BLENDED ENSEMBLE MODEL FOR PREDICTION OF INFECTIOUS DISEASES

XU DEREN

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF THESIS

Author's full name	: XU DEREN		
Matric No.	: pcs203038	Academic Session	: 20242025-1
Date of Birth	: 15 JUI 1980	UTM Email	: xuderen@graduate.utm.my
Thesis Title	: BLENDED ENSEMBLE MODEL FOR PREDICTION OF INFECTIOUS DISEASES		

I declare that this thesis is classified as:

☑ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the thesis belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this thesis for academic exchange.

Signature of Student:

Signature : xuderen

Full Name XU DEREN

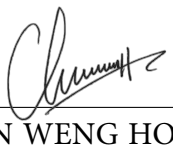
Date : 1 Dec 2024

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I	Full Name of Supervisor II
CHAN WENG HOWE	
Date : 1 Dec 2024	Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

"I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of the degree of Doctor of Philosophy (Computer Science)"

Signature	:	
Name of Supervisor	:	CHAN WENG HOWE
Date	:	1 DECEMBER 2024

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration _____and _____.

Certified by:

Signature :

Name :

Position :

Official Stamp

Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:** **Prof. Madya Dr. Shahreen binti Kasim**
Fakulti Sains Komputer dan Teknologi Maklumat
Universiti Tun Hussein Onn Malaysia
86400 Batu Pahat
Johor

Nama dan Alamat Pemeriksa Dalam **:** **Prof. Dr. Azlan bin Mohd Zain**
Fakulti Komputeran
Universiti Teknologi Malaysia
81310 Johor Bahru
Johor

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

BLENDED ENSEMBLE MODEL FOR PREDICTION OF INFECTIOUS DISEASES

XU DEREN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

Faculty of Computing
Universiti Teknologi Malaysia

DECEMBER 2024

DECLARATION

I declare that this thesis entitled *"BLENDED ENSEMBLE MODEL FOR PREDICTION OF INFECTIOUS DISEASES "* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	xuderen
Name	:	XU DEREN
Date	:	1 DECEMBER 2024

ACKNOWLEDGEMENT

On my journey to study, I have taken many detours and endured numerous challenges before submitting this doctoral thesis to you. Now, having entered an age of confusion, I continue to tirelessly seek knowledge in the temple of learning, where every effort feels like a pursuit of an elusive ideal. In my childhood, I experienced hardship, lacking proper clothing and sufficient food. Yet, there was no time more precious than those nights spent completing homework under the dim glow of a paraffin lamp. While poverty may stifle hope, adversity often serves as the most profound teacher.

I have drifted from place to place, worked as an apprentice, struggled to make ends meet, and once found myself held hostage by fate. Yet, I have never given up. At 36 years old, I still believe in my heart: if there is light within you, a path will emerge beneath your feet, and you will eventually flourish. I often long for the long-lost fragrance of books and ink, which ignited my desire to continue learning. Fortunately, my wife has always supported and encouraged me, allowing me to enroll in adult education. After navigating a few changes and alternating between work and study, I finally immersed myself in the sea of knowledge, with computers becoming my lifelong career partner. At the same time, cervical spondylosis and gastric hemorrhage have also become my unwavering companions.

The future is destined to be an even more complex road ahead of me. The sea of learning is endless, so I must row forward! I possess the courage and patience to face all difficulties and challenges. My aspirations are not ambitious; I simply hope that when I reach the age of 100, I can still return to this world with the spirit of youth and honor the arduous journey of this life. I sincerely thank my tutors and classmates who have supported me along the way. I do not need to ask whether there is a soulmate waiting for me; rather, I wonder who does not know you in this world! If I achieve anything in this life, I will always remember your support and love.

