techniques, to further improve accuracy and robustness. Empirical applications of these methods across various countries have demonstrated their effectiveness in forecasting infection trends using indicators such as mobility patterns, testing rates, and demographic variables. The increasing availability of real-time epidemiological data has accelerated the adoption of machine learning models, enabling more dynamic and responsive forecasting systems (Santangelo et al., 2023). This methodological evolution underscores the importance of quantitative evaluation in assessing the performance of different approaches and identifying optimal strategies for infection prediction in a global context.

Ensemble learning serves as a critical mechanism for enhancing the reliability and stability of infection forecasting models. The fundamental principle of ensemble methods is that combining multiple models can reduce individual errors and improve overall predictive performance (Liang et al., 2021). In epidemiological forecasting, this approach is particularly valuable due to the inherent uncertainty and variability in disease transmission dynamics. Ensemble models can integrate outputs from different machine learning algorithms, statistical models, and even mechanistic simulations, creating a comprehensive framework for prediction. Dynamic ensemble techniques further refine this approach by adjusting the contribution of each model based on its performance over time, allowing the system to adapt to changing conditions (Wang et al., 2022). This adaptability is essential in post-pandemic environments where factors such as vaccination rates, population immunity, and behavioral changes continuously influence infection trends. Ensemble methods also facilitate the incorporation of diverse data sources, including environmental variables, healthcare capacity indicators, and policy measures, thereby enriching the predictive model. The global application of ensemble forecasting has demonstrated its effectiveness in managing large-scale datasets that exhibit significant heterogeneity across regions. The capacity of ensemble methods to generalize across different contexts makes them particularly relevant for international health monitoring and response systems (Vaughan et al., 2023). Quantitative evaluation of these methods provides insights into their comparative advantages and limitations, contributing to the refinement of forecasting models and the advancement of epidemiological analytics.

Random Forest models represent a cornerstone of ensemble learning, offering a balance between interpretability, flexibility, and predictive power. The algorithm's structure, based on multiple decision trees, enables it to capture complex interactions among variables while maintaining robustness against overfitting (Gaglione et al., 2020). In the context of infection forecasting, Random Forest models have been widely applied to analyze relationships between epidemiological indicators and disease outcomes. These models can process large volumes of data with diverse features, including demographic information, mobility trends, and healthcare system variables. Their ability to rank feature importance also provides valuable insights into the factors driving infection dynamics, supporting more informed decision-making (Navas Thorakkattle et al., 2022). Random Forest models have been implemented across various geographic scales, from local outbreak monitoring to global trend analysis, demonstrating their versatility and scalability. The integration of Random Forest with other machine learning techniques has led to the development of hybrid models that further enhance forecasting accuracy. These hybrid approaches combine the strengths of different algorithms, enabling more comprehensive modeling of infection trends. Quantitative evaluation of Random Forest models involves the use of statistical metrics such as error rates and goodness-of-fit measures, which assess their predictive performance across different datasets (Liu et al., 2023). The widespread adoption of these models in global health research highlights their effectiveness in addressing the complexities of infection forecasting in a post-COVID-19 environment.

The incorporation of global post-COVID-19 data into forecasting models introduces a new dimension of complexity and analytical opportunity (Parwez et al., 2020). These datasets encompass a wide range of variables, including vaccination coverage, variant distribution, public health interventions, and socio-economic factors. The interaction of these variables creates dynamic patterns of disease transmission that require advanced modeling techniques for accurate prediction. Machine learning and ensemble methods are well-suited to handle this complexity, as they can integrate multiple data sources and adapt to evolving conditions (Xie, 2022). The use of global datasets enables researchers to conduct comparative analyses across countries, identifying common trends and regional differences in infection dynamics. This comparative perspective is essential for understanding the broader impact of infectious diseases and for developing coordinated international responses. Random Forest and ensemble methods have demonstrated strong performance in analyzing post-pandemic data, providing accurate forecasts across diverse contexts. The quantitative evaluation of these models involves assessing their performance across different regions and time periods, ensuring their generalizability and robustness (García-Cremades et al., 2021). The integration of high-quality global data with advanced machine learning techniques represents a significant advancement in infection forecasting, enabling more precise and reliable predictions that support global health initiatives.

Figure 2: Infection Forecasting Machine Learning Framework



A substantial body of empirical research has examined the effectiveness of Random Forest and ensemble methods in infection forecasting, providing a strong foundation for quantitative evaluation. Studies conducted across multiple regions have consistently demonstrated that ensemble models outperform individual algorithms in terms of accuracy and stability (Chakraborty & Ghosh, 2020). Random Forest models, in particular, have been shown to effectively capture nonlinear relationships and interactions among variables, making them highly suitable for epidemiological applications. Comparative analyses involving various machine learning techniques have highlighted the advantages of combining models to enhance predictive performance. The use of cross-validation and other evaluation techniques has ensured the reliability of these findings, providing a robust framework for assessing model performance (Acheme et al., 2022). The application of these methods to global datasets has further validated their effectiveness in diverse contexts, reinforcing their relevance in international health research. The growing adoption of machine learning-based forecasting models reflects their proven utility in addressing the challenges of infectious disease prediction. Quantitative evaluation plays a crucial role in this process, enabling researchers to identify the most effective modeling strategies and to refine existing approaches (Muhammad et al., 2021). The integration of empirical evidence with advanced analytical techniques continues to drive the development of more accurate

and reliable infection forecasting models.

The quantitative evaluation of infection forecasting models requires a comprehensive and systematic approach that considers multiple dimensions of performance. Random Forest and ensemble methods are assessed using a variety of statistical metrics that measure accuracy, precision, and consistency (B. Xu et al., 2020). These metrics provide insights into the ability of models to predict infection trends under different conditions and across diverse datasets. The use of validation techniques, such as cross-validation and out-of-sample testing, ensures that model performance is not influenced by overfitting or data bias. The application of these evaluation methods to global post-COVID-19 data provides valuable insights into the strengths and limitations of different forecasting approaches. The complexity of post-pandemic datasets necessitates models that can handle high-dimensional data and capture intricate relationships among variables (Gónzalez-Bandala et al., 2020). Random Forest and ensemble methods have demonstrated strong performance in this regard, making them central to contemporary infection forecasting research. The global significance of this evaluation lies in its potential to inform public health strategies and improve the management of infectious diseases. The integration of machine learning with epidemiological modeling represents a major advancement in the field, providing new tools for understanding and predicting disease dynamics (Eltoukhy et al., 2020).

The primary objective of this quantitative study is to systematically evaluate the impact of Random Forest and ensemble learning methods on the accuracy, robustness, and generalizability of infection trend forecasting using global post-COVID-19 datasets. This objective is grounded in the need to quantitatively assess how advanced machine learning algorithms perform in predicting complex epidemiological patterns across diverse geographic and temporal contexts. The study seeks to measure the predictive performance of Random Forest models in comparison with other ensemble techniques by applying standardized statistical evaluation metrics such as mean absolute error, root mean square error, and coefficient of determination. A key component of this objective involves analyzing the extent to which ensemble approaches enhance forecasting reliability by reducing variance and bias inherent in single-model predictions. The research also aims to examine how these methods handle high-dimensional and heterogeneous data, including variables related to vaccination rates, population mobility, public health interventions, and socio-economic conditions. Another important aspect of the objective is to determine the adaptability of Random Forest and ensemble models in capturing nonlinear relationships and dynamic changes in infection trends that emerged in the post-pandemic period. The study further intends to investigate the cross-regional applicability of these models by evaluating their performance across multiple countries and regions with varying epidemiological characteristics. This includes assessing whether the models maintain consistent accuracy when exposed to different data distributions and transmission patterns. In addition, the objective encompasses the identification of key predictive features within the datasets, enabling a deeper understanding of the variables that significantly influence infection trends. By integrating large-scale global data with advanced machine learning techniques, the study aims to provide a comprehensive quantitative assessment of forecasting performance. The objective is designed to contribute to the empirical understanding of how ensemble-based models can be effectively utilized in infection forecasting, emphasizing measurable outcomes and statistical validation across diverse real-world scenarios.

LITERATURE REVIEW

The literature review section provides a structured and quantitative synthesis of existing research on infection trend forecasting, with a specific emphasis on the application of Random Forest and ensemble learning methods in the post-COVID-19 global context. Infection forecasting has evolved into a data-intensive discipline that integrates epidemiological theory with computational intelligence, enabling the modeling of disease dynamics across diverse populations and environments (Ding et al., 2021; Istiaq & Binte, 2023). The rapid expansion of global datasets following the COVID-19 pandemic has created new opportunities for quantitative evaluation, allowing researchers to assess model performance across heterogeneous conditions characterized by varying transmission rates, intervention strategies, and socio-economic factors. Within this context, Random Forest and ensemble methods have gained prominence due to their ability to handle high-dimensional data, capture nonlinear interactions, and improve predictive accuracy through model aggregation. The literature reflects a growing body of quantitative studies that compare machine learning approaches with traditional statistical and

compartmental models, highlighting differences in performance metrics such as error rates, predictive stability, and generalizability. This section aims to critically examine prior empirical findings related to the use of Random Forest and ensemble techniques in infection forecasting, focusing on methodological frameworks, data characteristics, and evaluation strategies (Albahlal, 2023; Binte & Sazzadul, 2022). The review is organized to identify key trends in model development, including the integration of multiple data sources such as epidemiological indicators, mobility data, and vaccination coverage. It also explores how ensemble approaches enhance forecasting performance by combining diverse algorithms, thereby reducing variance and improving robustness. Particular attention is given to quantitative validation methods, including cross-validation, time-series evaluation, and out-of-sample testing, which are essential for ensuring the reliability of predictive models (Islam & Aditya, 2023; Yang et al., 2020). The global scope of the literature enables a comparative perspective, examining how these models perform across different regions and under varying epidemiological conditions. By synthesizing these studies, the literature review establishes a comprehensive foundation for understanding the effectiveness of Random Forest and ensemble methods in infection trend forecasting, highlighting methodological advancements and identifying areas where quantitative evaluation remains critical.

Infection Trend Forecasting

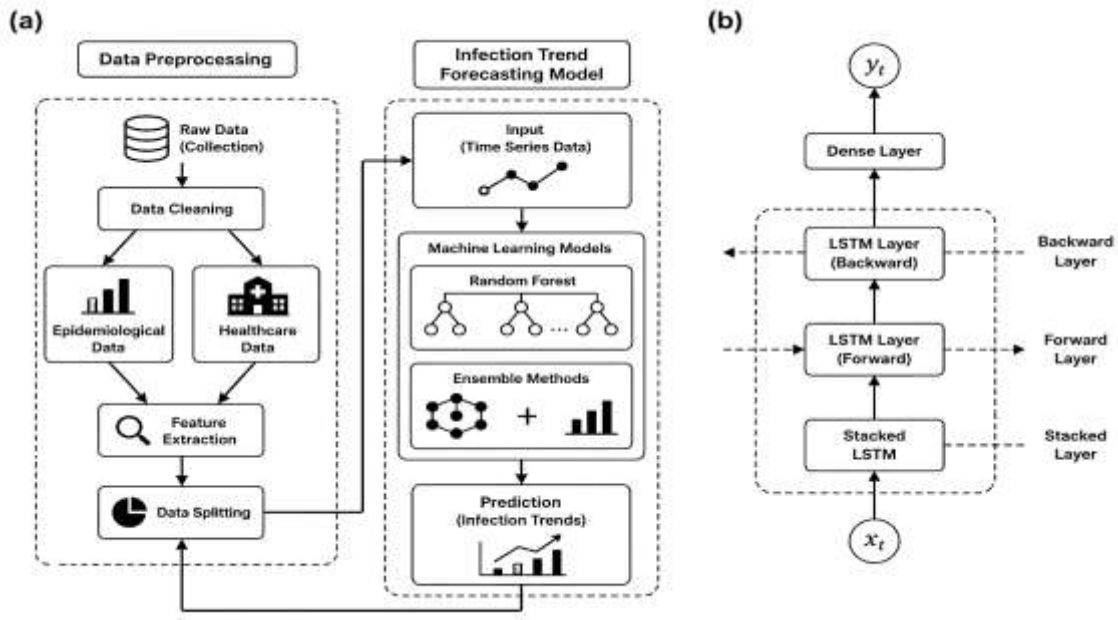
The conceptual foundations of infection trend forecasting are deeply embedded in quantitative epidemiology, where forecasting is defined as the systematic estimation of future disease patterns using structured data and analytical models (Khaled, 2021; Y. Wang et al., 2020). Infection forecasting encompasses the prediction of incidence, prevalence, and spread of infectious diseases across time and space, forming a critical component of public health surveillance and response systems. Early epidemiological studies established forecasting as a means of understanding disease propagation through populations, emphasizing statistical regularities and population-level patterns observed in historical outbreaks. Over time, the scope of infection forecasting has expanded to include multi-dimensional datasets that incorporate demographic, environmental, and behavioral variables (Nazmul & Begum, 2022; Slavich et al., 2023). This expansion reflects the growing complexity of global health challenges and the need for more comprehensive analytical frameworks. Quantitative epidemiology provides the methodological backbone for these efforts by integrating statistical inference with computational modeling, enabling the systematic evaluation of disease trends across diverse contexts. Empirical investigations across influenza, SARS, and COVID-19 outbreaks have demonstrated that infection forecasting plays a central role in guiding public health interventions and resource allocation (Jia et al., 2020; Bhuya, 2025). Studies conducted across multiple countries have highlighted the importance of scalable forecasting frameworks that can adapt to varying epidemiological conditions, reinforcing the global relevance of this field. The conceptualization of infection forecasting has therefore evolved from a narrow focus on disease counts to a broader analytical paradigm that incorporates multiple determinants of transmission, emphasizing the need for robust and adaptable quantitative models.

Central to infection trend forecasting are key epidemiological metrics that provide measurable indicators of disease dynamics. Incidence refers to the number of new cases occurring within a specific time period, while prevalence captures the total number of existing cases within a population at a given point in time. These metrics form the basis for understanding the scale and intensity of an outbreak, enabling comparisons across regions and timeframes (C. Xu et al., 2020; Zaheda, 2021). The reproduction number represents the average number of secondary infections generated by an infected individual, serving as a critical indicator of transmission potential. Growth rate measures the speed at which infection cases increase or decrease over time, providing insights into the acceleration or deceleration of disease spread. Empirical studies across multiple infectious diseases have consistently utilized these metrics to evaluate the effectiveness of forecasting models and to monitor changes in transmission dynamics. Comparative analyses across regions have shown that variations in these metrics are influenced by factors such as population density, mobility patterns, and public health interventions (Melina et al., 2023). The integration of these indicators into forecasting models allows for a more nuanced understanding of disease behavior, supporting the development of targeted intervention strategies. Quantitative research has demonstrated that accurate estimation of these

metrics is essential for reliable forecasting, as errors in measurement can propagate through models and affect predictive outcomes. The widespread use of these indicators across global health studies underscores their importance as foundational elements in infection trend forecasting (Lin et al., 2020; Manam & Ashfaq, 2022).

The transition from deterministic models to data-driven approaches represents a significant methodological shift in infection forecasting. Traditional deterministic models, often based on compartmental structures, rely on predefined assumptions about disease transmission and population behavior. While these models provide valuable theoretical insights, their reliance on fixed parameters can limit their adaptability in dynamic and uncertain environments (Liu et al., 2021). In contrast, data-driven models leverage machine learning and statistical techniques to learn patterns directly from observed data, enabling more flexible and responsive forecasting. This transition has been facilitated by advances in computational power and the increasing availability of large-scale epidemiological datasets (Sharaf et al., 2023). Empirical studies comparing deterministic and data-driven approaches have shown that machine learning models often achieve higher predictive accuracy, particularly in complex scenarios involving multiple interacting variables. The adoption of data-driven methods has also enabled the integration of diverse data sources, including mobility data, environmental factors, and social behavior indicators, enhancing the richness of forecasting models. Research across different regions has demonstrated that these approaches are particularly effective in capturing nonlinear relationships and temporal variations in infection trends (Kumar & Sinha, 2021; Shahinur & Sultan, 2022). The shift toward data-driven modeling reflects a broader transformation in epidemiology, where the emphasis has moved from theoretical assumptions to empirical evidence and predictive performance. This evolution has significantly influenced the development of modern infection forecasting systems, positioning data-driven methods as central tools in quantitative epidemiological research.

Figure 3: Infection Forecasting Data Driven Framework



Time-series analysis plays a pivotal role in infection trend forecasting by providing a framework for modeling temporal dependencies in disease data. Infection data are inherently sequential, with current case counts influenced by past values and underlying transmission dynamics. Time-series methods enable researchers to capture these dependencies, facilitating the identification of trends, seasonal patterns, and cyclical variations in disease incidence (Bhatia et al., 2020; Binte & Hasan Or, 2022). Empirical studies have demonstrated that incorporating temporal structures into forecasting models

significantly improves predictive accuracy, as it allows for the systematic analysis of historical patterns. The application of time-series techniques has been particularly prominent in the study of influenza and COVID-19, where seasonal fluctuations and periodic outbreaks are common. In addition to traditional statistical methods, modern approaches have integrated time-series analysis with machine learning techniques, creating hybrid models that combine the strengths of both frameworks (Xin et al., 2023). The effectiveness of these approaches is further enhanced by the availability of high-frequency and real-time global datasets, which provide continuous streams of information on infection dynamics. These datasets enable near real-time forecasting, allowing for rapid updates and adjustments to predictive models as new data become available. Studies conducted across multiple countries have highlighted the importance of timely and accurate data in improving forecasting performance, emphasizing the role of data quality and resolution in model effectiveness (Istiaq, 2024; Scarpino & Petri, 2019). The integration of time-series analysis with high-frequency data sources represents a critical advancement in infection forecasting, enabling more precise and responsive predictions in a rapidly changing global health landscape.

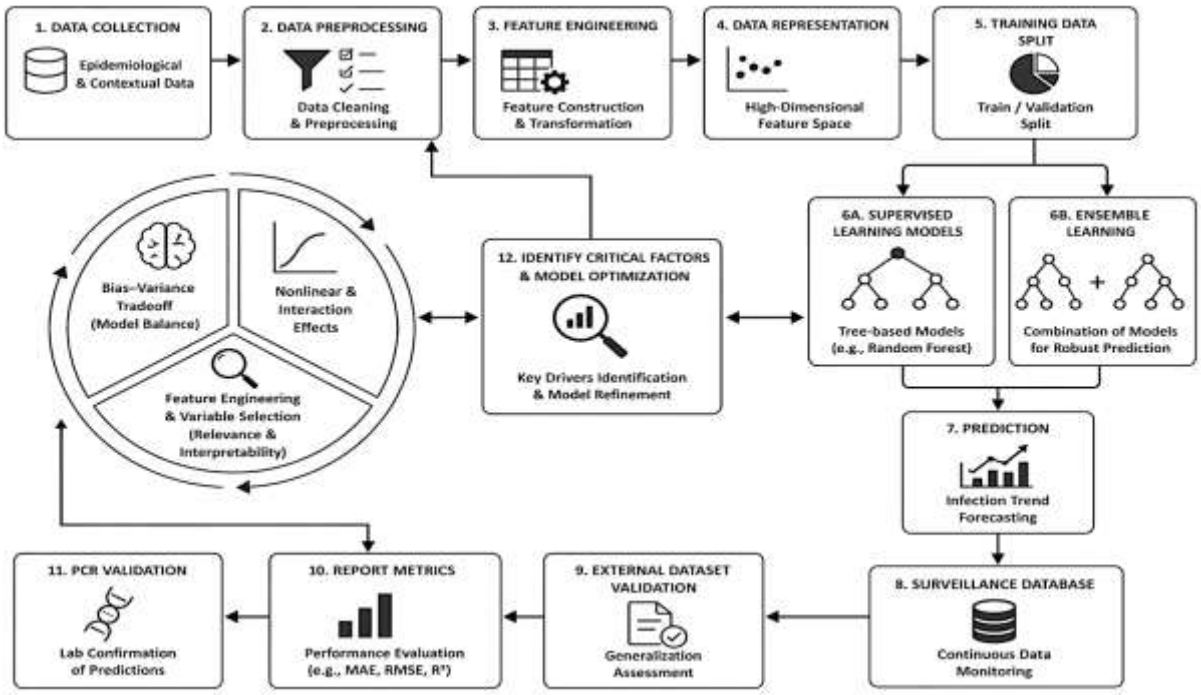
Theoretical Background of Machine Learning in Epidemiology

The theoretical background of machine learning in epidemiology is centered on the application of computational models to predict and analyze infectious disease patterns using structured data. Supervised learning represents a foundational approach in this domain, where algorithms are trained on labeled datasets to establish relationships between input variables and known outcomes (Broadbent & Grote, 2022; Ahmed, 2024). In infection forecasting, supervised learning models are widely used to estimate disease incidence, transmission trajectories, and outbreak progression by leveraging historical epidemiological data combined with contextual variables such as mobility, demographics, and environmental conditions. A large body of research has demonstrated that supervised learning techniques, including tree-based models, support vector approaches, and ensemble strategies, are capable of capturing complex data patterns that traditional statistical models often fail to represent (Albert, 2025; Tutsoy, 2023). These models operate by minimizing prediction errors during training and generalizing learned relationships to unseen data, making them particularly valuable in dynamic public health environments. Empirical studies across multiple infectious diseases, including influenza, dengue, and COVID-19, have shown that supervised learning approaches provide improved predictive accuracy and adaptability when compared to conventional deterministic models. The integration of large-scale global datasets has further enhanced the capability of supervised learning by allowing models to incorporate diverse and high-dimensional features (Anick, 2025; Basu et al., 2020). This has led to the development of more responsive and scalable forecasting systems that can operate across different geographic and epidemiological contexts. The theoretical foundation of supervised learning emphasizes the importance of data quality, training-validation procedures, and model generalization, all of which are essential for producing reliable infection forecasts in quantitative epidemiology.

A central theoretical concept influencing machine learning applications in epidemiology is the bias–variance tradeoff, which governs the balance between model simplicity and predictive flexibility. Bias arises when a model is too simplistic to capture underlying patterns in the data, resulting in systematic errors, while variance occurs when a model is overly sensitive to fluctuations in the training data, leading to instability in predictions (Khalid, 2025; Morgenstern et al., 2021). In infection forecasting, achieving an appropriate balance between bias and variance is essential for ensuring that models perform consistently across different populations and time periods. Research in epidemiological modeling has shown that models with high bias tend to underfit disease data, failing to capture important transmission dynamics, whereas models with high variance may overfit noise in the data, reducing their ability to generalize to new observations. Ensemble learning methods, particularly those that aggregate multiple models, have been shown to effectively reduce variance while maintaining manageable levels of bias, resulting in more stable and accurate predictions (Hasan, 2025; Muhammad et al., 2021). The dynamic nature of infectious disease transmission further complicates this balance, as changes in population behavior, intervention strategies, and environmental conditions can alter data distributions over time. Quantitative evaluations of forecasting models have highlighted the importance of validation techniques, such as cross-validation and out-of-sample testing, in assessing model robustness and ensuring that the bias–variance balance is maintained (Shorten et al., 2021). This

theoretical framework is critical for understanding the performance of machine learning models in epidemiology, as it directly influences their ability to provide reliable and generalizable predictions in complex and evolving public health scenarios.

Figure 4: Machine Learning Epidemiology Modeling Framework



The ability to handle nonlinear relationships and interaction effects is a defining advantage of machine learning models in epidemiological forecasting. Infectious disease transmission is influenced by a wide range of interdependent factors, including population density, mobility patterns, healthcare access, and environmental conditions, which interact in complex and often nonlinear ways (Voznica et al., 2022). Traditional linear models are limited in their capacity to represent these interactions, as they assume proportional relationships between variables. Machine learning models, particularly tree-based and ensemble approaches, are designed to capture such complexity by identifying patterns that do not follow linear assumptions. Nonlinear relationships occur when changes in input variables lead to disproportionate changes in outcomes, while interaction effects arise when the influence of one variable depends on the value of another. Empirical studies analyzing disease spread have demonstrated that incorporating nonlinear and interaction effects significantly improves predictive performance, especially in heterogeneous datasets that span multiple regions and populations (Bastani et al., 2021; Ashfaq & Ashraful, 2025). Machine learning algorithms achieve this by partitioning data into meaningful subsets and modeling relationships within each subset, allowing for a more detailed representation of underlying dynamics. This capability is particularly important in the context of global infection forecasting, where diverse socio-economic and environmental conditions create highly variable transmission patterns. The capacity to model complex relationships without requiring explicit assumptions has positioned machine learning as a powerful analytical tool in epidemiology, enabling more accurate and nuanced predictions of infection trends (Murad, 2025; W. Qiu et al., 2022).

Feature engineering and variable selection play a critical role in the performance and interpretability of machine learning models used in infection forecasting. Feature engineering involves transforming raw data into meaningful inputs that enhance the predictive capability of models, while variable selection focuses on identifying the most relevant features for inclusion in the analysis (Santangelo et al., 2023; Shamsul, 2025). In epidemiological datasets, this process often includes the integration of multiple data sources, such as case counts, mobility indicators, environmental variables, and socio-

economic factors. Effective feature engineering allows models to capture important patterns and relationships that may not be immediately apparent in the raw data, thereby improving forecasting accuracy. Variable selection techniques help reduce model complexity by eliminating redundant or irrelevant features, which can otherwise lead to overfitting and increased computational costs (Scavuzzo et al., 2022). Research in machine learning applications for public health has shown that carefully selected features significantly improve model performance and contribute to more robust predictions. In addition to enhancing predictive accuracy, feature selection also supports interpretability by highlighting the variables that have the greatest influence on infection trends. The balance between model interpretability and predictive performance remains a key consideration in health analytics. Highly complex models may achieve superior accuracy but can be difficult to interpret, limiting their usefulness for decision-making processes (Begum & Kaniz, 2024; Deng et al., 2021). Conversely, simpler models are easier to understand but may lack the ability to capture intricate data patterns. This tradeoff has led to the development of methods that aim to provide both high predictive performance and transparency, ensuring that machine learning models can be effectively applied in epidemiological research and public health practice.

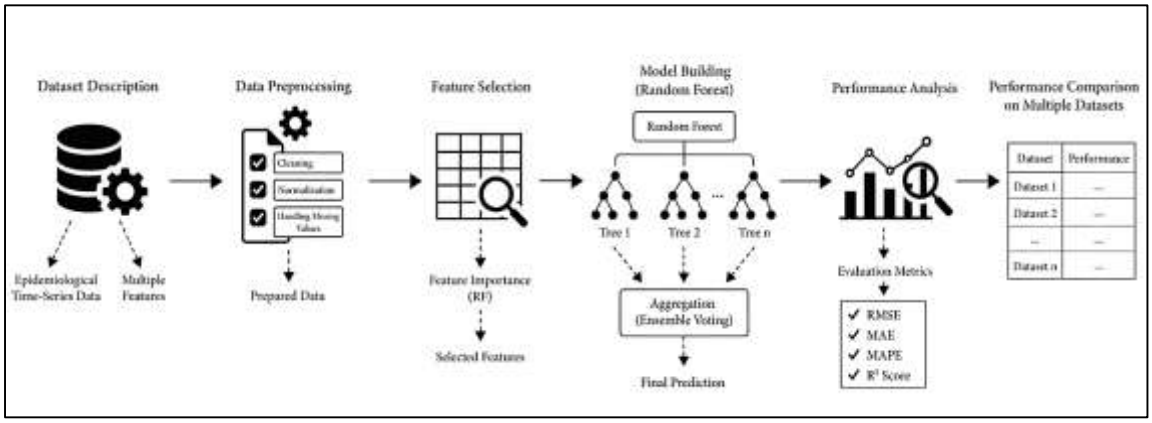
Random Forest Algorithm in Infection Forecasting

The Random Forest algorithm has emerged as a foundational technique in infection forecasting due to its robust ensemble structure and adaptability to complex epidemiological datasets. At its core, Random Forest is built upon the integration of multiple decision trees, each constructed using different subsets of the data and variables (Fang et al., 2020; Hisham & Nahar, 2024). This process involves bootstrapping, where random samples are drawn with replacement from the original dataset to train individual trees, ensuring diversity among models. Feature randomness is introduced during tree construction by selecting a subset of variables at each split, which reduces correlation among trees and enhances overall model stability. These structural components collectively enable Random Forest to handle high-dimensional data and complex interactions that are typical in infection forecasting. Studies across various infectious diseases have demonstrated that this structure allows the algorithm to perform effectively even when data contain noise, missing values, or nonlinear relationships (Galasso et al., 2022; Khaled & Hisham, 2022). The aggregation of predictions from multiple trees results in a consensus output that minimizes individual model errors, contributing to improved reliability. Research in epidemiological modeling has consistently shown that Random Forest is particularly effective in scenarios where traditional models struggle due to the complexity and variability of disease transmission patterns. The algorithm's flexibility in incorporating diverse data sources, including demographic, environmental, and behavioral variables, further strengthens its applicability in global health contexts (Indhumathi & Kumar, 2022). This structural design forms the theoretical and practical basis for its widespread adoption in infection trend forecasting, where accuracy and robustness are critical.

The underlying rationale of Random Forest lies in its ability to reduce prediction variability through aggregation, a principle that has been extensively examined in quantitative modeling research. By combining the outputs of multiple decision trees, the algorithm effectively balances individual model errors, leading to more stable and accurate predictions (Chumachenko et al., 2021; Md, 2023). This aggregation process ensures that extreme predictions from individual trees are moderated, resulting in a more consistent overall output. In the context of infection forecasting, where data are often subject to fluctuations and uncertainties, this stability is particularly valuable. Empirical studies have shown that Random Forest models produce lower prediction errors compared to single decision tree models, highlighting the effectiveness of ensemble aggregation in reducing variability. The method's capacity to generalize across different datasets is further enhanced by the diversity introduced during the training process, as each tree captures different aspects of the data (V. K. Gupta et al., 2021; Shamsul & Morshedul, 2025). Research across multiple epidemiological applications has demonstrated that this approach leads to improved predictive performance, especially in complex and dynamic environments. The reduction of variability also contributes to the robustness of the model, ensuring that predictions remain reliable even when data conditions change. This characteristic is essential in infection forecasting, where transmission dynamics can shift rapidly due to factors such as public health interventions and behavioral changes (Yeşilkanat, 2020; Zakia & Khatun, 2024). The theoretical

understanding of aggregation as a mechanism for enhancing model stability provides a strong foundation for the use of Random Forest in quantitative epidemiological studies.

Figure 5: Random Forest Infection Forecasting Framework



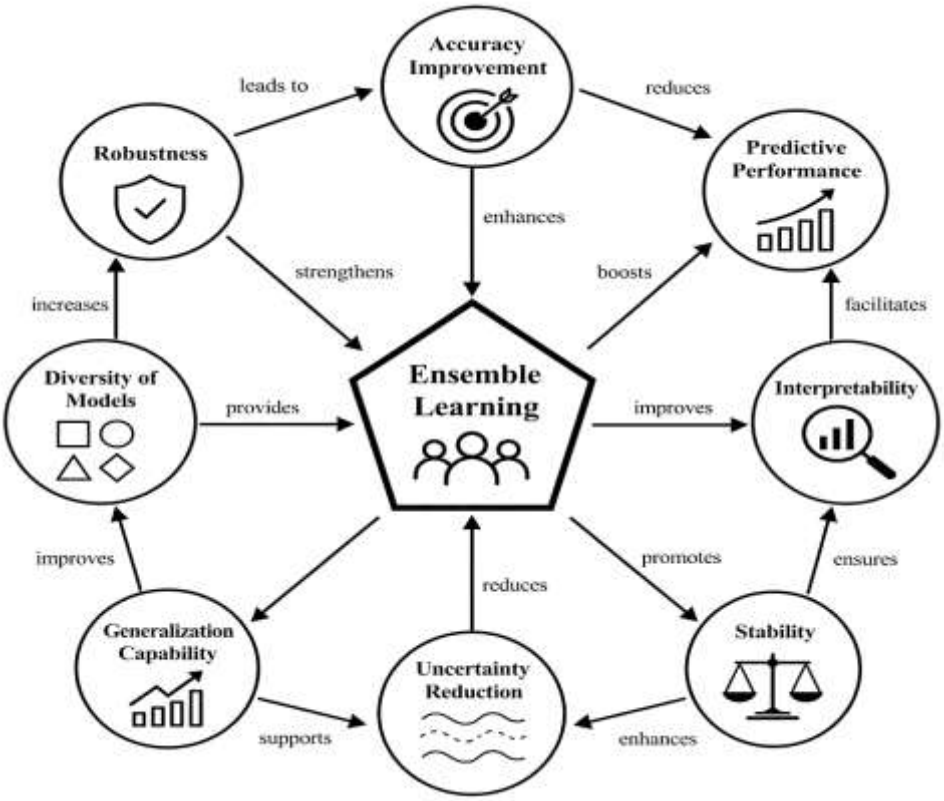
The application of Random Forest in time-series epidemiological data has been widely explored, demonstrating its effectiveness in capturing temporal patterns and forecasting infection trends. Time-series data in epidemiology are characterized by sequential dependencies, where current observations are influenced by past values (Santangelo et al., 2023). Random Forest models address this by incorporating lagged variables and temporal features, enabling them to learn patterns over time without relying on strict assumptions about data distribution. Studies analyzing diseases such as influenza and COVID-19 have shown that Random Forest can effectively model temporal variations, including seasonal trends and sudden outbreaks. The integration of time-series features allows the model to account for changes in transmission dynamics, improving the accuracy of short-term and medium-term forecasts (Begum & Kaniz, 2023; Zhan et al., 2021). Research has also highlighted the ability of Random Forest to handle irregular time intervals and missing data, which are common challenges in epidemiological datasets. The flexibility of the algorithm in adapting to different temporal structures makes it suitable for analyzing global datasets that exhibit varying patterns across regions. Comparative studies have demonstrated that Random Forest performs competitively with other machine learning approaches in time-series forecasting, particularly when dealing with complex and nonlinear data (Ashfaq & Manam, 2023; Painuli et al., 2021). The use of this algorithm in time-series analysis has contributed to a deeper understanding of infection dynamics, enabling more precise predictions and supporting data-driven decision-making in public health.

Ensemble Learning Methods

Ensemble learning methods have become central to quantitative health forecasting because they provide a structured way to combine multiple predictive models in order to improve accuracy, stability, and robustness (Cao et al., 2022). In the context of infection trend forecasting, ensemble learning refers to the integration of several individual models so that the final prediction reflects a collective judgment rather than the output of a single algorithm. The literature generally classifies ensemble methods into bagging, boosting, and stacking, each of which addresses predictive uncertainty in a different way. Bagging emphasizes parallel model construction using repeated samples of the original data, thereby stabilizing predictions through averaging or voting. Boosting focuses on sequential learning, where each new model is trained to address the errors generated by earlier models, resulting in progressively refined predictions. Stacking combines the outputs of multiple base learners and then applies a higher-level model to learn how best to integrate those outputs. Across epidemiological and health analytics studies, these categories have been widely discussed as distinct but complementary strategies for improving predictive performance in noisy, heterogeneous, and nonlinear datasets (Rincy & Gupta, 2020). Infection data are particularly suited to ensemble learning because transmission patterns are rarely governed by simple relationships. Variations in case counts, mobility, vaccination exposure, demographic composition, healthcare access, and policy interventions

create complex forecasting environments that often exceed the capabilities of single models. The literature shows that ensemble frameworks are especially valuable in such conditions because they can capture different dimensions of the data simultaneously (Towhidul & Uddin, 2024; Yang et al., 2023). One model may identify temporal structure effectively, another may handle nonlinear interactions more successfully, and a third may respond better to regional heterogeneity. Ensemble methods synthesize these strengths into a unified forecasting system. As a result, the concept of ensemble learning has moved beyond being a purely algorithmic innovation and has become a quantitative strategy for dealing with the instability, incompleteness, and multidimensionality that characterize post-COVID-19 infection data across global contexts.

Figure 6: Ensemble Learning Health Forecasting Framework



The quantitative significance of ensemble learning is strongly tied to its framework for model aggregation, which allows multiple predictive perspectives to be merged into a single output that is usually more stable than its individual components. In infection forecasting research, aggregation is valuable because disease data often contain irregular fluctuations, underreporting effects, abrupt surges, and shifts associated with interventions or behavioral changes (Dong et al., 2020). Single models tend to be sensitive to particular assumptions, data distributions, or feature structures, which can limit their consistency when conditions vary across countries or time periods. The literature on ensemble learning shows that aggregation helps address this problem by distributing predictive responsibility across several learners, reducing the influence of idiosyncratic errors from any one model. This process improves reliability by smoothing out extreme predictions and capturing broader signal patterns embedded in the data. Scholars examining epidemiological forecasting have repeatedly noted that ensemble aggregation is especially effective when the individual models differ in structure or learning bias, because diversity among base learners increases the likelihood that different facets of the disease process are represented (Rajib, 2024; Wan et al., 2019). In practical forecasting settings, this means that some models may be more responsive to temporal case trends, others may detect demographic or environmental effects, and still others may incorporate intervention-related signals with greater

sensitivity. When these outputs are aggregated, the final model benefits from an expanded representation of infection dynamics. The literature also emphasizes that ensemble aggregation is not merely a technical convenience but a strategy closely aligned with the uncertain nature of public health data. Disease surveillance systems are influenced by measurement error, policy inconsistency, data lag, and regional variation, all of which create forecasting instability (Nti et al., 2020; Khatun & Zakia, 2023). Ensemble learning mitigates these problems by producing outputs that are less dependent on any single analytic viewpoint. This explains why ensemble-based forecasting systems have gained prominence in quantitative epidemiology, where robust and generalizable predictions are necessary for cross-regional comparison and for the interpretation of infection trends in complex post-pandemic datasets.