ABSTRACT

Infectious diseases have caused significant disruption to humanity, severely hindering societal progress and global health security. Since the twentieth century, the increasing prevalence of infectious diseases that driven by globalization, urbanization, and environmental changes has highlighted their extensive socio-economic and public health impacts. As a result, managing infectious diseases, particularly through early detection and prediction, has become a fundamental aspect of global public health initiatives. In this modern era, owing to the unprecedented advances in artificial intelligence and the growing importance of interdisciplinary approaches, various models have been developed to analyze and predict trends in infectious diseases. These models utilize machine learning, statistical techniques, and large datasets to produce actionable insights. However, early prediction of emerging infectious diseases presents significant challenges. Key issues include difficulties in selecting appropriate models for novel outbreaks, inherent limitations in prediction accuracy and generalization, and the scarcity of reliable early-stage data, which hinders effective decision-making during critical early phases of disease outbreak. To address these issues, this study introduces a novel multi-model ensemble based on the concept of blending, that integrates ridge regression, decision tree, and XGBoost models. The blended ensemble model is specifically designed and trained for predicting COVID-19 cases with improved accuracy and generalization. The blended ensemble model is further enhanced with transfer learning, incremental learning, and epidemic feature, Rt, which enable better generalization to new emerging infectious disease. The enhanced blended ensemble model is tested on monkeypox case prediction using 14 days of monkeypox cases. The outcome reveals that the enhanced blended model significantly outperforms the pre-optimized ensemble model, achieving root mean square error (RMSE) and mean absolute error (MAE) values of 3.14 and 2.3, respectively, compared to the latter's RMSE of 18.48 and MAE of 16.99. These findings underscore the efficacy of the proposed blended ensemble model in addressing data scarcity and improving predictive accuracy. Academically, this research enhances the theoretical understanding of machine learning applications in infectious disease modeling. Practically, it establishes a foundation for sustainable, high-accuracy predictive systems to help public health authorities mitigate future outbreaks, bridging critical gaps in global health crisis response.

ABSTRAK

Penyakit berjangkit telah menyebabkan gangguan besar kepada umat manusia, menghalang kemajuan masyarakat dan keselamatan kesihatan global. Sejak abad ke-20, peningkatan kelaziman penyakit berjangkit yang didorong oleh globalisasi, urbanisasi dan perubahan persekitaran telah membawa kesan sosio-ekonomi dan kesihatan awam yang meluas. Oleh itu, pengurusan penyakit berjangkit melalui pengesanan awal dan ramalan, telah menjadi satu aspek asas dalam inisiatif kesihatan awam global. Dalam era moden ini, susulan daripada kemajuan luar biasa dalam kecerdasan buatan serta kepentingan pendekatan antara disiplin yang semakin meningkat, pelbagai model telah dibangunkan untuk menganalisis dan meramalkan trend penyakit berjangkit. Model-model ini menggunakan pembelajaran mesin, teknik statistik, dan set data besar untuk menghasilkan cerapan yang boleh diambil tindakan. Walau bagaimanapun, terdapat cabaran yang ketara dalam ramalan awal bagi penyakit berjangkit yang baru muncul. Isu utama termasuk kesukaran dalam memilih model yang sesuai untuk wabak baru, keterbatasan dalam ketepatan ramalan dan generalisasi, serta kekurangan data yang boleh dipercayai pada peringkat awal, telah menghalang pembuatan keputusan yang berkesan semasa fasa awal yang kritikal bagi sesuatu wabak penyakit. Bagi menangani isu-isu ini, kajian ini memperkenalkan satu gabungan baru berdasarkan konsep pengadunan yang menggabungkan pelbagai model iaitu regresi permatang, pepohon keputusan, dan model XGBoost. Model gabungan ini direka khas dan dilatih untuk meramalkan kes COVID-19 dengan ketepatan dan generalisasi yang lebih baik. Model gabungan ini dipertingkatkan lagi dengan pembelajaran pemindahan, pembelajaran tambahan dan ciri epidemik, Rt, yang membolehkan generalisasi yang lebih baik terhadap penyakit berjangkit baru muncul. Model gabungan yang dipertingkatkan telah diuji pada ramalan kes cacar monyet dengan menggunakan data kes cacar monyet selama 14 hari. Hasil ujian menunjukkan bahawa prestasi model gabungan yang dipertingkatkan lebih baik, dengan capaian nilai punca min ralat kuasa dua (RMSE) dan punca ralat mutlak (MAE) sebanyak 3.14 dan 2.3, berbanding dengan model gabungan sebelum peningkatan, dengan capaian RMSE dan MAE masing-masing sebanyak 18.48 dan 16.99. Penemuan ini menekankan keberkesanan model gabungan yang dicadangkan dalam menangani kekurangan data dan meningkatkan ketepatan ramalan. Dari segi akademik, penyelidikan ini meningkatkan pemahaman teori tentang aplikasi pembelajaran mesin dalam pemodelan penyakit berjangkit. Secara praktikal, ia mewujudkan asas untuk sistem ramalan berketepatan tinggi yang mampan bagi membantu pihak berkuasa kesihatan awam mengurangkan wabak masa depan dan merapatkan jurang dalam tindak balas krisis kesihatan global.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACF	–	Auto Correlation Function
ANN	–	Artificial Neural Network
ARIMA	–	Auto Regressive Integrated Moving Average
ARMA	–	Auto Regressive Moving Average
BIC	–	Bayesian Information Criterion
CNN	–	Convolutional Neural Network
COVID-19	–	Coronavirus Disease 2019
DBN	–	Deep Belief Network
DT	–	Decision Tree
FNN	–	Feedforward Neural Network
GA	–	Genetic Algorithm
GBDT	–	Gradient Boosting Decision Tree
GRU	–	Gated Recurrent Unit
H1N1	–	Influenza A virus subtype H1N1
HIV	–	Human Immunodeficiency Virus
KNN	–	K-Nearest Neighbors
LSTM	–	Long Short-Term Memory
MAD	–	Mean Absolute Deviation
MAE	–	Mean Absolute Error
MAPE	–	Mean Absolute Percentage Error
MERS	–	Middle East Respiratory Syndrome
NSGA-II	–	Non-dominated Sorting Genetic Algorithm II
PACF	–	Partial Autocorrelation Function

PCA	–	Principal Component Analysis
RFE	–	Recursive Feature Elimination
RF	–	Random Forest
RNN	–	Recurrent Neural Network
RMSE	–	Root Mean Squared Error
SARS	–	Severe Acute Respiratory Syndrome
SARS-CoV-2	–	Severe Acute Respiratory Syndrome Coronavirus 2
SARIMA	–	Seasonal Autoregressive Integrated Moving Average
SEIR	–	Susceptible-Exposed-Infectious-Recovered
SEIRS	–	Susceptible-Exposed-Infectious-Recovered-Susceptible
SEIS	–	Susceptible-Exposed-Infectious-Susceptible
SIR	–	Susceptible-Infectious-Recovered
SIS	–	Susceptible-Infectious-Susceptible
SVM	–	Support Vector Machine
TCN	–	Temporal Convolutional Network
WHO	–	World Health Organization
XGBoost	–	Extreme Gradient Boosting
	–	

LIST OF SYMBOLS

ϕ_i	–	Autoregressive coefficients
y_t	–	Actual
c_t	–	Current price
ε_t	–	Error Term
x_i	–	Eigenvectors
ϵ	–	Error
g_i	–	Gradient
l	–	Hysteresis operator
$l(\cdot)$	–	Loss Function
θ_i	–	Moving Average Factor
d	–	Number of Differences
h_i	–	Second derivative
x_t	–	Time series data
c_{t-1}	–	Upfront price
r_t	–	Yield
	–	

CHAPTER 1

INTRODUCTION

1.1 Overview

Throughout human history, infectious diseases have been persistent threats due to their ongoing presence across centuries and continents, causing repeated outbreaks and reemergence despite advancements in medicine. They are significant influencing factors because their impact extends beyond public health, influencing social structures, economic stability, and cultural practices. For instance, pandemics like the Black Death and smallpox not only drastically reduced populations but also reshaped labor systems, trade routes, and societal norms. Understanding these persistent and significant influences highlights the importance of early detection and response to infectious diseases, as effective control can mitigate such profound impacts on society (Ravì *et al.*, 2016; Shashvat *et al.*, 2019).