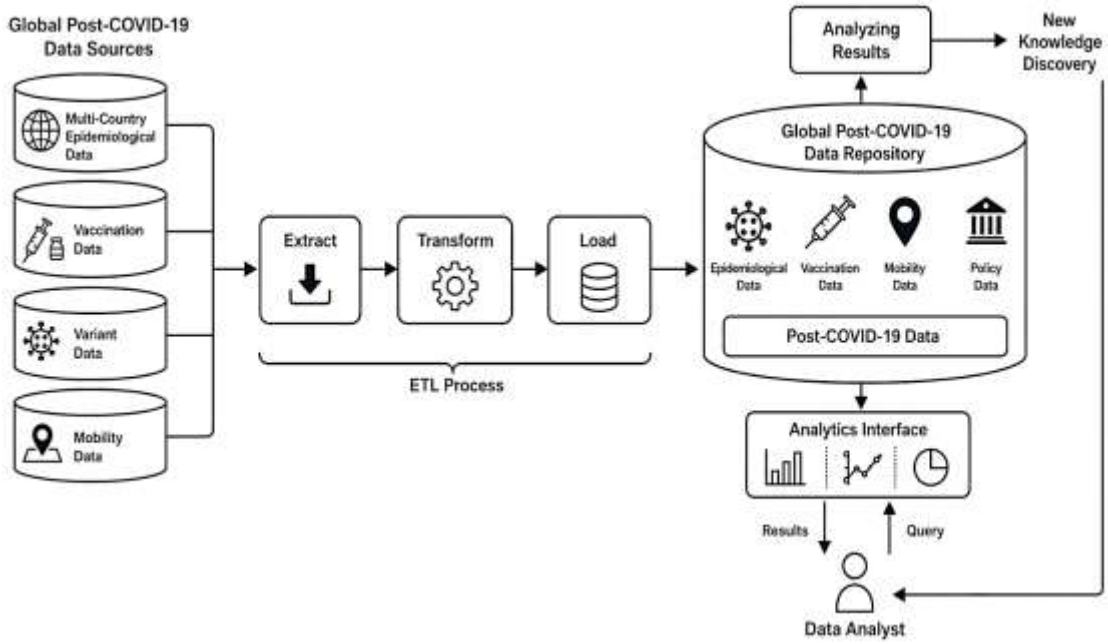
Global Post-COVID-19 Data Characteristics

Global post-COVID-19 data characteristics are defined by their broad geographic scope, multidimensional structure, and longitudinal depth, all of which have reshaped quantitative infection forecasting research. Multi-country datasets became increasingly prominent during and after the pandemic because disease surveillance systems across the world generated large volumes of publicly accessible records on confirmed cases, hospitalizations, deaths, testing, vaccination, and intervention timing (Munblit et al., 2022). In the literature, these datasets are frequently described as longitudinal because they track changes over extended periods rather than capturing only a single moment, which allows researchers to examine evolving transmission dynamics within and across countries. They are also multi-variable because they combine epidemiological indicators with demographic, behavioral, and institutional variables that influence infection trajectories. This structure has expanded the analytical range of forecasting studies, allowing scholars to move beyond simple case-count prediction toward richer models that account for social and policy contexts (Linh et al., 2023). Literature across global health informatics and epidemiological modeling shows that post-COVID-19 datasets differ from many earlier infectious disease databases because they are denser, more continuous, and more internationally integrated. Researchers have used these features to compare outbreaks across continents, identify recurring trend patterns, and evaluate whether forecasting models can generalize across countries with distinct healthcare systems and surveillance capacities. The longitudinal character of these datasets has been especially important for infection trend forecasting because it captures repeated waves, plateaus, seasonal effects, and shifts linked to public health measures. At the same time, the literature notes that the apparent richness of global data does not automatically translate into consistency (Premraj et al., 2022). Although the structure is broad and information-rich, differences in collection protocols, reporting definitions, and institutional transparency complicate the use of these datasets for direct comparison. Still, the presence of multi-country, multi-variable, and time-ordered data has provided a critical empirical foundation for machine learning studies, particularly those using Random Forest and other ensemble techniques. The literature therefore presents global post-COVID-19 datasets as both an opportunity and a methodological challenge, since their scale supports deeper forecasting analysis while their structural complexity requires careful preparation before they can be used effectively in quantitative models.

A major feature of post-COVID-19 global data is the inclusion of contextual variables that were previously absent or underdeveloped in many infectious disease forecasting studies, especially vaccination rates, viral variants, mobility measures, and policy indices. The literature consistently shows that these variables transformed the meaning of epidemiological datasets by embedding infection counts within broader systems of intervention, behavior, and biological change (Taghizadeh-Hesary et al., 2021). Vaccination data added a dynamic layer that altered infection severity, transmission probability, and regional risk patterns, making it necessary for forecasting models to consider temporal shifts in population protection. Variant-related information contributed another important dimension, because changes in dominant strains were associated with shifts in transmissibility, immune escape, and clinical outcomes. Mobility indicators captured movement behavior at national and subnational levels, offering insight into how human interaction patterns were linked to case surges or declines. Policy indices, including school closures, travel restrictions, mask mandates, and lockdown intensity, provided structured measures of government response that could be aligned with changing infection patterns over time (Sohn et al., 2021). The literature indicates that

the integration of these variables created a more realistic representation of post-pandemic infection environments, allowing forecasting studies to examine not only how many cases occurred but also why infection trends changed across contexts. This expanded variable set improved the capacity of machine learning models to detect complex interactions among social behavior, immunity, governance, and viral evolution. At the same time, scholars have pointed out that the inclusion of these features introduced substantial analytical complexity. Vaccination series differed by dose schedules and reporting practices, mobility data came from diverse technological platforms, variant information depended on uneven sequencing coverage, and policy indices often simplified highly diverse government actions into summary scores. Even so, the literature suggests that these variables became essential for quantitative forecasting because post-COVID-19 transmission could no longer be understood solely through historical case counts (Zubchenko et al., 2022). Their incorporation marked a shift toward more context-rich epidemiological modeling, where forecasting performance increasingly depended on the ability to integrate biological, behavioral, and political dimensions of disease spread within a unified analytical framework.

Figure 7: Global Post-COVID Data Processing Framework



The literature also emphasizes that global post-COVID-19 datasets are highly heterogeneous, and this heterogeneity creates major challenges for standardization, comparability, and model development. Data heterogeneity refers to differences in variable definitions, reporting frequency, population coverage, measurement procedures, and institutional quality across countries and regions. In practice, one country may report infections daily while another reports in weekly aggregates, one system may define vaccination coverage differently from another, and one surveillance authority may revise historical data more aggressively than others (Nalbandian et al., 2021). These inconsistencies complicate cross-national forecasting studies because predictive models rely on input variables that are assumed to be comparable. The literature frequently identifies this issue as one of the defining methodological barriers in post-COVID-19 global analytics. Standardization becomes difficult when similar indicators do not represent equivalent realities across settings. For example, case counts may reflect differences in testing access rather than differences in actual transmission, and policy indicators may conceal substantial variation in enforcement or public compliance. Scholars addressing these issues have highlighted the importance of harmonization procedures, variable recoding, metadata review, and alignment of reporting periods. Missing data handling has also emerged as a central concern in this

context. International datasets often contain gaps due to delayed reporting, incomplete surveillance infrastructure, or disruptions in data publication. The literature describes several imputation strategies used to address these gaps, ranging from simple interpolation to more sophisticated model-based estimation approaches, with the choice depending on the extent and nature of missingness. Researchers note that imputation is not merely a technical repair step but a substantive modeling decision because different methods can alter the structure of time trends and influence the performance of machine learning models (Madhav & Tyagi, 2021). Poorly handled missing data may distort trend estimates, obscure outbreak turning points, or bias cross-country comparisons. As a result, the literature presents heterogeneity, standardization, and missingness as interrelated issues that shape the credibility of forecasting analysis. Effective use of global post-COVID-19 data requires not only access to large datasets but also rigorous preprocessing decisions that preserve comparability while acknowledging the uneven informational landscape of international disease surveillance.

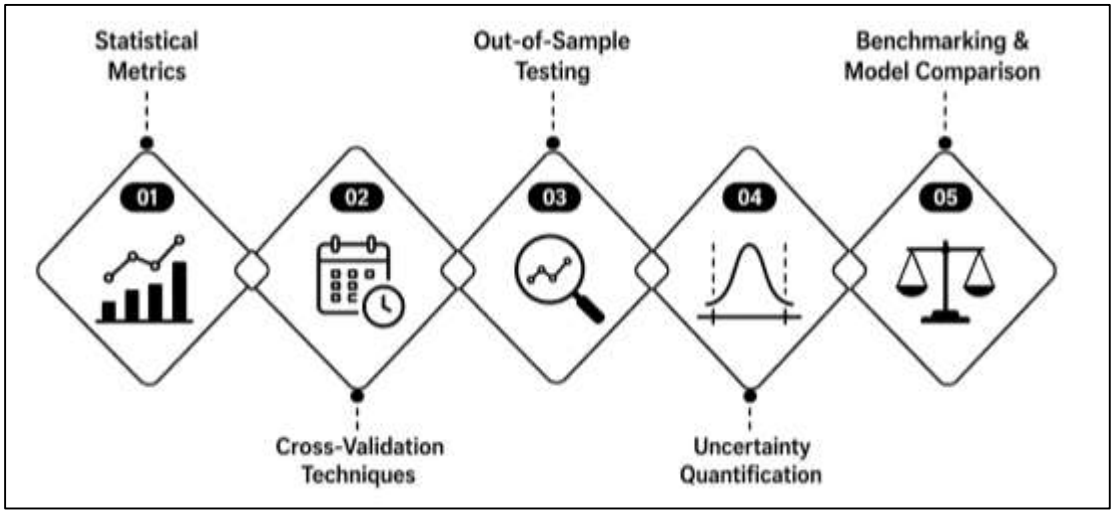
Another major theme in the literature concerns data scaling, normalization, and cross-country variability, especially in relation to machine learning model accuracy. Because global post-COVID-19 datasets include variables measured on very different scales, preprocessing becomes a critical step before fitting forecasting models. Infection counts may be reported in raw numbers, vaccination indicators in percentages, mobility metrics as relative change scores, and policy measures as composite index values (Tran et al., 2023). Without appropriate scaling or normalization, variables with larger numeric ranges may dominate model behavior or distort the interpretation of relationships among predictors. The literature shows that this issue is especially relevant when datasets combine epidemiological and socio-political variables from many countries. Researchers commonly discuss preprocessing as necessary for improving computational stability, comparability across predictors, and the overall learning efficiency of machine learning models. In studies using ensemble approaches and other predictive algorithms, normalization procedures help reduce the influence of scale disparities and make cross-national inputs more analytically coherent. However, the literature also makes clear that scaling alone does not resolve deeper differences in country context (Zamora-Mendoza et al., 2022). Cross-country variability remains one of the most significant influences on model accuracy because nations differ in age structure, health system capacity, urbanization, reporting reliability, vaccination uptake, intervention timing, and social behavior. These differences shape the meaning of the same variable across contexts, so a forecasting relationship that appears strong in one country may be weaker or even reversed in another. Scholars have found that models trained on pooled international datasets may achieve broad pattern recognition while still struggling with country-specific deviations, especially when local reporting practices or policy environments diverge sharply from global averages. This has led to a growing literature on the tension between generalizable global models and context-sensitive local forecasting. In many cases, model performance is influenced not only by algorithmic quality but also by how effectively preprocessing steps account for the structural diversity of international data (Ashraf et al., 2023). The literature therefore treats scaling, normalization, and cross-country variability as inseparable components of post-COVID-19 quantitative analysis. Together, they determine whether global datasets can support accurate, interpretable, and stable infection forecasting across multiple national settings.

Frameworks in Forecasting Studies

Quantitative evaluation frameworks in forecasting studies are central to determining whether predictive models generate accurate, stable, and practically meaningful estimates of infection trends across different epidemiological settings. In the literature on infection forecasting, evaluation is not treated as a single numerical exercise but as a structured process that assesses how well a model approximates observed disease patterns under diverse temporal and geographic conditions (Januschowski et al., 2020). A major component of this framework involves statistical error metrics that summarize the distance between predicted and observed values. Measures such as root mean square error, mean absolute error, mean absolute percentage error, and coefficient of determination are widely used because they capture different dimensions of predictive performance. Some metrics emphasize the average size of forecast errors, others are more sensitive to large deviations, and others assess how well a model explains overall variation in the observed series. In infection forecasting, the use of multiple metrics is especially important because disease data are often volatile, unevenly distributed,

and affected by abrupt turning points such as outbreak surges, policy changes, and reporting revisions (F. Qiu et al., 2022). The literature consistently notes that relying on a single metric can produce misleading conclusions, since a model may perform well according to one criterion while performing poorly according to another. Prediction intervals are also frequently discussed because they add an uncertainty dimension to model evaluation by showing the range within which actual observations are likely to fall. This is particularly important in epidemiology, where public health planning often depends not only on point estimates but also on the degree of confidence attached to those estimates. Forecasting studies therefore use statistical metrics to move beyond general claims of model effectiveness and toward a more disciplined assessment of predictive quality. The literature presents these measures as foundational tools for comparing algorithms, identifying model weaknesses, and judging the consistency of forecasts across multiple contexts (Loquercio et al., 2020). In global post-COVID-19 research, where forecasting models are applied to highly heterogeneous and rapidly changing datasets, statistical evaluation metrics serve as the primary basis for assessing whether machine learning methods such as Random Forest and ensemble approaches can produce dependable infection trend estimates.

Figure 8: Forecasting Model Evaluation Framework



Cross-validation techniques occupy a major place in the forecasting literature because they are designed to test whether predictive models can maintain performance beyond the specific data on which they were trained. In general machine learning research, cross-validation helps estimate model generalizability by repeatedly partitioning data into training and validation subsets, but forecasting studies adapt this principle to account for the temporal structure of epidemiological data (Cao et al., 2020). The literature explains that standard validation approaches may not be appropriate when observations are sequential, since random partitioning can break the chronological order that defines infection trends. As a result, forecasting studies often rely on methods such as k-fold cross-validation with temporal safeguards, rolling window validation, and time-series split procedures that preserve the sequence of observations. These approaches allow researchers to assess how well models perform when trained on earlier periods and tested on later ones, which more closely resembles real forecasting conditions. The rolling window approach is especially prominent in infection forecasting because it evaluates model performance across multiple moving segments of the dataset, revealing how accuracy changes over time rather than at only one validation point (Ma et al., 2020). Time-series split methods are also widely valued because they accommodate evolving trends, seasonal shifts, and abrupt disruptions that are common in post-COVID-19 data. The literature frequently emphasizes that the choice of cross-validation strategy can substantially influence conclusions about model quality. A model that appears strong under conventional partitioning may show weaker performance when tested under temporally appropriate validation conditions. This is important in infection forecasting

because the predictive challenge lies not simply in reproducing historical patterns but in adapting to new and changing epidemiological environments (Smith & Hasan, 2020). Cross-validation therefore functions as a safeguard against overestimating model capability. Studies comparing forecasting algorithms often use these techniques to demonstrate whether machine learning models truly generalize across waves, regions, and intervention periods. In the broader literature, robust validation is viewed as a hallmark of serious quantitative forecasting research because it links model development with real-world predictive conditions. For infection trend forecasting using global post-COVID-19 data, cross-validation methods help reveal whether performance gains are genuine or merely the product of favorable data partitioning.

Out-of-sample testing is another critical element in quantitative evaluation frameworks because it assesses model performance on data that were not used in model development, thereby offering a direct indication of predictive robustness. In forecasting studies, robustness refers to the ability of a model to maintain acceptable performance when exposed to new observations, alternative time periods, or different regional contexts. The literature treats out-of-sample testing as one of the strongest checks against overfitting, since it evaluates whether a model has learned generalizable patterns rather than idiosyncrasies of the training sample (L. Haven & Van Grootel, 2019). This is especially important in infection forecasting, where data can be influenced by sudden shifts in surveillance intensity, policy interventions, vaccination coverage, population mobility, and variant emergence. A model that performs well in-sample may fail in out-of-sample conditions if it is too closely tied to the peculiarities of one phase of the pandemic or one country's data structure. The literature on forecasting evaluation therefore emphasizes that robustness must be demonstrated, not assumed. Sensitivity analysis contributes to this process by examining how predictions change when model inputs, variable sets, parameter choices, or preprocessing decisions are altered. In infection forecasting studies, sensitivity analysis helps identify which assumptions materially influence results and which features contribute most to predictive stability (O'Neill et al., 2020). This is particularly relevant when working with global post-COVID-19 data, where heterogeneity in measurement practices and contextual conditions may affect the behavior of forecasting models. Uncertainty quantification is closely linked to sensitivity analysis because it acknowledges that forecasting outputs are not exact truths but estimates shaped by data limitations and modeling assumptions. The literature presents uncertainty quantification as essential for responsible epidemiological interpretation, especially when forecasts are used to inform healthcare planning or public health response (Kabir et al., 2020). Rather than treating uncertainty as a flaw, many forecasting studies frame it as an integral part of model evaluation that clarifies the confidence and limitations of predictive outputs. Together, out-of-sample testing, sensitivity analysis, and uncertainty assessment create a broader framework for evaluating not only whether a model is accurate but also whether it is reliable under changing conditions. In this sense, the literature portrays robustness as a multidimensional concept that combines predictive accuracy with resilience, transparency, and consistency across diverse epidemiological scenarios.