From the 20th century to the present, new infectious diseases have emerged worldwide every year, with some developing into large-scale epidemics such as COVID-19, H1N1, and HIV (Milligan *et al.*, 2021). For example, in 2003, severe acute respiratory syndrome (SARS) and HIV outbreaks posed serious challenges to public health in China. In 2009, the H1N1 influenza pandemic subsequently occurred in the United States, resulting in tens of thousands of deaths. Furthermore, from 2012 -- 2015, the Middle East Respiratory Syndrome (MERS) epidemic led to tens of thousands of infections. More recently, the novel coronavirus (COVID-19) first appeared in Wuhan, China, in 2019 and quickly spread globally.

According to the World Health Organization, over 1,000 infectious diseases have been identified to date. Historically, the number of casualties caused by these diseases has even exceeded that caused by wars (Pradines and Rogier, 2018). Moreover, these diseases can trigger a range of social problems such as social unrest,

economic decline, and famine, all of which severely hinder social development. Therefore, comprehensive research on infectious diseases and the development of effective prevention measures are significant challenges.

The global pandemic of the novel coronavirus (COVID-19) in 2019, as well as outbreaks of monkeypox in multiple regions in 2022, the emergence of poliovirus in New York sewage, and the discovery of the Marburg virus in Ghana, have sounded an alarm worldwide. These events further confirm the notion that the frequency of infectious disease occurrence is accelerating, and that prevention is more crucial than treatment. With the continuous development of artificial intelligence technology and its increasingly widespread application in various disciplines, it has played an increasingly important role in the prevention and control of infectious diseases. An increasing number of models are being used for the analysis and prediction of infectious diseases, providing scientific evidence for disease prevention, control, and treatment. However, the transmission modes of infectious diseases are becoming more diverse, viruses are mutating frequently, and the incubation period of new infectious diseases has a high level of uncertainty. In addition to the rapid spread of these diseases, preventing and controlling infectious diseases is challenging and complex (Zhong *et al.*, 2020). Therefore, the prevention and control of infectious diseases pose a significant challenge to humanity. Although significant efforts have been made in the prevention and control of infectious diseases, there is still a long way to go (Hong *et al.*, 2022).

1.2 Problem Background

In the long history of the struggle between humanity and infectious diseases, mathematicians have been attempting to utilize mathematical models to study the transmission of these diseases (Choisy *et al.*, 2007). As early as 1906, Hamer designed a mathematical model for the spread of measles in a discrete-time system, introducing the concept of the incidence rate, which represents the frequency of new cases occurring in a specific population during a specific time period (Hamer, 1906). In 1911, Ross constructed a differential equation model to explore the relationship

between the transmission of malaria in populations and mosquitoes and discovered that controlling the mosquito population could effectively control the spread of malaria.

In 1927, Dr. McKendrick and biochemist Kermack proposed the classical susceptible-infectious-recovered (SIR) model for the spread of infectious diseases. This model divides the population into three categories: susceptible individuals, infectious individuals, and recovered individuals. In the same year, the study also introduced the susceptible-infectious-susceptible (SIS) model, laying the foundation for the dynamics of infectious diseases (Kermack and McKendrick, 1927).

Since then, infectious disease models based on compartmental models have rapidly developed, giving rise to various variant models such as SIRS, SEIR, SEIRS, and SEIS. These models play a significant role in studying the transmission mechanisms of different infectious diseases, predicting epidemic development, and formulating prevention and control strategies.