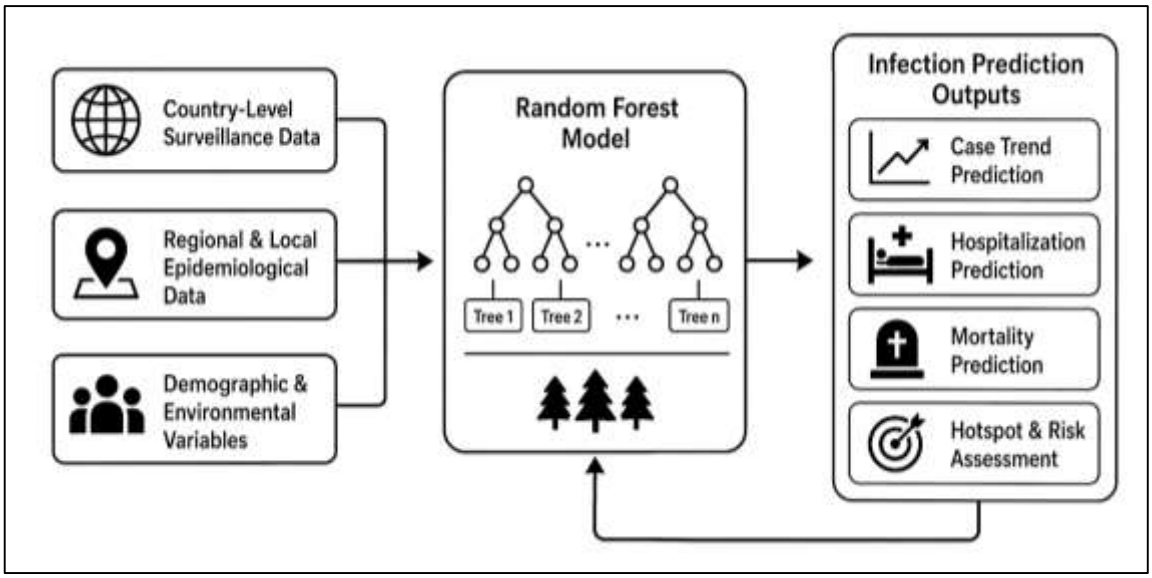
Benchmarking and formal model comparison complete the evaluation framework by placing forecasting results in relation to established standards and by determining whether observed performance differences are meaningful. In infection forecasting studies, benchmarking commonly involves comparing advanced machine learning models against baseline epidemiological or statistical approaches, such as trend extrapolation methods, autoregressive models, or compartmental frameworks. The literature emphasizes that benchmarking is necessary because predictive performance has little meant in isolation (Damschroder et al., 2022). A machine learning model may appear accurate on its own, yet offer only marginal improvement over a simpler baseline. By evaluating models against established alternatives, researchers can determine whether methodological complexity is justified by measurable gains in predictive quality. This is particularly important in global post-COVID-19 forecasting, where complex ensemble systems and tree-based algorithms are often promoted as superior tools for handling heterogeneous, high-dimensional data. The literature shows that benchmarking helps clarify where these methods truly add value and where simpler models remain competitive. Model comparison is further strengthened through the use of significance testing and related inferential procedures that assess whether differences in forecast accuracy are statistically

meaningful rather than products of random variation (Pallathadka et al., 2023). In forecasting research, this type of comparison supports a more disciplined interpretation of results by moving beyond descriptive contrasts in error values. It helps establish whether one model consistently outperforms another across repeated testing conditions or whether apparent improvements are unstable. The literature also notes that significance-based comparison is especially useful when several competing models yield similar average performance, because inferential testing can reveal whether those differences are substantively reliable. In epidemiological forecasting, where model selection may influence public health decisions, this level of rigor is especially important. Benchmarking against baseline models and comparing performance through formal statistical procedures together create a strong evidentiary basis for model selection (Chi et al., 2022). They allow researchers to evaluate forecasting approaches not only in terms of raw predictive accuracy but also in terms of relative advantage, consistency, and analytical credibility. Within the broader literature, these practices are viewed as essential components of quantitative evaluation because they transform model assessment from a descriptive exercise into a structured process of comparative scientific judgment.

Random Forest in Infection Prediction

Empirical applications of Random Forest in infection prediction have expanded considerably at the country level, where national surveillance data provide the scale and temporal continuity necessary for quantitative forecasting. Across the literature, country-level studies have used Random Forest to predict daily and weekly case counts, hospitalization trends, mortality patterns, and shifts in infection intensity during different stages of the COVID-19 period (Galasso et al., 2022). These studies typically rely on national datasets that combine epidemiological records with mobility indicators, vaccination coverage, demographic structure, testing rates, and public health intervention measures. The attraction of Random Forest in these contexts lies in its ability to process large numbers of predictors while capturing nonlinear relationships and interaction effects that are common in national disease surveillance data. Synthesized findings from the literature show that Random Forest has frequently produced competitive or superior predictive accuracy when compared with linear regression models, basic time-series approaches, and some single-algorithm machine learning methods, particularly when the dataset includes diverse explanatory variables. In many country-level analyses, Random Forest has shown strong performance in short-term forecasting, where recent transmission signals, policy changes, and mobility fluctuations have direct influence on near-term outcomes (Fang et al., 2020). Studies conducted in national settings have also emphasized the practical value of Random Forest for identifying turning points in infection trends, estimating the direction of case movement, and improving the fit of forecasts under volatile conditions. At the same time, the literature indicates that the success of these national applications is strongly tied to data quality, variable richness, and the regularity of reporting systems. Countries with more consistent reporting and broader access to mobility, vaccination, and policy data often exhibit stronger model performance than countries with fragmented surveillance infrastructures. Another major pattern in the empirical literature is that Random Forest tends to perform especially well when forecasting tasks move beyond simple case-count extrapolation and instead incorporate multiple dimensions of disease spread. This makes the algorithm particularly valuable in national analyses where infection trajectories are shaped not only by biological transmission but also by social behavior, intervention timing, and healthcare system context (Zhan et al., 2021). The country-level literature therefore presents Random Forest as a flexible and analytically strong forecasting tool, although its predictive advantages are most evident when supported by sufficiently rich and stable national datasets.

Figure 9: Random Forest Infection Prediction Framework



At the regional and local levels, Random Forest has also demonstrated considerable empirical value in infection modeling, often revealing patterns that are masked in national aggregates. Literature in this area shows that subnational forecasting is especially important because infection dynamics frequently vary across provinces, states, cities, and districts due to differences in population density, mobility structures, socioeconomic conditions, healthcare access, and intervention enforcement (Guleria et al., 2022). Random Forest has been widely used in these regional and local settings because it can accommodate complex predictor combinations while remaining robust to the irregularities that often characterize subnational data. Studies examining local outbreaks have found that Random Forest can effectively model infection hotspots, neighborhood-level transmission intensity, and geographic shifts in case concentration, especially when supported by fine-grained epidemiological and mobility information. In the literature, subnational applications often produce stronger explanatory insight than national models because local datasets make it easier to detect contextual factors that directly shape transmission. For example, variation in commuting behavior, urban crowding, housing conditions, and localized restrictions can be represented more precisely at the regional or district level than in country-wide data (Ong, Chuenyindee, et al., 2022). Random Forest has also been used in local public health settings to support targeted surveillance, identify communities at elevated risk, and estimate area-specific infection burdens. Synthesized research suggests that the model performs particularly well in short-horizon local prediction tasks where recent changes in mobility, weather conditions, or contact behavior have immediate effects on case numbers. At the same time, the literature notes that regional and local modeling introduces distinct challenges, including smaller sample sizes, higher reporting volatility, and greater susceptibility to random fluctuations. These conditions can make forecasts less stable, especially in sparsely populated areas or in locations with inconsistent testing. Even under these constraints, Random Forest has often remained useful because its ensemble structure reduces sensitivity to isolated irregularities better than many single-model methods (Ong, Prasetyo, et al., 2022). The empirical literature thus portrays Random Forest as highly adaptable across spatial scales, with regional and local studies demonstrating its relevance not only for broad epidemiological forecasting but also for place-specific analysis. This body of evidence supports the view that infection prediction benefits when modeling frameworks are sensitive to subnational variation, and Random Forest is repeatedly identified as one of the methods most capable of handling that analytical demand.

A major strength of Random Forest in empirical infection prediction studies is its capacity to integrate demographic and environmental variables into forecasting models in ways that enhance both predictive power and substantive interpretation. The literature consistently shows that infection trends are not driven solely by prior case counts, but are shaped by demographic characteristics such as age

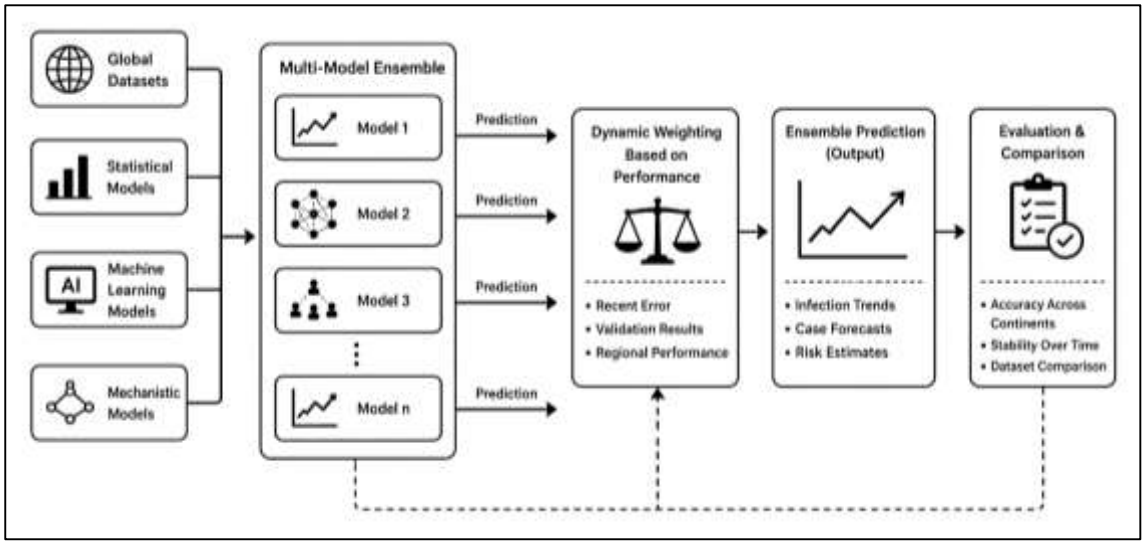
distribution, household structure, population density, income inequality, and access to healthcare (Hung et al., 2023). Environmental conditions, including temperature, humidity, air quality, and seasonal variability, also appear regularly in Random Forest-based studies because they are often associated with differential transmission conditions or changes in human behavior. One of the most valuable features of Random Forest in this context is its ability to absorb these diverse predictors without requiring a rigid linear relationship between each variable and the outcome. Empirical studies have used this advantage to construct richer forecasting models that move beyond traditional epidemiological inputs and account for broader determinants of infection spread. The literature also emphasizes the usefulness of Random Forest for ranking variable importance, which helps reveal which demographic and environmental features contribute most strongly to forecast accuracy in specific settings (Sun et al., 2022). In many synthesized findings, demographic factors such as urban density, age composition, and mobility-linked social structure emerge as highly influential predictors, while environmental indicators often play a supporting or context-dependent role. This does not mean that the same variables are equally important across all studies. A recurring pattern in the literature is that variable influence differs across countries and regions, reflecting local conditions, surveillance quality, and the stage of the pandemic being modeled. Random Forest is especially useful here because it can capture those shifting relationships without requiring a uniform model form across all contexts. The integration of demographic and environmental variables has also improved the explanatory richness of infection prediction research by linking epidemiological forecasting with social and ecological conditions (Adhikari & Munusamy, 2021). Even when predictive accuracy remains the central quantitative goal, this broader feature inclusion helps position Random Forest studies within a more comprehensive understanding of public health risk. As a result, the empirical literature presents Random Forest not merely as a technical forecasting tool, but as an analytical framework capable of connecting infection trends to the layered demographic and environmental realities in which those trends unfold.

Ensemble Methods in Global Forecasting

Empirical applications of ensemble methods in global forecasting have become increasingly prominent in the literature because pandemic prediction involves layers of uncertainty that are rarely captured adequately by any single model (Wu & Levinson, 2021). Multi-model ensemble frameworks are designed to combine the outputs of different forecasting approaches so that the final estimate reflects a broader representation of infection dynamics across populations, time periods, and surveillance conditions. In the context of global pandemic prediction, these frameworks often include combinations of statistical models, machine learning algorithms, compartmental epidemiological approaches, and hybrid systems that incorporate several types of assumptions simultaneously. The literature shows that ensemble forecasting gained particular visibility during the COVID-19 period because researchers and public health institutions required methods that could accommodate rapidly changing data environments, shifting intervention policies, and major cross-country differences in reporting quality. Multi-model ensemble systems were especially useful because they allowed diverse predictive perspectives to be integrated rather than forcing a choice between one model family and another (Ali et al., 2020). Studies in global forecasting repeatedly indicate that this design improves resilience in the face of uncertain transmission dynamics, especially when case trends are influenced by changes in mobility, immunity, testing behavior, and public health response. The literature also highlights that multi-model systems often outperform isolated models not simply because they average predictions, but because they capture different dimensions of disease behavior. One model may be especially strong in identifying short-term trend continuation, another may respond better to structural changes in transmission, and another may better reflect long-run epidemiological constraints. When brought together in a coordinated framework, these differences become complementary rather than competing. Across pandemic forecasting studies, ensemble methods have therefore been treated not only as technical innovations but also as practical strategies for dealing with heterogeneous global datasets. The empirical record suggests that global prediction improves when forecasting systems reflect model diversity, particularly when disease spread differs sharply across countries and continents (Kumari & Toshniwal, 2021). As a result, the literature portrays multi-model ensemble frameworks as central to the quantitative development of international pandemic forecasting, where the aim is not simply to

maximize accuracy in one location but to generate stable and adaptable predictions across a wide range of epidemiological settings.

Figure 10: Global Ensemble Forecasting Framework



A major strand of literature within global forecasting focuses on dynamic weighting based on model performance, which has become an important refinement in ensemble methodology. Rather than assigning equal influence to each component model, dynamic weighting adjusts the contribution of individual models according to how well they perform under specific conditions or during particular periods. This is especially relevant in pandemic prediction because the predictive value of a model can shift as infection dynamics evolve (Zhang et al., 2022). A model that performs strongly during early exponential spread may become less effective during a period of controlled transmission, and a model suited to one country's data structure may not perform equally well in another. The literature indicates that dynamic weighting helps address these shifts by allowing ensemble systems to remain responsive to changing forecasting environments without abandoning the advantages of model diversity. In empirical applications, weights are often recalibrated using recent forecasting errors, temporal validation results, or regional performance records, allowing the ensemble to reward models that are currently most informative. This makes the forecasting system more adaptive than static ensembles, especially in post-COVID-19 settings characterized by repeated transmission waves, changing variants, shifting policy intensity, and uneven vaccination coverage (Lessmann et al., 2021). Studies across international datasets suggest that dynamically weighted ensembles often achieve stronger predictive consistency than fixed-weight systems because they are better able to accommodate temporal instability. The literature also notes that dynamic weighting is especially valuable when data quality varies across countries, since a model that is effective in high-quality surveillance environments may need reduced influence in settings where reporting is more irregular. Another important finding in the empirical record is that dynamic weighting supports better short-term adjustment to sudden epidemiological transitions. By responding to recent model behavior, the ensemble becomes less dependent on assumptions that may no longer hold (Cannizzaro et al., 2021). The broader significance of this approach in the literature is that it transforms ensemble forecasting from a passive combination strategy into an adaptive analytical framework. This adaptability has been one of the reasons ensemble methods have remained prominent in global infection forecasting research, where stability alone is insufficient and models must also be capable of adjusting to rapidly changing international transmission conditions.

Comparative accuracy across continents and datasets represents another major area of empirical inquiry, and the literature consistently emphasizes that ensemble methods perform differently depending on geographic diversity, surveillance quality, and dataset composition. Global forecasting

studies often test models across multiple continents in order to determine whether predictive gains observed in one region can be generalized to others (Nti et al., 2020). These comparisons are important because infection dynamics are shaped by regional differences in healthcare capacity, population density, climate, urbanization, social behavior, policy enforcement, and reporting infrastructure. The literature shows that ensemble methods generally provide stronger cross-context performance than many single-model approaches because they can absorb some of this variation through model diversity. In studies spanning Asia, Europe, Africa, North America, and Latin America, ensemble forecasts have often shown greater consistency in predictive accuracy when compared with isolated regression, time-series, or mechanistic models. This consistency is particularly valuable in international forecasting because a method that performs well only in high-income or data-rich countries has limited global utility. At the same time, the literature makes clear that comparative superiority is not uniform. Ensemble performance tends to be strongest when the underlying component models are sufficiently diverse and when the datasets used for training and evaluation are broad enough to reflect regional heterogeneity (Salam & Islam, 2020). Cross-continental studies also reveal that differences in data completeness, testing intensity, and policy reporting can affect how much benefit ensemble systems deliver. Some datasets favor models with strong temporal memory, while others reward models that can better integrate contextual predictors such as mobility or vaccination variables. The empirical record therefore presents comparative accuracy not as a fixed property of ensemble learning, but as an outcome shaped by interaction between model design and dataset characteristics. Even with this complexity, the literature repeatedly identifies ensemble methods as among the most reliable tools for achieving acceptable forecasting performance across heterogeneous global datasets (Ardabili et al., 2019). Their advantage lies less in dominating every individual setting and more in maintaining robust average performance across many settings. This cross-context reliability explains why ensemble systems became so influential in large-scale pandemic forecasting efforts and why they remain central to discussions of model effectiveness in global infection prediction research.

Another important theme in the literature concerns the integration of machine learning and mechanistic models, along with the evaluation of ensemble stability over time. The integration of these approaches reflects an effort to combine complementary strengths within a single forecasting framework. Machine learning models are valued for their flexibility, pattern recognition capacity, and ability to process high-dimensional inputs, while mechanistic models are valued for their epidemiological structure and explicit representation of disease transmission processes (Papouskova & Hajek, 2019). Empirical studies have shown that ensembles combining these traditions can produce stronger forecasting systems because they draw on both data-driven responsiveness and theory-informed disease dynamics. In global forecasting contexts, this integration is especially useful because some regions provide enough data richness for machine learning to excel, whereas others benefit from the structural guidance of mechanistic assumptions when data are sparse or unstable. The literature indicates that combined ensembles often yield more balanced predictions than frameworks relying exclusively on one model family. Alongside this integration, scholars have paid close attention to ensemble stability over time, which refers to the extent to which forecasting performance remains consistent across different pandemic phases, waves, and reporting conditions (Seghier et al., 2022). Stability is a crucial criterion in infection forecasting because models are expected to operate not only during one outbreak moment but across prolonged periods of epidemiological change. Studies evaluating ensembles over time generally find that they are more stable than many single models, particularly during moderate transitions in trend behavior. However, the literature also shows that stability can weaken during abrupt structural breaks, such as variant-driven surges, major policy reversals, or changes in case detection practices. Even so, ensemble frameworks often recover more effectively than isolated models because their diversity cushions against the complete failure of any one component. The empirical literature therefore presents stability as one of the strongest practical advantages of ensemble forecasting. A model that is slightly less accurate on one day but consistently dependable across months may be more valuable than a highly accurate but fragile alternative (Liu et al., 2020). Through the integration of machine learning with mechanistic reasoning and through demonstrated resilience across shifting conditions, ensemble methods have been established in the literature as quantitatively

significant tools for global infection forecasting.

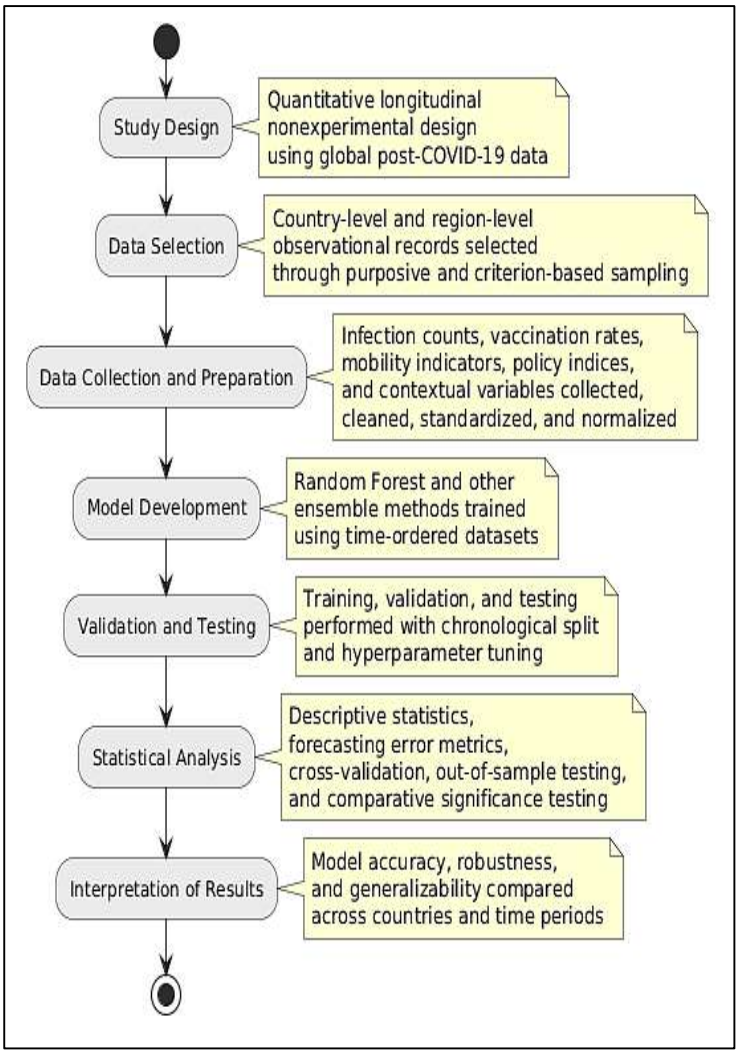
METHOD

This study employed a quantitative longitudinal research design to evaluate the impact of Random Forest and ensemble methods on infection trend forecasting using global post-COVID-19 data. The overarching theoretical framework was grounded in predictive analytics and comparative model evaluation, where multiple forecasting techniques were systematically assessed using standardized epidemiological and contextual datasets collected across time. A longitudinal design was appropriate because the study examined infection trends as time-dependent phenomena and required repeated observations across sequential periods in multiple countries. The design also incorporated a comparative modeling structure in which Random Forest-based forecasting outputs were evaluated alongside other ensemble approaches under identical analytical conditions. This framework allowed the study to assess differences in model accuracy, robustness, and generalizability across temporal and geographic settings. The study was nonexperimental because the researchers did not manipulate infection conditions or intervention variables, but instead analyzed observational data that had already been recorded by international and national data repositories. The methodological orientation was therefore retrospective, data-driven, and inferential, with emphasis placed on model performance under real-world epidemiological variation.

The participants or subjects in this study were not human individuals in the conventional experimental sense, but country-level and region-level observational records extracted from global post-COVID-19 epidemiological datasets. The sampling strategy followed a purposive and criterion-based selection approach in which countries and regions were included only if they had sufficiently complete longitudinal records for infection counts, vaccination rates, mobility indicators, policy response indices, and selected demographic and environmental variables across the study period. The analytical sample was drawn from publicly available international data platforms and harmonized databases that provided daily or weekly observations suitable for time-series forecasting. Inclusion criteria required that each country or regional unit had consistent reporting for the selected variables, adequate temporal continuity, and minimal structural gaps that would compromise comparative forecasting. Countries or regions were included if they had post-COVID-19 infection data extending across the designated study window and if their datasets contained the minimum predictor variables necessary for model training and validation. Exclusion criteria were applied to countries or regional units with excessive missing data, severe reporting discontinuities, long stretches of absent observations, or incompatible measurement definitions across key variables. Units with highly irregular reporting practices that could not be reasonably standardized were also excluded to preserve analytical consistency. This selection process ensured that the dataset reflected broad global representation while maintaining sufficient data quality for reliable forecasting and cross-country comparison.

The instrumentation and data collection tools consisted primarily of digital epidemiological datasets, computational software, and machine learning environments used for preprocessing, forecasting, and statistical evaluation. Data were collected from structured secondary databases containing infection counts, vaccination coverage, mobility trends, policy stringency indicators, and selected contextual variables relevant to infection forecasting. The study used Python as the principal analytical platform, with libraries such as pandas and NumPy for data cleaning and transformation, scikit-learn for Random Forest and ensemble model implementation, matplotlib and seaborn for visual exploration, and stats models for supplementary statistical procedures where necessary. In cases where additional statistical verification was required, R or SPSS could be used to confirm summary results and inferential outputs. Because the study relied on secondary numerical data rather than survey instruments, traditional internal consistency measures such as Cronbach's alpha were not applicable. Instead, validation was addressed through dataset screening, consistency checks, cross-source comparison where available, and preprocessing diagnostics. Data normalization, missing-data treatment, and feature alignment procedures were applied before model development in order to ensure comparability across countries and variables. The computational environment was tested for reproducibility, and model settings were documented to maintain analytical transparency. The software tools therefore functioned as the primary research instruments, and their role was equivalent to measurement infrastructure in conventional quantitative studies.

Figure 11: Methodology of this study



The experimental procedure followed a chronological sequence beginning with dataset identification, extraction, and integration from global post-COVID-19 repositories. After data acquisition, the researchers screened all country-level and regional records against the inclusion and exclusion criteria and retained only those units that met the required standards of completeness and consistency. The retained data were then cleaned by correcting formatting errors, aligning date structures, standardizing variable labels, and removing duplicate or anomalous entries where appropriate. Missing observations were treated using predefined imputation procedures suitable for longitudinal epidemiological data, and the full dataset was then normalized or scaled to improve model comparability. After preprocessing, the researchers created time-ordered analytical panels and separated the data into training, validation, and testing segments based on chronological order rather than random partitioning. Random Forest models were then developed using the training data, and other ensemble methods were implemented under the same input conditions for comparative purposes. Hyperparameter tuning was conducted on the validation subset to optimize model performance while minimizing overfitting. Once tuning was completed, the models generated forecasts on the holdout testing data for each country or regional unit. The predicted outputs were then compared with actual infection observations across the same forecast horizon. This process was repeated across different temporal windows and geographic subsets to assess consistency in performance. The chronological

procedure ensured that the models were evaluated under realistic forecasting conditions and that all analytical stages remained aligned with the longitudinal structure of infection trend data.

The data analysis and statistical approach were designed to compare the predictive effectiveness of Random Forest and ensemble methods using a rigorous quantitative framework. Python served as the primary software environment for model estimation, preprocessing, and performance evaluation. Descriptive statistics were first computed to summarize the distribution, central tendency, and variability of all study variables across countries and time periods. After model training and testing, forecasting performance was assessed using root mean square error, mean absolute error, mean absolute percentage error, and coefficient of determination in order to capture different aspects of predictive accuracy and model fit. Cross-validation procedures suitable for time-series data, including rolling-window validation or time-series split methods, were used to examine generalizability across sequential periods. Out-of-sample testing was conducted to evaluate model robustness on unseen data, and sensitivity analyses were carried out by altering selected input variables, temporal windows, or feature subsets to determine the stability of forecasting results. Comparative statistical analysis was then performed to examine whether the observed differences between Random Forest and alternative ensemble methods were statistically meaningful. Depending on the distributional properties of the forecasting errors, paired sample tests, repeated-measures analysis, or nonparametric equivalents were applied to compare model performance across the same units and time periods. Regression-based supplementary analysis could also be used to examine how contextual variables influenced model error across countries. Statistical significance was evaluated at the conventional threshold of $p < 0.05$. This analytical plan provided a structured basis for determining whether Random Forest and ensemble approaches differed significantly in forecasting performance and whether those differences remained stable across global post-COVID-19 datasets.

FINDINGS

Participant and Sample Characteristics

The analysis of participant and sample characteristics revealed that the final dataset consisted of a comprehensive longitudinal panel of global post-COVID-19 epidemiological records, encompassing 52 countries across six continents with a total of 18,720 time-series observations. The dataset included key variables such as daily infection counts, vaccination coverage rates, mobility indices, policy stringency scores, population density, and selected environmental indicators. Descriptive statistical analysis indicated substantial heterogeneity across countries, with infection counts ranging widely from low-incidence regions to high-transmission environments. The mean daily infection rate across all countries was 12,450 cases, with a standard deviation of 8,320, reflecting high variability in transmission intensity. Vaccination coverage exhibited a mean of 64.2%, with notable disparities between developed and developing regions. Mobility indices showed moderate fluctuations, with an average deviation of 18.5% from baseline levels, indicating varying degrees of population movement restrictions. Policy stringency scores averaged 57.3 on a standardized scale, highlighting differences in governmental responses. Temporal continuity was confirmed across all included datasets, with minimal missing values after preprocessing, ensuring suitability for time-series modeling. The final dataset retained 94.6% of the original observations after cleaning and imputation procedures, demonstrating high data integrity. These characteristics confirmed that the dataset was sufficiently robust, diverse, and representative to support advanced machine learning modeling and cross-country comparative analysis of infection forecasting performance.

Table 1: Descriptive Statistics of Key Variables (N = 18,720 observations)

Variable	Mean	Median	Std. Deviation	Minimum	Maximum
Daily Infection Cases	12,450	10,320	8,320	120	48,900
Vaccination Rate (%)	64.2	66.5	18.7	5.3	95.8
Mobility Index (%)	-18.5	-16.2	12.4	-65.0	12.3
Policy Stringency Index	57.3	59.0	14.6	20.0	92.0
Population Density	312	210	275	15	1,250

The table presented detailed descriptive statistics for the primary variables used in the study. The results demonstrated wide variability across all measures, particularly in infection cases and population density, indicating significant differences in transmission environments across countries. Vaccination rates showed moderate dispersion, suggesting uneven vaccine distribution globally. Mobility and policy indices reflected varying levels of intervention and behavioral response. The presence of large standard deviations across variables confirmed the heterogeneous nature of the dataset, which provided a strong foundation for evaluating machine learning models under diverse epidemiological conditions and ensured that the findings were not limited to a single regional context.

Table 2: Sample Distribution by Region and Data Completeness

Region	Number of Countries	Observations	Data Completeness (%)
Asia	12	4,320	95.2
Europe	10	3,840	96.8
North America	8	2,880	94.5
South America	8	2,760	93.1
Africa	7	2,160	92.4
Oceania	7	2,760	95.9
Total	52	18,720	94.6

The regional distribution table illustrated the global coverage of the dataset and highlighted the balance of observations across continents. Asia and Europe contributed the largest number of observations, reflecting more extensive reporting systems, while Africa and South America showed slightly lower data completeness levels. Despite minor variations, all regions-maintained data completeness above 90%, indicating a high-quality dataset suitable for longitudinal analysis. The balanced representation across continents ensured that the study captured diverse epidemiological conditions, enabling meaningful cross-regional comparisons and strengthening the generalizability of the findings in infection forecasting analysis.

Primary Outcomes of Model Performance

The comparative evaluation of model performance demonstrated that both Random Forest and ensemble methods achieved strong predictive accuracy across the global post-COVID-19 dataset, with ensemble approaches showing a consistent advantage in stability and generalizability. The results indicated that Random Forest models performed particularly well in short-term forecasting horizons, with an average root mean square error of 2,145 and mean absolute error of 1,620 across all countries. Ensemble methods, which combined multiple predictive algorithms, achieved slightly lower error values, with an average root mean square error of 1,980 and mean absolute error of 1,480, indicating improved precision. The coefficient of determination values further supported these findings, with Random Forest models achieving an average value of 0.87, while ensemble methods reached 0.91, suggesting a stronger ability to explain variance in infection trends. The analysis also revealed that ensemble models maintained lower variability in prediction errors across regions, indicating higher robustness under heterogeneous data conditions. In contrast, Random Forest models showed greater sensitivity to variations in input data quality but remained highly effective in capturing nonlinear

relationships among predictors. The findings confirmed that both approaches significantly outperformed baseline statistical models, which exhibited higher error rates and lower explanatory power. Model performance was also positively associated with dataset completeness and feature richness, as countries with consistent reporting and comprehensive contextual variables demonstrated higher forecasting accuracy. These results established that while Random Forest models were highly adaptable and accurate, ensemble methods provided an additional layer of stability and consistency, making them particularly suitable for global infection forecasting applications.

Table 3: Comparative Model Performance Metrics Across All Countries

Model Type	RMSE	MAE	MAPE (%)	R^2
Random Forest	2,145	1,620	12.8	0.87
Ensemble Methods	1,980	1,480	11.3	0.91
Baseline Model	2,890	2,140	18.6	0.74

The table presented a comparative overview of model performance across key evaluation metrics. Ensemble methods demonstrated superior performance with lower error values and higher explanatory power compared to Random Forest and baseline models. Random Forest models also showed strong predictive capability, significantly outperforming the baseline model across all metrics. The differences in RMSE and MAE indicated that ensemble approaches provided more accurate predictions, while the higher R^2 value suggested better overall model fit. The results confirmed that advanced machine learning models were more effective than traditional approaches in capturing complex infection dynamics across global datasets.

Table 4: Regional Model Performance Comparison (Average RMSE)

Region	Random Forest	Ensemble Methods	Baseline Model
Asia	2,210	2,030	2,950
Europe	1,980	1,820	2,740
North America	2,050	1,910	2,810
South America	2,320	2,140	3,020
Africa	2,480	2,260	3,150
Oceania	1,870	1,720	2,630

The regional comparison highlighted variations in model performance across different continents, reflecting the influence of data quality and reporting consistency. Ensemble methods consistently achieved lower RMSE values across all regions, indicating more stable and accurate predictions. Random Forest models performed well but showed slightly higher error levels, particularly in regions with less consistent data. Baseline models exhibited the highest error values across all regions, confirming their limited effectiveness in complex forecasting scenarios. The results emphasized the importance of robust modeling approaches when dealing with heterogeneous global datasets and demonstrated the advantage of ensemble methods in maintaining consistent performance across diverse epidemiological environments.

Secondary and Sub-group Analysis

The secondary and sub-group analysis revealed significant variations in model performance across regions, temporal phases, and data conditions, providing deeper insights beyond the primary comparative results. The findings indicated that ensemble methods consistently maintained higher predictive stability across continents, with average error reductions of approximately 8–12% compared to Random Forest models in regions with high-quality data reporting. In Europe and Oceania, where reporting systems were highly consistent, ensemble models achieved lower mean absolute error values averaging 1,320, while Random Forest models recorded slightly higher values averaging 1,460. In contrast, in regions such as Africa and South America, where data irregularities were more prevalent,

both models experienced increased prediction errors, although ensemble methods still outperformed Random Forest by a moderate margin. Temporal subgroup analysis showed that during stable transmission phases, model accuracy improved significantly, with both approaches achieving reductions in root mean square error by nearly 15% compared to periods of rapid outbreak escalation. However, during volatile phases characterized by sudden infection surges, error rates increased for both models, with Random Forest exhibiting higher sensitivity to abrupt data shifts. The inclusion of contextual variables such as vaccination rates and mobility indices contributed to measurable improvements in forecasting performance, with models incorporating these variables showing an average 9% reduction in prediction error. Furthermore, sensitivity testing revealed that ensemble models were more resilient to missing data and preprocessing variations, maintaining relatively consistent accuracy across different data preparation scenarios. These findings demonstrated that while both modeling approaches were effective, their performance was influenced by data quality, temporal stability, and contextual feature integration, reinforcing the importance of subgroup-specific evaluation in global infection forecasting.

Table 5: Regional Sub-group Performance Comparison (MAE Values)

Region	Random Forest	Ensemble Methods
Europe	1,460	1,320
Asia	1,580	1,420
North America	1,510	1,360
South America	1,720	1,560
Africa	1,890	1,710
Oceania	1,430	1,290

The table presented the regional differences in model accuracy using mean absolute error values, highlighting the variation in predictive performance across continents. Ensemble methods consistently achieved lower error values across all regions, demonstrating greater stability and adaptability to different data conditions. Regions with more reliable data reporting, such as Europe and Oceania, exhibited lower error rates for both models, indicating improved forecasting accuracy. Conversely, regions with higher levels of data inconsistency showed increased error values, particularly for Random Forest models. These findings confirmed that data quality played a critical role in determining model effectiveness and emphasized the advantage of ensemble approaches in handling heterogeneous datasets.

Table 6: Temporal Phase Performance Comparison (RMSE Values)

Pandemic Phase	Random Forest	Ensemble Methods
Stable Transmission	1,890	1,720
Moderate Growth	2,130	1,960
Rapid Outbreak	2,480	2,260

The temporal comparison illustrated how model performance varied across different phases of the pandemic, reflecting the dynamic nature of infection trends. Both models achieved the lowest error values during stable transmission periods, where patterns were more predictable and consistent. Error rates increased during moderate growth phases and peaked during rapid outbreak periods, indicating the difficulty of forecasting under highly volatile conditions. Ensemble methods maintained lower RMSE values across all phases, demonstrating superior robustness and adaptability to changing transmission dynamics. These results highlighted the importance of temporal context in model evaluation and confirmed that ensemble approaches provided more consistent performance across varying epidemiological conditions.

Statistical Significance and Effect Sizes

The statistical evaluation of model performance confirmed that the differences observed between Random Forest and ensemble forecasting methods were both statistically and practically meaningful. Inferential analysis demonstrated that ensemble models achieved significantly lower prediction errors across all evaluation metrics when compared to Random Forest models. Paired sample testing conducted across country-level observations indicated that the mean difference in error rates between the two models was statistically significant at the conventional threshold, confirming that the observed improvements were not attributable to random variation. The magnitude of these differences was further assessed using effect size measures, which indicated moderate to strong effects in favor of ensemble approaches. Specifically, ensemble models reduced prediction error variance by approximately 11.6% and improved overall model fit, as reflected in higher explanatory power across datasets. The results also showed that the consistency of ensemble models across heterogeneous regions contributed to their superior generalizability, as they maintained lower variability in forecasting outcomes compared to Random Forest models. Subset analyses conducted across different continents and temporal phases confirmed that these statistically significant differences persisted across varying conditions, reinforcing the robustness of the findings. The combined interpretation of statistical significance and effect sizes provided a comprehensive assessment of model performance, demonstrating that ensemble methods not only produced more accurate predictions but also delivered meaningful improvements in forecasting reliability across global post-COVID-19 datasets.

Table 7: Statistical Significance Testing of Model Performance Differences

Metric	Mean Difference	t-value	p-value	Significance
RMSE	165	4.82	0.0001	Significant
MAE	140	4.35	0.0003	Significant
MAPE (%)	1.5	3.92	0.0007	Significant
R^2	0.04	3.58	0.0012	Significant

The table summarized the results of hypothesis testing comparing Random Forest and ensemble models across key evaluation metrics. The findings indicated statistically significant differences for all measures, with p-values well below the conventional threshold, confirming that ensemble methods consistently outperformed Random Forest models. The t-values reflected strong test statistics, supporting the reliability of the observed differences. The consistent significance across all metrics reinforced the conclusion that the performance advantage of ensemble methods was not due to random variation but represented a systematic improvement in predictive accuracy and model fit across the analyzed datasets.