Postnikov (2020) utilized the SIR model provided by Kermack and McKendrick to estimate the COVID-19 pandemic and evaluate whether this simplest SIR model could generate accurate parameters and predictions. The research findings demonstrated that the simplest SIR model could accurately describe the outbreak of COVID-19 at the average statistical level nationwide. However, owing to the constant transmission rate in the SIR model, it does not fit well when interventions such as quarantine measures are introduced in cities or communities since these measures alter the transmission rate. To address this issue, Wangping *et al.* (2020) developed an extended SIR model (eSIR) that introduces a transmission modifier, allowing the transmission rate within the population to vary over time. In their study, the eSIR model was employed to assess the impact of interventions on the COVID-19 situation in Italy. However, the study noted that further expansion of the eSIR model is needed to incorporate factors such as the incubation period and other unforeseen elements for more accurate predictions.

Padhi *et al.* (2020) utilized various features and sophisticated computational facilities to accurately model infectious diseases. the study developed a quantum

circuit based on infection and recovery rate parameters and demonstrated the benefits of this quantum computing technique in reducing the space and time complexity of epidemic data processing.

In addition, Nisar *et al.* (2021) demonstrated the dynamics of a fractional-order SIRD (susceptible-infectious-recovered-death) mathematical model in the Caputo sense through numerical simulation methods. Jahanshahi *et al.* (2021) introduced memory effects (both long-term and short-term) in the fractional-order SIRD model to explain the multifractal characteristics of COVID-19 progression. Pacheco and de Lacerda (2021) extended the SIRD model by combining the Levenberg-Marquardt technique with Tikhonov regularization to describe its behavior without the need for a priori specifications of the underlying functions.

Recent research has aimed to improve another fundamental mathematical model used for epidemiological prediction, the susceptible-exposed-infectious-recovered (SEIR) model. To create a dynamic SEIR model, researchers have utilized China's mobile data along with the latest COVID-19 epidemiological data to predict the progression of the epidemic. Additionally, the study incorporated machine learning techniques based on the 2003 SARS coronavirus outbreak data into the model (Yang *et al.*, 2020c; Yarsky, 2021).

Yarsky (2021) achieved model consistency with actual data by simulating adjustments to parameters that depend on specific conditions. In addition to applying or extending the basic SIR or SEIR models, some studies have proposed new specific mathematical models for the transmission of COVID-19. Examples include the susceptible-exposed-infectious-hospitalized-recovered-death (SEHRD) model (Ivorra *et al.*, 2020). the susceptible-exposed-infectious-quarantined-recovered (SEIQR) model (Mandal *et al.*, 2020). and the two-layer SEIR/V-UA model (Zhao *et al.*, 2021). However, given the high uncertainty and complexity of the COVID-19 pandemic, mathematical models used for epidemic trend forecasting still have limitations in capturing and robustly identifying the complex relationships that evolve over time, thus offering more reliable results.

Time series models are commonly used for predicting disease incidence rates, and among them, the autoregressive integrated moving average (ARIMA) model is a classic approach widely employed to forecast the future behavior of stationary time series. The ARIMA model is an extension of the autoregressive moving average (ARMA) model and combines the characteristics of autoregressive (AR) and moving average (MA) models (de Lima *et al.*, 2020).

Owing to its statistical requirements and Box–Jenkins methodology, the ARIMA model has become the preferred choice for many researchers in time series analysis. The ARIMA model evaluates the stability and seasonal trends of a series through time plots, autocorrelation function (ACF) plots, partial autocorrelation function (PACF) plots, and unit root tests. For nonstationary time series, the ARIMA model transforms them into stationary states through differencing operations. The stationary time series is subsequently modelled, and the fit of the model is assessed via the Box-Ljung test. The ARIMA time series method has been used for the stability and growth prediction of COVID-19 (Ghafouri-Fard *et al.*, 2021).