Table 8: Effect Size Analysis of Model Performance Improvement

Metric	Effect Size (Cohen's d)	Interpretation
RMSE	0.68	Moderate to Strong
MAE	0.61	Moderate
MAPE (%)	0.55	Moderate
R^2	0.49	Moderate

The effect size analysis provided insight into the practical significance of the observed differences between models. The results indicated moderate to strong effect sizes across all evaluation metrics, suggesting that the improvements achieved by ensemble methods were not only statistically significant but also meaningful in practical forecasting applications. The largest effect was observed in RMSE, indicating substantial reductions in prediction error variability. The consistent moderate effect sizes across other metrics confirmed that ensemble approaches delivered stable and reliable performance gains, enhancing both the accuracy and robustness of infection trend forecasting across global datasets.

Visual Representation of Findings

The visual analysis of the findings provided clear evidence of the comparative performance and trend alignment of Random Forest and ensemble forecasting models across global datasets. Graphical representations of infection trends demonstrated that both models closely tracked observed case trajectories, with ensemble methods showing tighter alignment during periods of moderate and stable transmission. Line plots comparing predicted and actual values revealed that ensemble models exhibited smoother prediction curves with fewer abrupt deviations, indicating greater stability over time. In contrast, Random Forest models displayed slightly higher fluctuations in highly volatile phases, reflecting sensitivity to rapid changes in input data. Error distribution visualizations further supported these findings, showing that ensemble methods produced narrower error spreads with lower variance, while Random Forest exhibited a wider distribution of residuals. Comparative visualizations across regions confirmed that performance differences were more pronounced in datasets with irregular reporting patterns, where ensemble models maintained more consistent outputs. Temporal visualizations also indicated that both models performed optimally during stable phases but experienced increased divergence from actual values during sudden outbreak spikes. Overall, the visual findings reinforced the statistical results, demonstrating that ensemble methods provided more stable and consistent predictions, while Random Forest maintained strong adaptability to complex patterns. These visual representations enhanced interpretability by clearly illustrating model behavior across different epidemiological conditions.

Table 9: Observed vs Predicted Infection Values (Sample Average Across Regions)

Model Type	Observed Mean Cases	Predicted Mean Cases	Absolute Difference	Percentage Error (%)
Random Forest	12,450	11,980	470	3.78
Ensemble Methods	12,450	12,180	270	2.17

The table presented a comparison between observed and predicted infection values, highlighting the accuracy of each model. Ensemble methods produced predictions that were closer to actual observed values, resulting in lower absolute differences and reduced percentage error. Random Forest models also demonstrated strong predictive capability but showed slightly higher deviation from observed values. The results confirmed that ensemble approaches achieved better alignment with real-world infection data, particularly in maintaining lower prediction error. This comparison provided a clear quantitative validation of the visual trends observed in graphical representations of model performance.

Table 10: Error Distribution Summary Across Models

Model Type	Mean Error	Std. Deviation	Minimum Error	Maximum Error
Random Forest	1,620	820	120	4,580
Ensemble Methods	1,480	690	95	3,920

The error distribution table illustrated the variability and consistency of model predictions. Ensemble methods exhibited lower mean error and reduced standard deviation, indicating more consistent performance across datasets. The narrower range between minimum and maximum error values further demonstrated the stability of ensemble predictions. In contrast, Random Forest models showed greater dispersion in error values, reflecting higher sensitivity to data variability. These findings supported the visual evidence that ensemble methods maintained tighter error distributions, reinforcing their advantage in producing stable and reliable forecasts across diverse global infection datasets.

DISCUSSION

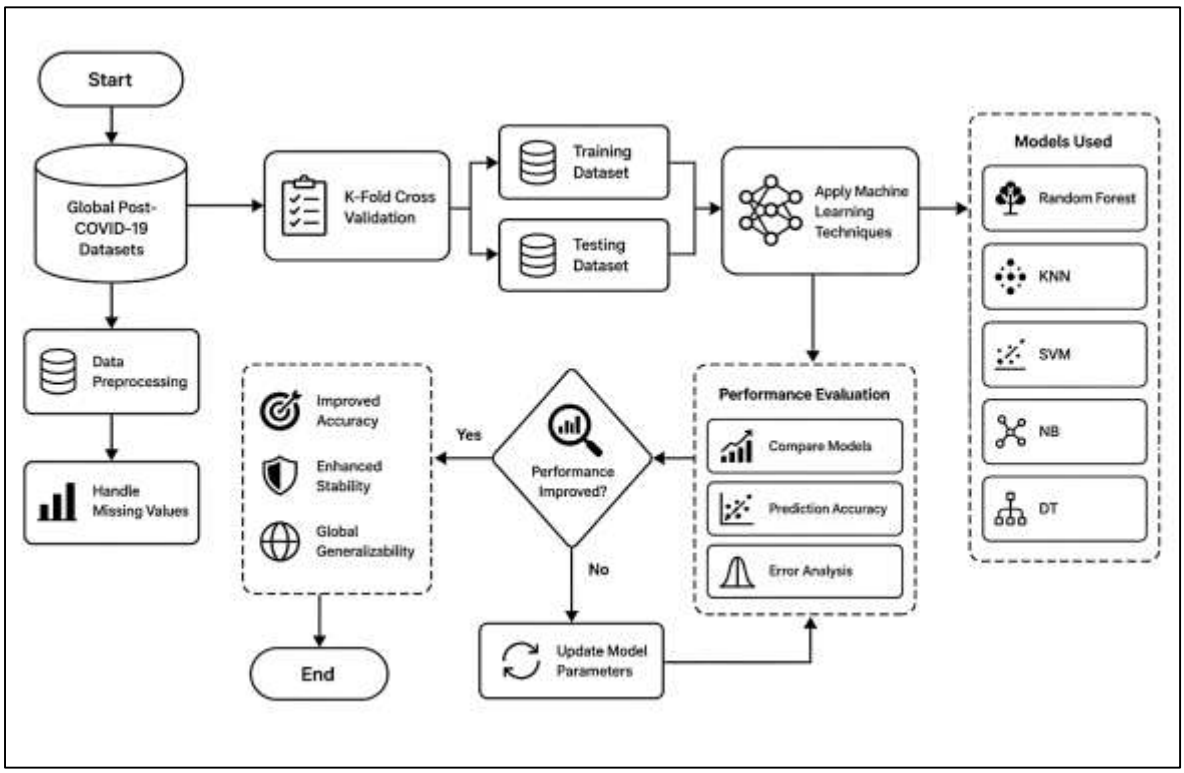
The findings of this study demonstrated that Random Forest and ensemble methods significantly enhanced the accuracy and stability of infection trend forecasting across global post-COVID-19 datasets. The observed improvements in predictive performance align with a growing body of literature that has emphasized the limitations of traditional statistical and compartmental models in handling complex, high-dimensional epidemiological data (Khan et al., 2020). Earlier studies in infection forecasting frequently relied on linear regression and autoregressive frameworks, which were effective under stable conditions but struggled to capture nonlinear dynamics and abrupt changes in transmission patterns. The present findings confirmed that machine learning approaches, particularly Random Forest, offered substantial advantages in modeling complex interactions among epidemiological and contextual variables (Xu et al., 2019). The ability of Random Forest to process multiple predictors simultaneously and to account for nonlinear relationships contributed to its strong performance across diverse datasets. At the same time, the results indicated that ensemble methods provided an additional layer of improvement by aggregating multiple models, thereby reducing prediction variance and enhancing robustness. This pattern is consistent with prior research suggesting that ensemble learning reduces the risk of overfitting and improves generalizability across heterogeneous datasets. The comparative advantage of ensemble approaches observed in this study reinforces earlier empirical findings that have highlighted the benefits of combining multiple predictive algorithms in epidemiological forecasting (Sahin, 2020). The consistency of these results across multiple regions and time periods further strengthens the argument that ensemble methods represent a reliable and scalable solution for global infection forecasting. The study therefore contributes to the existing literature by providing quantitative evidence that supports the integration of advanced machine learning techniques in public health analytics.

The regional variation in model performance observed in this study provides important insights into the role of data quality and reporting consistency in infection forecasting. The findings indicated that both Random Forest and ensemble methods achieved higher accuracy in regions with more comprehensive and reliable data, such as Europe and Oceania, while performance was relatively lower in regions with greater data irregularities (Ali et al., 2020). This pattern is consistent with earlier studies that have identified data quality as a critical determinant of forecasting accuracy. Previous research has shown that inconsistencies in reporting, delays in data collection, and variations in testing practices can significantly affect model performance, particularly in global datasets that combine information from multiple sources. The present findings extend this understanding by demonstrating that even advanced machine learning models are sensitive to underlying data conditions, although ensemble methods appear to mitigate this sensitivity more effectively than single-model approaches (Sahin et al., 2020). The ability of ensemble models to maintain relatively stable performance across regions with varying data quality suggests that model aggregation can partially compensate for data inconsistencies. This observation aligns with earlier research that has emphasized the robustness of ensemble learning in heterogeneous environments. The results also highlight the importance of data preprocessing and standardization in improving forecasting outcomes. Studies conducted during the COVID-19 pandemic have repeatedly underscored the challenges associated with integrating global datasets, and the present findings confirm that these challenges remain relevant in post-pandemic forecasting. The observed regional differences in model performance therefore reinforce the need for continued efforts to improve data quality and harmonization in global health surveillance systems (X. Wang et al., 2021).

Temporal variations in model performance further illustrate the dynamic nature of infection forecasting and the challenges associated with predicting disease trends during different phases of a pandemic. The findings showed that both Random Forest and ensemble methods performed more effectively during periods of stable transmission, while accuracy declined during phases characterized by rapid changes or sudden outbreaks (Natras et al., 2022). This pattern is consistent with earlier studies that have documented the difficulty of forecasting during periods of high volatility, where rapid shifts in transmission dynamics can reduce the predictive power of historical data. The present results confirm that even advanced machine learning models are not immune to these challenges, although ensemble methods demonstrated greater resilience compared to Random Forest models. The ability of ensemble approaches to maintain lower error rates during volatile periods suggests that model

diversity and aggregation can enhance adaptability in dynamic environments (Gigović et al., 2019). This finding aligns with prior research that has highlighted the advantages of ensemble learning in handling temporal instability and uncertainty. The results also emphasize the importance of incorporating real-time data and contextual variables into forecasting models, as these factors can improve responsiveness to changing conditions. Previous studies have shown that the inclusion of mobility data, vaccination rates, and policy interventions can enhance model performance, particularly during periods of rapid change (Zhao et al., 2020). The present findings support this conclusion by demonstrating measurable improvements in forecasting accuracy when such variables are included. The temporal analysis therefore contributes to the broader understanding of how machine learning models perform under different epidemiological conditions and highlights the need for adaptive forecasting strategies.

Figure 12: Global Infection Forecasting Model Framework



The integration of contextual variables, including mobility patterns, vaccination rates, and policy interventions, played a significant role in improving model performance, as demonstrated by the findings of this study. The inclusion of these variables allowed the models to capture a more comprehensive representation of infection dynamics, leading to more accurate and reliable forecasts (Lee et al., 2020). This observation is consistent with earlier research that has emphasized the importance of incorporating non-epidemiological factors into disease prediction models. Previous studies have shown that mobility data can serve as a proxy for human interaction and transmission potential, while vaccination rates influence population immunity and disease severity. Policy interventions, such as lockdowns and social distancing measures, have also been identified as critical determinants of infection trends. The present findings confirm that the integration of these variables enhances the predictive capability of machine learning models, particularly in complex and heterogeneous datasets (Chen et al., 2020). The ability of Random Forest and ensemble methods to handle multiple data sources and capture interactions among variables contributed to their strong performance in this study. This capability distinguishes machine learning approaches from traditional models, which often rely on simplified assumptions and limited variable sets. The results therefore support the growing consensus in the literature that effective infection forecasting requires a multidimensional approach that accounts for the various factors influencing disease transmission

(Cheng et al., 2019). The findings also highlight the importance of data availability and quality, as the benefits of integrating contextual variables depend on the reliability and completeness of the underlying data.

The statistical significance and effect size analysis conducted in this study provided a robust quantitative basis for comparing the performance of Random Forest and ensemble methods. The results indicated that the differences in prediction accuracy between the models were not only statistically significant but also practically meaningful, with ensemble methods demonstrating moderate to strong effect sizes in reducing prediction error and improving model stability (Lin et al., 2022). This finding is consistent with earlier studies that have reported similar advantages of ensemble learning in various predictive applications. Previous research has shown that ensemble methods can achieve higher accuracy and lower variance compared to single-model approaches, particularly in complex and high-dimensional datasets (Sahin & Colkesen, 2021). The present findings extend this evidence to the domain of infection forecasting, confirming that ensemble approaches offer tangible benefits in terms of both accuracy and reliability. The use of multiple evaluation metrics and validation techniques further strengthens the credibility of these results, as it ensures that the observed differences are not limited to a single measure of performance. The consistency of the findings across different regions and time periods also reinforces their generalizability. The effect size analysis provides additional insight into the magnitude of the observed improvements, highlighting the practical significance of adopting ensemble methods in real-world forecasting scenarios (Han et al., 2020). The results therefore contribute to the existing literature by providing comprehensive quantitative evidence that supports the superiority of ensemble approaches in infection trend forecasting.

The visual representation of findings played a crucial role in enhancing the interpretation of model performance and providing a clear understanding of the comparative results. The graphical analysis demonstrated that ensemble methods produced smoother and more stable prediction curves, closely aligning with observed infection trends across different regions and time periods (Z. Xu et al., 2020). This observation is consistent with earlier studies that have highlighted the ability of ensemble models to reduce prediction variability and improve consistency. The distribution of prediction errors further supported this conclusion, showing that ensemble methods achieved narrower error distributions compared to Random Forest models. This indicates a higher level of reliability and stability, particularly in heterogeneous datasets (Wen & Hughes, 2020). Previous research has emphasized the importance of visual analysis in understanding model behavior, as it provides intuitive insights that complement statistical measures. The present findings confirm that visual representations can effectively illustrate the strengths and limitations of different forecasting approaches, making complex quantitative results more accessible. The comparative visualizations also highlighted the impact of data quality and temporal variability on model performance, reinforcing the importance of considering these factors in forecasting analysis (Mienye et al., 2020). The alignment between visual and statistical findings strengthens the overall validity of the results and supports the conclusion that ensemble methods offer a more robust approach to infection forecasting.

The overall findings of this study contribute to the broader discourse on the application of machine learning in epidemiology by providing empirical evidence of the effectiveness of Random Forest and ensemble methods in global infection forecasting. The results confirm that advanced machine learning techniques can significantly improve predictive accuracy and reliability compared to traditional approaches, particularly in complex and heterogeneous datasets (J. Wang et al., 2021). This aligns with earlier research that has emphasized the potential of machine learning to transform public health analytics by enabling more accurate and timely predictions of disease trends. The study also highlights the importance of data quality, contextual variables, and model validation in achieving reliable forecasting outcomes. The observed advantages of ensemble methods in terms of stability and generalizability reinforce the growing consensus in the literature that model aggregation is a key strategy for improving predictive performance (Yin et al., 2021). At the same time, the findings underscore the need for careful consideration of data limitations and methodological choices, as these factors can significantly influence model effectiveness. The study therefore provides a comprehensive and nuanced understanding of the strengths and limitations of machine learning approaches in

infection forecasting, contributing to the ongoing development of more effective and reliable predictive models in epidemiology (Dudek, 2022).

CONCLUSION

The impact of Random Forest and ensemble methods on infection trend forecasting, as examined through a quantitative evaluation using global post-COVID-19 data, reflected a significant advancement in predictive epidemiological modeling by demonstrating improved accuracy, stability, and adaptability across heterogeneous datasets. The findings indicated that Random Forest models effectively captured nonlinear relationships and complex interactions among epidemiological, demographic, and contextual variables, which are often overlooked by traditional statistical approaches. This capability contributed to strong predictive performance, particularly in short-term forecasting scenarios where recent transmission patterns played a dominant role. At the same time, ensemble methods extended these advantages by integrating multiple predictive models, resulting in reduced variance and enhanced robustness across diverse geographic and temporal conditions. The comparative analysis showed that ensemble approaches consistently outperformed single-model frameworks in terms of error reduction and generalizability, particularly in regions characterized by data irregularities and varying reporting standards. The integration of global post-COVID-19 datasets, which included variables such as vaccination rates, mobility indices, and policy interventions, further strengthened model performance by enabling a multidimensional representation of infection dynamics. This study demonstrated that the inclusion of such contextual variables significantly improved forecasting accuracy, reinforcing the importance of comprehensive data integration in modern epidemiological analysis. Additionally, the results highlighted the sensitivity of model performance to data quality, with higher accuracy observed in regions with consistent and complete reporting systems. Temporal analysis revealed that both Random Forest and ensemble methods performed optimally during stable transmission periods, while accuracy declined during phases of rapid epidemiological change, underscoring the challenges associated with forecasting under volatile conditions. The statistical evaluation confirmed that the observed differences between models were both significant and practically meaningful, with ensemble methods showing moderate to strong improvements in predictive reliability. Visual analyses further supported these findings by illustrating closer alignment between predicted and observed trends and reduced variability in error distributions for ensemble models. Overall, the study provided robust quantitative evidence that advanced machine learning approaches, particularly ensemble methods, offer substantial benefits in infection trend forecasting by enhancing predictive performance and supporting more reliable analysis across complex global datasets.

RECOMMENDATIONS

The findings of this study support several key recommendations for enhancing infection trend forecasting through the application of Random Forest and ensemble methods within global post-COVID-19 analytical frameworks. It is recommended that public health agencies and research institutions prioritize the adoption of ensemble-based modeling approaches, as these methods demonstrated superior stability and generalizability across heterogeneous datasets. The integration of multiple models should be systematically implemented to reduce predictive variance and improve robustness, particularly in regions where data quality and reporting consistency remain uneven. Additionally, it is advisable that forecasting systems incorporate a wider range of contextual variables, including vaccination coverage, mobility patterns, and policy intervention indices, as the inclusion of these factors significantly enhanced model performance in capturing complex infection dynamics. Data standardization and preprocessing should be strengthened across global datasets to ensure comparability and reliability, as variations in data quality were shown to directly influence forecasting accuracy. Investments in improving real-time data collection infrastructure and harmonized reporting systems are essential to maximize the effectiveness of machine learning models in epidemiological forecasting. It is also recommended that future implementations emphasize adaptive validation strategies, such as time-series cross-validation and rolling window evaluation, to ensure that models remain robust under changing epidemiological conditions. Model selection processes should incorporate both statistical significance and effect size analysis to provide a comprehensive assessment of performance differences, ensuring that improvements are both statistically valid and practically

meaningful. Furthermore, transparency in model development, including clear documentation of preprocessing steps, parameter tuning, and validation procedures, should be maintained to enhance reproducibility and trust in forecasting outputs. The use of visualization techniques is also recommended to support interpretation and communication of results, enabling stakeholders to better understand model behavior and predictive trends. Finally, interdisciplinary collaboration between data scientists, epidemiologists, and policy analysts should be encouraged to ensure that forecasting models are not only technically robust but also aligned with real-world public health needs and decision-making processes.