Indeed, a primary limitation of traditional time series models is their assumption that time series data exhibit a linear correlation structure, which may not always hold true for real-world time series data. Therefore, when dealing with nonlinear time series data, the modeling and predictive performance of traditional time series models may be suboptimal.

The emergence of machine learning and deep learning algorithms has propelled further advancements in infectious disease prediction research. In the field of machine learning, artificial neural networks (ANN) are one of the most commonly used classical methods. ANN outperform traditional time series models in predicting time series data because of their strong modeling capabilities and lower data requirements (Tripto *et al.*, 2020). Time series data originating from various domains exhibit nonlinearity and variability, which pose challenges in modeling such data. Deep learning algorithms, on the other hand, have the ability to recognize structures in multidimensional complex data. Therefore, the study enhance the performance of time series prediction models when dealing with irregular and nonstationary time series datasets. Among them, recurrent neural networks (RNN) are a promising class of deep learning models

that have attracted increasing attention from researchers across various fields and are gradually replacing many traditional statistical models (Shin *et al.*, 2021).

One variant of the RNN is the long short-term memory (LSTM) network, which is designed to capture and model long-term dependencies in time series data (Zhang and Fu, 2020). Specifically, in modeling the long-term dependencies in time series data, RNN, particularly LSTM, have shown excellent performance (Wang *et al.*, 2020c). LSTM has special gate structures that enable it to store input information from long sequences and address the issue of vanishing gradients (Li, 2022).

In the field of infectious disease prediction, deep learning techniques hold great promise for time series forecasting. However, the prediction accuracy of individual deep learning models may be limited, even with ideal parameter settings and applications, due to the inherent characteristics of time series data. Any fluctuations in the data can increase the instability of predictions (Mahajan *et al.*, 2022a). Therefore, an increasing number of studies advocate for ensemble prediction models, which outperform individual models in terms of improving prediction accuracy (Hong *et al.*, 2022; Jia *et al.*, 2018).

Ensemble learning combines multiple independent models to achieve better generalization performance. Currently, deep learning architectures have demonstrated superior performance compared with shallow or traditional models. Deep ensemble learning models combine the advantages of deep learning models with those of ensemble learning, resulting in improved generalization performance of the final model (Ganaie and Hu, 2021). Based on the development of ensemble learning, ensemble learning methods are mainly categorized into five basic categories: bagging, boosting, voting stacking and blending. The core idea of the stacking method is to train a series of base models and use another model to train the predictions of the base models, thereby reducing the generalization error and improving the model's robustness. Stacking architecture techniques have been shown to outperform baseline models such as bagging or boosting in terms of predictive performance (Nath and Sahu, 2019). The blending model has been widely used in the financial and medical fields (Xu *et al.*, 2024). However, few blending model have been applied in infectious disease prediction studies.

Gupta *et al.* (2022) showed that several studies in recent years have used stacked ensemble-based deep neural network models to overcome the limitations of COVID-19 complication data and provide a feasible method to accurately predict cardiac complications. Mahajan *et al.* (2022b) further verified that the prediction accuracy of such stacked ensemble models is higher than that of standard gradient boosting models. The stacked ensemble learning model optimized by the genetic algorithm proposed by Ismail *et al.* (2023) performed well in responding to COVID-19 epidemic prediction, with an overall accuracy rate of 99.99% in the three selected countries/regions.

However, most of these studies focus on single-objective optimization, mainly improving prediction accuracy, and do not fully balance the computational efficiency and the adaptability of the model, which may limit the practical application of the model in specific public health scenarios (Ferrández *et al.*, 2023; Mahajan *et al.*, 2022c; Hasan *et al.*, 2024a; Dervishi, 2024a; Shen *et al.*, 2024a). The balance between accuracy, versatility, and computational efficiency is critical in early modeling of infectious diseases, which makes the introduction of multi-objective optimization particularly necessary (Diraco *et al.*, 2023). At present, the application of multi-objective optimization methods in the early prediction of infectious diseases is still limited, especially compared with single-objective optimization, and its potential in solving multiple needs needs to be further explored (Deng *et al.*, 2022).