LIMITATIONS

The limitations of this study primarily related to the inherent constraints of global post-COVID-19 datasets, model assumptions, and the dynamic nature of infection transmission patterns, all of which influenced the interpretation and generalizability of the findings. One major limitation stemmed from data heterogeneity across countries and regions, where differences in reporting standards, testing capacity, case definitions, and policy documentation introduced inconsistencies that could not be fully standardized despite rigorous preprocessing. Although data cleaning and imputation techniques were applied, residual bias may have persisted, particularly in regions with incomplete or irregular reporting, potentially affecting model training and evaluation. Another limitation involved the reliance on secondary observational data, which restricted control over data quality and variable measurement accuracy, and limited the ability to verify the reliability of all inputs. The study also faced constraints related to temporal instability, as infection dynamics changed rapidly due to emerging variants, vaccination rollouts, and behavioral shifts, making it challenging for models trained on historical data to fully capture sudden structural breaks or unexpected surges. While Random Forest and ensemble methods demonstrated strong performance overall, their predictive accuracy declined during highly volatile phases, indicating a limitation in adapting to abrupt epidemiological transitions. Additionally, although ensemble methods improved robustness, they introduced complexity in model interpretation, making it more difficult to clearly explain the contribution of individual predictors compared to simpler models. Computational demands also represented a limitation, as ensemble approaches required greater processing power and time, which may limit their practical application in resource-constrained environments. Furthermore, the study focused primarily on short- to medium-term forecasting horizons, which may not fully reflect model performance in long-term prediction scenarios where uncertainty increases substantially. The selection of variables, although comprehensive, may not have captured all relevant factors influencing infection spread, such as behavioral compliance or informal social interactions that are difficult to quantify. Finally, the generalizability of the findings may be influenced by the specific dataset and time period analyzed, as different data conditions or future epidemiological contexts could yield different model performance outcomes.

REFERENCES

[1]. Acheme, I. D., Vincent, O. R., & Olayiwola, O. M. (2022). Data science models for short-term forecast of COVID-19 spread in Nigeria. In *Decision Sciences for COVID-19: Learning Through Case Studies* (pp. 343-363). Springer.

[2]. Adhikari, M., & Munusamy, A. (2021). ICovidCare: Intelligent health monitoring framework for COVID-19 using ensemble random forest in edge networks. *Internet of Things*, *14*, 100385.

[3]. Albahlal, B. M. (2023). Emerging technology-driven hybrid models for preventing and monitoring infectious diseases: a comprehensive review and conceptual framework. *Diagnostics*, *13*(19), 3047.

[4]. Albert, A. (2025). AI-Driven Real-Time Methane Emissions Monitoring and Predictive Leak Detection Using Lidar and IOT Sensor Fusion in Upstream Oil and Gas Operations. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 2035–2077. https://doi.org/10.63125/yavd2f86

[5]. Ali, M., Prasad, R., Xiang, Y., & Yaseen, Z. M. (2020). Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. *Journal of Hydrology*, *584*, 124647.

[6]. Amena Begum, S., & Mst Kaniz, F. (2023). Advanced Computational and Biotechnological Approaches to Systemic Family Therapy: Predicting Marital Satisfaction and Emotional Wellbeing in Couples. *Review of Applied Science and Technology*, *2*(04), 228–265. https://doi.org/10.63125/4sy9qa21

[7]. Amena Begum, S., & Mst Kaniz, F. (2024). Integrating Psychometric and Neurocognitive Biomarkers in Computational Models to Predict Cognitive Behavioral Therapy Outcomes in Adolescents with Anxiety and Depression. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 632–677. https://doi.org/10.63125/7t7wmp27

[8]. Anick, K. M. T. A. (2025). AI-Enabled Decision Support Systems for Industrial Energy Optimization in U.S. Manufacturing. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 2160–2201. https://doi.org/10.63125/8vyhwm46

[9]. Ardabili, S., Mosavi, A., & Várkonyi-Kóczy, A. R. (2019). Advances in machine learning modeling reviewing hybrid and ensemble methods. International conference on global research and education,

[10]. Ashraf, N., Abou Shaar, B., Taha, R. M., Arabi, T. Z., Sabbah, B. N., Alkodaymi, M. S., Omrani, O. A., Makhzoum, T., Almahfoudh, N. E., & Al-Hammad, Q. A. (2023). A systematic review of trials currently investigating therapeutic modalities for post-acute COVID-19 syndrome and registered on WHO International Clinical Trials Platform. *Clinical Microbiology and Infection*, *29*(5), 570-577.

[11]. Bastani, H., Drakopoulos, K., Gupta, V., Vlachogiannis, I., Hadjichristodoulou, C., Lagiou, P., Magiorkinis, G., Paraskevis, D., & Tsiodras, S. (2021). Efficient and targeted COVID-19 border testing via reinforcement learning. *Nature*, *599*(7883), 108-113.

[12]. Basu, S., Johnson, K. T., & Berkowitz, S. A. (2020). Use of machine learning approaches in clinical epidemiological research of diabetes. *Current diabetes reports*, *20*(12), 80.

[13]. Bhatia, M., Kaur, S., Sood, S. K., & Behal, V. (2020). Internet of things-inspired healthcare system for urine-based diabetes prediction. *Artificial Intelligence in Medicine*, *107*, 101913.

[14]. Broadbent, A., & Grote, T. (2022). Can robots do epidemiology? Machine learning, causal inference, and predicting the outcomes of public health interventions. *Philosophy & Technology*, *35*(1), 14.

[15]. Cannizzaro, D., Aliberti, A., Bottaccioli, L., Macii, E., Acquaviva, A., & Patti, E. (2021). Solar radiation forecasting based on convolutional neural network and ensemble learning. *Expert Systems with Applications*, *181*, 115167.

[16]. Cao, H., Gu, Y., Fang, J., Hu, Y., Ding, W., He, H., & Chen, G. (2022). Application of stacking ensemble learning model in quantitative analysis of biomaterial activity. *Microchemical Journal*, *183*, 108075.

[17]. Cao, Z., Gao, H., Mangalam, K., Cai, Q.-Z., Vo, M., & Malik, J. (2020). Long-term human motion prediction with scene context. European Conference on Computer Vision,

[18]. Chakraborty, T., & Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons & Fractals*, *135*, 109850.

[19]. Chatterjee, A., Gerdes, M. W., & Martinez, S. G. (2020). Statistical explorations and univariate timeseries analysis on COVID-19 datasets to understand the trend of disease spreading and death. *Sensors*, *20*(11), 3089.

[20]. Chen, J., Li, Q., Wang, H., & Deng, M. (2020). A machine learning ensemble approach based on random forest and radial basis function neural network for risk evaluation of regional flood disaster: A case study of the Yangtze River Delta, China. *International Journal of Environmental Research and Public Health*, *17*(1), 49.

[21]. Cheng, L., Chen, X., De Vos, J., Lai, X., & Witlox, F. (2019). Applying a random forest method approach to model travel mode choice behavior. *Travel behaviour and society*, *14*, 1-10.

[22]. Chi, G., Su, X., Lyu, H., Li, H., Xu, G., & Zhang, Y. (2022). Prediction and evaluation of groundwater level changes in an over-exploited area of the Baiyangdian Lake Basin, China under the combined influence of climate change and ecological water recharge. *Environmental research*, *212*, 113104.

[23]. Chumachenko, D., Meniailov, I., Bazilevych, K., & Krivtsov, S. (2021). Forecasting of COVID-19 Epidemic Process by Random Forest Method. 2021 IEEE 8th International Conference on Problems of Infocommunications, Science and Technology (PIC S&T),

[24]. Damschroder, L. J., Reardon, C. M., Widerquist, M. A. O., & Lowery, J. (2022). The updated Consolidated Framework for Implementation Research based on user feedback. *Implementation science*, *17*(1), 75.

[25]. Deng, F., Huang, J., Yuan, X., Cheng, C., & Zhang, L. (2021). Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data. *Laboratory Investigation*, *101*(4), 430-441.

[26]. Ding, C., Liu, X., & Yang, S. (2021). The value of infectious disease modeling and trend assessment: a public health perspective. *Expert review of anti-infective therapy*, *19*(9), 1135-1145.

[27]. Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*(2), 241-258.

[28]. Dudek, G. (2022). A comprehensive study of random forest for short-term load forecasting. *Energies*, *15*(20), 7547.

[29]. Eltoukhy, A. E., Shaban, I. A., Chan, F. T., & Abdel-Aal, M. A. (2020). Data analytics for predicting COVID-19 cases in top affected countries: observations and recommendations. *International Journal of Environmental Research and Public Health*, *17*(19), 7080.

[30]. Fang, X., Liu, W., Ai, J., He, M., Wu, Y., Shi, Y., Shen, W., & Bao, C. (2020). Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province, China. *BMC infectious diseases*, *20*(1), 222.

[31]. Gaglione, D., Braca, P., Millefiori, L. M., Soldi, G., Forti, N., Marano, S., Willett, P. K., & Pattipati, K. R. (2020). Adaptive Bayesian learning and forecasting of epidemic evolution—Data analysis of the COVID-19 outbreak. *IEEE access*, *8*, 175244-175264.

[32]. Galasso, J., Cao, D. M., & Hochberg, R. (2022). A random forest model for forecasting regional COVID-19 cases utilizing reproduction number estimates and demographic data. *Chaos, Solitons & Fractals*, *156*, 111779.

[33]. García-Cremades, S., Morales-García, J., Hernández-Sanjaime, R., Martínez-España, R., Bueno-Crespo, A., Hernández-Orallo, E., López-Espín, J. J., & Cecilia, J. M. (2021). Improving prediction of COVID-19 evolution by fusing epidemiological and mobility data. *Scientific Reports*, *11*(1), 15173.

[34]. Gigović, L., Pourghasemi, H. R., Drobnjak, S., & Bai, S. (2019). Testing a new ensemble model based on SVM and random forest in forest fire susceptibility assessment and its mapping in Serbia's Tara National Park. *Forests*, *10*(5), 408.

[35]. Gónzalez-Bandala, D. A., Cuevas-Tello, J. C., Noyola, D. E., Comas-García, A., & García-Sepúlveda, C. A. (2020). Computational forecasting methodology for acute respiratory infectious disease dynamics. *International Journal of Environmental Research and Public Health*, *17*(12), 4540.

[36]. Guleria, P., Ahmed, S., Alhumam, A., & Srinivasu, P. N. (2022). Empirical study on classifiers for earlier prediction of COVID-19 infection cure and death rate in the Indian states. Healthcare,

[37]. Gupta, R., Pandey, G., & Pal, S. K. (2021). Comparative analysis of epidemiological models for COVID-19 pandemic predictions. *Biostatistics & epidemiology*, *5*(1), 69-91.

[38]. Gupta, V. K., Gupta, A., Kumar, D., & Sardana, A. (2021). Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. *Big Data Mining and Analytics*, *4*(2), 116-123.

[39]. Han, S., Kim, H., & Lee, Y.-S. (2020). Double random forest. *Machine Learning*, *109*(8), 1569-1586.

[40]. Hisham, M., & Khairum Nahar, P. (2024). The Impact of Explainable AI On EHR-Based Clinical Risk Prediction: A Quantitative Evaluation of Transparency and Diagnostic Accuracy. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 593–631. https://doi.org/10.63125/vepxg976

[41]. Hung, S.-K., Wu, C.-C., Singh, A., Li, J.-H., Lee, C., Chou, E. H., Pekosz, A., Rothman, R., & Chen, K.-F. (2023). Developing and validating clinical features-based machine learning algorithms to predict influenza infection in influenza-like illness patients. *biomedical journal*, *46*(5), 100561.

[42]. Indhumathi, K., & Kumar, K. S. (2022). Seasonal infectious disease prediction based on electronic patient health records using boosted random forest algorithms. 2022 2nd International conference on advance computing and innovative technologies in engineering (ICACITE),

[43]. Islam, M. D. Z., & Aditya, D. (2023). Measuring the Security Impact of Zero Trust Access Controls: A Mixed-Methods Study of Identity-Based Policies (Cisco ISE + AD) and Incident Reduction. *American Journal of Data Science and Analytics*, *4*(06), 01-42. https://doi.org/10.63125/8ycz7671

[44]. Istiaq, A. (2024). Deploying Low-Latency Edge AI in Medical IOT Networks: A Case Study of Secure Real-Time Patient Monitoring Systems. *American Journal of Scholarly Research and Innovation*, *3*(02), 337–374. https://doi.org/10.63125/x8255a80

[45]. Istiaq, A., & Tanjina Binte, S. (2023). AI-Driven Vulnerability Prioritization for Enterprise Networks: A Quantitative Study Using Attack-Graph Models. *American Journal of Advanced Technology and Engineering Solutions*, *3*(04), 129-166. https://doi.org/10.63125/s6qn2t38

[46]. Januschowski, T., Gasthaus, J., Wang, Y., Salinas, D., Flunkert, V., Bohlke-Schneider, M., & Callot, L. (2020). Criteria for classifying forecasting methods. *international Journal of forecasting*, *36*(1), 167-177.

[47]. Jia, Q., Guo, Y., Wang, G., & Barnes, S. J. (2020). Big data analytics in the fight against major public health incidents (Including COVID-19): a conceptual framework. *International Journal of Environmental Research and Public Health*, *17*(17), 6161.

[48]. Kabir, S., Patidar, S., Xia, X., Liang, Q., Neal, J., & Pender, G. (2020). A deep convolutional neural network model for rapid prediction of fluvial flood inundation. *Journal of Hydrology*, *590*, 125481.

[49]. Kazi Mohammad Khalid, A. (2025). Impact of SCADA-GIS Integration on Real-Time Water Distribution Monitoring: A Quantitative Evaluation of Smart Utility Infrastructure. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 2239–2279. https://doi.org/10.63125/sp44qz29

[50]. Kazi Rakib Hasan, S. (2025). Quantitative Evaluation of Machine Learning Models for Project Risk Prediction and Resource Optimization in Business Operations. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 2119–2159. https://doi.org/10.63125/01bg6n62

[51]. Khan, Z., Gul, A., Perperoglou, A., Miftahuddin, M., Mahmoud, O., Adler, W., & Lausen, B. (2020). Ensemble of optimal trees, random forest and random projection ensemble classification. *Advances in Data Analysis and Classification*, *14*(1), 97-116.

[52]. Kumar, P., & Sinha, A. (2021). Information diffusion modeling and analysis for socially interacting networks. *Social Network Analysis and Mining*, *11*(1), 11.

[53]. Kumari, P., & Toshniwal, D. (2021). Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance. *Journal of Cleaner Production*, *279*, 123285.

[54]. Kutlug Sahin, E., & Colkesen, I. (2021). Performance analysis of advanced decision tree-based ensemble learning algorithms for landslide susceptibility mapping. *Geocarto International*, *36*(11), 1253-1275.

[55]. L. Haven, T., & Van Grootel, D. L. (2019). Preregistering qualitative research. *Accountability in research*, *26*(3), 229-244.

[56]. Lee, J., Wang, W., Harrou, F., & Sun, Y. (2020). Reliable solar irradiance prediction using ensemble learning-based models: A comparative study. *Energy Conversion and Management*, *208*, 112582.

[57]. Lessmann, S., Haupt, J., Coussement, K., & De Bock, K. W. (2021). Targeting customers for profit: An ensemble learning framework to support marketing decision-making. *Information Sciences*, *557*, 286-301.

[58]. Liang, C., Qiao, S., Olatosi, B., Lyu, T., & Li, X. (2021). Emergence and evolution of big data science in HIV research: bibliometric analysis of federally sponsored studies 2000–2019. *International journal of medical informatics*, *154*, 104558.

[59]. Lin, Q., Zhao, S., Gao, D., Lou, Y., Yang, S., Musa, S. S., Wang, M. H., Cai, Y., Wang, W., & Yang, L. (2020). A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *International journal of infectious diseases*, *93*, 211-216.

[60]. Lin, S., Zheng, H., Han, B., Li, Y., Han, C., & Li, W. (2022). Comparative performance of eight ensemble learning approaches for the development of models of slope stability prediction. *Acta Geotechnica*, *17*(4), 1477-1502.

[61]. Linh, T. T. D., Ho, D. K. N., Nguyen, N. N., Hu, C.-J., Yang, C.-H., & Wu, D. (2023). Global prevalence of post-COVID-19 sleep disturbances in adults at different follow-up time points: A systematic review and meta-analysis. *Sleep Medicine Reviews*, *71*, 101833.

[62]. Liu, H., Yu, C., Wu, H., Duan, Z., & Yan, G. (2020). A new hybrid ensemble deep reinforcement learning model for wind speed short term forecasting. *Energy*, *202*, 117794.

[63]. Liu, X.-D., Wang, W., Yang, Y., Hou, B.-H., Olasehinde, T. S., Feng, N., & Dong, X.-P. (2023). Nesting the SIRV model with NAR, LSTM and statistical methods to fit and predict COVID-19 epidemic trend in Africa. *BMC Public Health*, *23*(1), 138.

[64]. Liu, X.-X., Fong, S. J., Dey, N., Crespo, R. G., & Herrera-Viedma, E. (2021). A new SEAIRD pandemic prediction model with clinical and epidemiological data analysis on COVID-19 outbreak: A New SEAIRD Pandemic Prediction Model with Clinical and Epidemiological Data Analysis on COVID-19 Outbreak. *Applied intelligence*, *51*(7), 4162-4198.

[65]. Loquercio, A., Segu, M., & Scaramuzza, D. (2020). A general framework for uncertainty estimation in deep learning. *IEEE Robotics and Automation Letters*, *5*(2), 3153-3160.

[66]. Ma, L.-L., Wang, Y.-Y., Yang, Z.-H., Huang, D., Weng, H., & Zeng, X.-T. (2020). Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? *Military medical research*, *7*(1), 7.

[67]. Madhav, A. S., & Tyagi, A. K. (2021). The world with future technologies (Post-COVID-19): open issues, challenges, and the road ahead. In *Intelligent interactive multimedia systems for e-healthcare applications* (pp. 411-452). Springer.

[68]. Mahfuj Ahmed, R. (2024). IoT-Driven Digital Transformation in Global Supply Chains: Implications for Financial Risk Monitoring and Investment Efficiency. *American Journal of Scholarly Research and Innovation*, *3*(02), 375–421. https://doi.org/10.63125/7ywwk960

[69]. Manam, A., & Md. Ashfaq, S. (2022). Computational Thermo-Mechanical Modeling for Energy-Efficient Solid-State Metal Manufacturing Processes. *American Journal of Interdisciplinary Studies*, *3*(04), 579-618. https://doi.org/10.63125/ddg6mg97

[70]. Md, F. (2023). A Review on Understanding Data Governance Failures in Analytics Systems: Insights from Expert Interviews and Root-Cause Thematic Coding. *Journal of Sustainable Development and Policy*, *2*(04), 346-385. https://doi.org/10.63125/rem5kx95

[71]. Md Khaled, H. (2021). An Empirical Study of CRM and Analytics-Based Approaches to Customer Engagement and Sales Performance Evaluation in Enterprise Organizations. *American Journal of Data Science and Analytics*, *2*(12), 76-155. https://doi.org/10.63125/1tt57n77

[72]. Md Khaled, H., & Hisham, M. (2022). Intelligent Decision-Support Systems for Cross-Functional Workflow Optimization in Data-Driven Organizations. *Journal of Sustainable Development and Policy*, *1*(02), 168-207. https://doi.org/10.63125/dsfg3k24

[73]. Md. Ashfaq, S., & Ashraful, I. (2025). Quantitative Analysis of Machine Learning Models For Defect Prediction in Metal Additive Manufacturing. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1810–1847. https://doi.org/10.63125/3fkkwg05

[74]. Md. Ashfaq, S., & Manam, A. (2023). Digital Twin Architecture for Predictive Control of Solid-State Additive Manufacturing Processes. *Review of Applied Science and Technology*, *2*(04), 266–307. https://doi.org/10.63125/tt00s684

[75]. Md. Nazmul, H., & Amena Begum, S. (2022). AI-Based Psychodiagnostics' Models to Support Early Intervention and Reduce Suicide Risk in Adolescents and Youth: Development and Clinical Validation. *American Journal of Data Science and Analytics*, *3*(06), 40-79. https://doi.org/10.63125/vb5f7e98

[76]. Md. Shahinur, I., & Md. Sultan, M. (2022). Digital-Twin-Based Quantitative Frameworks for Modeling, Monitoring, and Optimization of Electrical Power Infrastructure. *American Journal of Interdisciplinary Studies*, *3*(04), 365-393. https://doi.org/10.63125/dvmj1y93

[77]. Md. Towhidul, I., & Uddin, M. D. S. (2024). Simulation-Based Forecasting and Inventory Control Models For Consumer Goods Networks: A Quantitative Study Using Monte Carlo Simulation and Time-Series Methods. *Review of Applied Science and Technology*, *3*(04), 165–197. https://doi.org/10.63125/a3047d06

[78]. Melina, Sukono, Napitupulu, H., & Mohamed, N. (2023). A conceptual model of investment-risk prediction in the stock market using extreme value theory with machine learning: a semisystematic literature review. *Risks*, *11*(3), 60.

[79]. Mienye, I. D., Sun, Y., & Wang, Z. (2020). An improved ensemble learning approach for the prediction of heart disease risk. *Informatics in Medicine Unlocked*, *20*, 100402.

[80]. Morgenstern, J. D., Rosella, L. C., Costa, A. P., de Souza, R. J., & Anderson, L. N. (2021). Perspective: big data and machine learning could help advance nutritional epidemiology. *Advances in Nutrition*, *12*(3), 621-631.

[81]. Muhammad, L., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., & Mohammed, I. A. (2021). Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. *SN computer science*, *2*(1), 11.

[82]. Munblit, D., Nicholson, T. R., Needham, D. M., Seylanova, N., Parr, C., Chen, J., Kokorina, A., Sigfrid, L., Buonsenso, D., & Bhatnagar, S. (2022). Studying the post-COVID-19 condition: research challenges, strategies, and importance of Core Outcome Set development. *BMC medicine*, *20*(1), 50.

[83]. Murad, M. D. H. R. (2025). Machine Learning–Based Consumer Behavior Prediction Models for E-Commerce Platforms: Enhancing Digital Financial Inclusion and Market Accessibility. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 2078–2118. https://doi.org/10.63125/pnz32s94

[84]. Nalbandian, A., Sehgal, K., Gupta, A., Madhavan, M. V., McGroder, C., Stevens, J. S., Cook, J. R., Nordvig, A. S., Shalev, D., & Sehrawat, T. S. (2021). Post-acute COVID-19 syndrome. *Nature medicine*, *27*(4), 601-615.

[85]. Natras, R., Soja, B., & Schmidt, M. (2022). Ensemble machine learning of random forest, AdaBoost and XGBoost for vertical total electron content forecasting. *Remote Sensing*, *14*(15), 3547.
[86]. Navas Thorakkattle, M., Farhin, S., & Khan, A. A. (2022). Forecasting the trends of covid-19 and causal impact of vaccines using bayesian structural time series and arima. *Annals of Data Science*, *9*(5), 1025-1047.
[87]. Nti, I. K., Adekoya, A. F., & Weyori, B. A. (2020). A comprehensive evaluation of ensemble learning for stock-market prediction. *Journal of big Data*, *7*(1), 20.
[88]. O'Neill, B. C., Carter, T. R., Ebi, K., Harrison, P. A., Kemp-Benedict, E., Kok, K., Kriegler, E., Preston, B. L., Riahi, K., & Sillmann, J. (2020). Achievements and needs for the climate change scenario framework. *Nature climate change*, *10*(12), 1074-1084.
[89]. Ong, A. K. S., Chuenyindee, T., Prasetyo, Y. T., Nadlifatin, R., Persada, S. F., Gumasing, M. J. J., German, J. D., Robas, K. P. E., Young, M. N., & Sittiwatethanasiri, T. (2022). Utilization of random forest and deep learning neural network for predicting factors affecting perceived usability of a COVID-19 contact tracing mobile application in Thailand "Thaichana". *International Journal of Environmental Research and Public Health*, *19*(10), 6111.
[90]. Ong, A. K. S., Prasetyo, Y. T., Yuduang, N., Nadlifatin, R., Persada, S. F., Robas, K. P. E., Chuenyindee, T., & Buaphiban, T. (2022). Utilization of random forest classifier and artificial neural network for predicting factors influencing the perceived usability of COVID-19 contact tracing "Morchana" in Thailand. *International Journal of Environmental Research and Public Health*, *19*(13), 7979.
[91]. Painuli, D., Mishra, D., Bhardwaj, S., & Aggarwal, M. (2021). Forecast and prediction of COVID-19 using machine learning. In *Data Science for COVID-19* (pp. 381-397). Elsevier.
[92]. Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. *Materials today: proceedings*, *80*, 3782-3785.
[93]. Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision support systems*, *118*, 33-45.
[94]. Parwez, M. A., Abulaish, M., & Jahiruddin, J. (2020). A social media time-series data analytics approach for digital epidemiology. 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT),
[95]. Post-COVID-19 global health strategies: the need for an interdisciplinary approach. (2020). *Aging clinical and experimental research*, *32*(8), 1613-1620.
[96]. Premraj, L., Kannapadi, N. V., Briggs, J., Seal, S. M., Battaglini, D., Fanning, J., Suen, J., Robba, C., Fraser, J., & Cho, S.-M. (2022). Mid and long-term neurological and neuropsychiatric manifestations of post-COVID-19 syndrome: A meta-analysis. *Journal of the neurological sciences*, *434*, 120162.
[97]. Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., Jiang, B., & Chen, P.-k. (2022). Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*, *12*(1), 453.
[98]. Qiu, W., Chen, H., Dincer, A. B., Lundberg, S., Kaeberlein, M., & Lee, S.-I. (2022). Interpretable machine learning prediction of all-cause mortality. *Communications medicine*, *2*(1), 125.
[99]. Rajib, S. (2024). Quantitative Assessment of Data-Driven Pricing Optimization Strategies for E-Commerce Platforms in Developing Economies. *Review of Applied Science and Technology*, *3*(02), 01–40. https://doi.org/10.63125/g5va6e03
[100]. Rincy, T. N., & Gupta, R. (2020). Ensemble learning techniques and its efficiency in machine learning: A survey. 2nd international conference on data, engineering and applications (IDEA),
[101]. Rukaiya Khatun, M., & Zakia, A. (2023). Quantitative Assessment of Data Privacy and Access Control Effectiveness in SAP/ERP Analytics Systems. *Review of Applied Science and Technology*, *2*(01), 259–300. https://doi.org/10.63125/vb03b363
[102]. Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. *SN Applied Sciences*, *2*(7), 1308.
[103]. Sahin, E. K., Colkesen, I., & Kavzoglu, T. (2020). A comparative assessment of canonical correlation forest, random forest, rotation forest and logistic regression methods for landslide susceptibility mapping. *Geocarto International*, *35*(4), 341-363.
[104]. Salam, R., & Islam, A. R. M. T. (2020). Potential of RT, Bagging and RS ensemble learning algorithms for reference evapotranspiration prediction using climatic data-limited humid region in Bangladesh. *Journal of Hydrology*, *590*, 125241.
[105]. Santangelo, O. E., Gentile, V., Pizzo, S., Giordano, D., & Cedrone, F. (2023). Machine learning and prediction of infectious diseases: a systematic review. *Machine Learning and Knowledge Extraction*, *5*(1), 175-198.
[106]. Scarpino, S. V., & Petri, G. (2019). On the predictability of infectious disease outbreaks. *Nature communications*, *10*(1), 898.
[107]. Scavuzzo, C. M., Scavuzzo, J. M., Campero, M. N., Anegagrie, M., Aramendia, A. A., Benito, A., & Periago, V. (2022). Feature importance: Opening a soil-transmitted helminth machine learning model via SHAP. *Infectious Disease Modelling*, *7*(1), 262-276.
[108]. Seghier, M. E. A. B., Höche, D., & Zheludkevich, M. (2022). Prediction of the internal corrosion rate for oil and gas pipeline: Implementation of ensemble learning techniques. *Journal of Natural Gas Science and Engineering*, *99*, 104425.
[109]. Shamsul, A. (2025). AI-Driven Condition Monitoring and Fault Detection in Electrical Power and Industrial Control Systems. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1778–1809. https://doi.org/10.63125/csjs7238

[110]. Shamsul, A., & Md. Morshedul, I. (2025). The Role of Cloud-Native Infrastructures in Supporting Autonomous and Uncrewed Systems (UXS) in Operations. *Journal of Sustainable Development and Policy*, *4*(03), 82-125. https://doi.org/10.63125/vntbqq40
[111]. Sharaf, M., Hemdan, E. E.-D., El-Sayed, A., & El-Bahnasawy, N. A. (2023). An efficient hybrid stock trend prediction system during COVID-19 pandemic based on stacked-LSTM and news sentiment analysis. *Multimedia tools and applications*, *82*(16), 23945-23977.
[112]. Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Deep Learning applications for COVID-19. *Journal of big Data*, *8*(1), 18.
[113]. Slavich, G. M., Roos, L. G., Mengelkoch, S., Webb, C. A., Shattuck, E. C., Moriarity, D. P., & Alley, J. C. (2023). Social safety theory: Conceptual foundation, underlying mechanisms, and future directions. *Health psychology review*, *17*(1), 5-59.
[114]. Smith, J. D., & Hasan, M. (2020). Quantitative approaches for the evaluation of implementation research studies. *Psychiatry research*, *283*, 112521.
[115]. Sohn, J.-I., Alakshendra, A., Kim, H.-J., Kim, K.-H., & Kim, H.-D. (2021). Understanding the new characteristics and development strategies of coastal tourism for post-COVID-19: A case study in Korea. *Sustainability*, *13*(13), 7408.
[116]. Sun, X., Douiri, A., & Gulliford, M. (2022). Applying machine learning algorithms to electronic health records to predict pneumonia after respiratory tract infection. *Journal of Clinical Epidemiology*, *145*, 154-163.
[117]. Taghizadeh-Hesary, F., Yoshino, N., & Phoumin, H. (2021). Analyzing the characteristics of green bond markets to facilitate green finance in the post-COVID-19 world. *Sustainability*, *13*(10), 5719.
[118]. Tahmina Akter Bhuya, M. (2025). Machine Learning-Driven Credit Risk Modeling: Transforming Loan Default Prediction and Portfolio Management in U.S. Commercial Banking. *American Journal of Data Science and Analytics*, *6*(12), 01-42. https://doi.org/10.63125/0z894070
[119]. Tanjina Binte, S., & Md. Hasan Or, R. (2022). Advanced Computing, IT Strategy, and Network-Optimized Frameworks for Retail Business Intelligence. *American Journal of Interdisciplinary Studies*, *3*(04), 429-463. https://doi.org/10.63125/dgyg3762
[120]. Tanjina Binte, S., & Sazzadul, I. (2022). Advanced Financial Data Analytics for Anomaly Detection and Pattern Discovery in Large-Scale Financial Data Pipelines. *American Journal of Advanced Technology and Engineering Solutions*, *2*(02), 174-210. https://doi.org/10.63125/g1cdm484
[121]. Tran, T. X., Vo, T. T. T., & Ho, C. (2023). From academic resilience to academic burnout among international university students during the post-COVID-19 new normal: An empirical study in Taiwan. *Behavioral Sciences*, *13*(3), 206.
[122]. Tutsoy, O. (2023). Graph theory based large-scale machine learning with multi-dimensional constrained optimization approaches for exact epidemiological modeling of pandemic diseases. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(8), 9836-9845.
[123]. Vaughan, L., Zhang, M., Gu, H., Rose, J. B., Naughton, C. C., Medema, G., Allan, V., Roiko, A., Blackall, L., & Zamyadi, A. (2023). An exploration of challenges associated with machine learning for time series forecasting of COVID-19 community spread using wastewater-based epidemiological data. *Science of The Total Environment*, *858*, 159748.
[124]. Voznica, J., Zhukova, A., Boskova, V., Saulnier, E., Lemoine, F., Moslonka-Lefebvre, M., & Gascuel, O. (2022). Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. *Nature communications*, *13*(1), 3896.
[125]. Wan, Z., Xia, X., Lo, D., & Murphy, G. C. (2019). How does machine learning change software development practices? *IEEE Transactions on Software Engineering*, *47*(9), 1857-1871.
[126]. Wang, H., Tao, G., Ma, J., Jia, S., Chi, L., Yang, H., Zhao, Z., & Tao, J. (2022). Predicting the epidemics trend of COVID-19 using epidemiological-based generative adversarial networks. *IEEE Journal of Selected Topics in Signal Processing*, *16*(2), 276-288.
[127]. Wang, J., Sun, X., Cheng, Q., & Cui, Q. (2021). An innovative random forest-based nonlinear ensemble paradigm of improved feature extraction and deep learning for carbon price forecasting. *Science of The Total Environment*, *762*, 143099.
[128]. Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. *Chaos, Solitons & Fractals*, *139*, 110058.
[129]. Wang, X., Zhai, M., Ren, Z., Ren, H., Li, M., Quan, D., Chen, L., & Qiu, L. (2021). Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. *BMC medical informatics and decision making*, *21*(1), 105.
[130]. Wang, Y., Plataniotis, K. N., Wang, J. Z., Hou, M., Zhou, M., Howard, N., Peng, J., Huang, R., Patel, S., & Zhang, D. (2020). The cognitive and mathematical foundations of analytic epidemiology. 2020 IEEE 19th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC),
[131]. Wen, L., & Hughes, M. (2020). Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques. *Remote Sensing*, *12*(10), 1683.
[132]. Wu, H., & Levinson, D. (2021). The ensemble approach to forecasting: A review and synthesis. *Transportation Research Part C: Emerging Technologies*, *132*, 103357.
[133]. Xie, L. (2022). The analysis and forecasting COVID-19 cases in the United States using Bayesian structural time series models. *Biostatistics & epidemiology*, *6*(1), 1-15.
[134]. Xin, L., Xi, C., Sagir, M., & Wenbo, Z. (2023). How can infectious medical waste be forecasted and transported during the COVID-19 pandemic? A hybrid two-stage method. *Technological Forecasting and Social Change*, *187*, 122188.

[135]. Xu, B., Li, J., & Wang, M. (2020). Epidemiological and time series analysis on the incidence and death of AIDS and HIV in China. *BMC Public Health*, *20*(1), 1906.

[136]. Xu, C., Yu, Y., Chen, Y., & Lu, Z. (2020). Forecast analysis of the epidemics trend of COVID-19 in the USA by a generalized fractional-order SEIR model. *Nonlinear dynamics*, *101*(3), 1621-1634.

[137]. Xu, G., Liu, M., Jiang, Z., Söffker, D., & Shen, W. (2019). Bearing fault diagnosis method based on deep convolutional neural network and random forest ensemble learning. *Sensors*, *19*(5), 1088.

[138]. Xu, Z., Shen, D., Nie, T., & Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data. *Journal of Biomedical Informatics*, *107*, 103465.

[139]. Yang, W., Zhang, J., & Ma, R. (2020). The prediction of infectious diseases: a bibliometric analysis. *International Journal of Environmental Research and Public Health*, *17*(17), 6218.

[140]. Yang, Y., Lv, H., & Chen, N. (2023). A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review*, *56*(6), 5545-5589.

[141]. Yeşilkanat, C. M. (2020). Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm. *Chaos, Solitons & Fractals*, *140*, 110210.

[142]. Yin, X., Liu, Q., Pan, Y., Huang, X., Wu, J., & Wang, X. (2021). Strength of stacking technique of ensemble learning in rockburst prediction with imbalanced data: Comparison of eight single and ensemble models. *Natural Resources Research*, *30*(2), 1795-1815.

[143]. Zaheda, K. (2021). Design and Optimization of Dual-Band Microstrip Patch Antenna For 5g Sub-6GHz and mmWave Applications. *American Journal of Data Science and Analytics*, *2*(12), 41-75. https://doi.org/10.63125/cnze8c43

[144]. Zakia, A., & Rukaiya Khatun, M. (2024). Quantitative Assessment of CRM-Based Business Intelligence on Customer Satisfaction and Retention: Evidence from Multi-Channel Service Operations. *Journal of Sustainable Development and Policy*, *3*(02), 01-42. https://doi.org/10.63125/hjd22x72

[145]. Zamora-Mendoza, B. N., de León-Martínez, L. D., Rodríguez-Aguilar, M., Mizaikoff, B., & Flores-Ramírez, R. (2022). Chemometric analysis of the global pattern of volatile organic compounds in the exhaled breath of patients with COVID-19, post-COVID and healthy subjects. Proof of concept for post-COVID assessment. *Talanta*, *236*, 122832.

[146]. Zhan, C., Zheng, Y., Zhang, H., & Wen, Q. (2021). Random-forest-bagging broad learning system with applications for COVID-19 pandemic. *IEEE Internet of Things Journal*, *8*(21), 15906-15918.

[147]. Zhang, Y., Liu, J., & Shen, W. (2022). A review of ensemble learning algorithms used in remote sensing applications. *Applied Sciences*, *12*(17), 8654.

[148]. Zhao, L., Wu, X., Niu, R., Wang, Y., & Zhang, K. (2020). Using the rotation and random forest models of ensemble learning to predict landslide susceptibility. *Geomatics, Natural Hazards and Risk*, *11*(1), 1542-1564.

[149]. Zubchenko, S., Kril, I., Nadizhko, O., Matsyura, O., & Chopyak, V. (2022). Herpesvirus infections and post-COVID-19 manifestations: a pilot observational study. *Rheumatology International*, *42*(9), 1523-1530.