These studies have several limitations. First, most models have adopted single-objective optimization approaches, which focus primarily on enhancing the prediction accuracy while neglecting other critical factors such as the model's generalizability, computational efficiency, and feasibility in practical applications (Li *et al.*, 2023b; Khan *et al.*, 2023). This pursuit of a single objective may lead to limitations in the application of the model under specific circumstances, as it fails to fully meet the complex demands of the public health sector (Akbulut *et al.*, 2023; Sassano *et al.*, 2022). Second, It is challenging for a single model to accommodate both the non-linear characteristics of the data and the dynamically changing requirements, and there is a lack of widely applicable models. Data scarcity is a prevalent issue for emerging infectious diseases, particularly during the initial stages of an outbreak. These diseases often lack adequate historical data to facilitate effective model training, leading to

unstable predictions due to insufficient information. To address this challenge, transfer learning has been progressively implemented. Transfer learning is a machine learning technique that enables a model trained on one task (the source task) to apply its knowledge to another related task (the target task). This approach is particularly valuable for emerging diseases, as it facilitates the development of models for new diseases using data from established diseases (e.g., SARS or COVID-19) without the necessity of constructing models from the ground up. Additionally, incremental learning presents another viable solution to this issue. Incremental learning allows a model to dynamically update itself as new data becomes available, eliminating the need for complete retraining. This capability is crucial in the context of infectious diseases, which can evolve rapidly. With incremental learning, models can adapt in real time to reflect changes in the spread or nature of an outbreak, thereby improving the accuracy and timeliness of predictions.

The aim of this study is to employ multiobjective optimization techniques to identify models that exhibit superior predictive performance as foundational models. A blending model is subsequently established, that integrates transfer learning and incremental learning methodologies to address the evolving nature of emerging infectious diseases in their early stages. This approach aims to enable rapid adaptation of the model to potential novel infectious diseases in the future.

1.3 Problem Statement

In the early stages of the COVID-19 outbreak, 27 cases of COVID-19 were reported in Wuhan, China, and the spread of this coronavirus in mainland China gained global attention. One month later, the number of cases in China rapidly increased to 11,791, with 7,153 cases in Hubei Province (Al-Qahtani, 2020). If China had had access to mature transfer learning models that leveraged data from past epidemics such as SARS, combined with a stacking framework to predict COVID-19 spread, this could have provided actionable insights for government decision-makers. Such insights would have enabled timely isolation and preventive measures, significantly

slowing the virus's transmission and lessening its societal impact. However, achieving this level of early prediction and control presents distinct challenges:

(a) ***Difficulty in model selection:*** Infectious disease modeling requires selecting the most appropriate model, balancing accuracy, computational efficiency, and interpretability. Simple models like linear regression may fail to capture complex nonlinear disease dynamics, whereas complex deep learning models can be resource-intensive, often requiring significant computational power and time. The wide range of prediction tasks—such as predicting short-term cases versus long-term trends—demands different models, making it difficult to select a model that meets the specific needs of each scenario.

(b) ***Lack of widely applicable models:*** The complex nature of infectious disease transmission, influenced by factors like mutation rates, incubation periods, and varied transmission methods, makes it challenging to find a model that fits multiple diseases. Different diseases may require tailored models to account for unique factors, which limits the generalizability of any single model and requires adaptation for each new outbreak.

(c) ***Time-consuming data collection and model reconstruction for emerging infectious diseases:*** For new infectious diseases, data collection and model training are often lengthy processes. Since real-time responses are crucial in epidemic control, the time required to collect comprehensive data and reconstruct a model can hinder effective intervention. Moreover, gathering sufficient, high-quality data to train reliable models is often a significant barrier, especially in the early phases of an outbreak. This time delay not only hampers immediate response but can also affect the model's predictive performance, as delays lead to outdated information and less accurate predictions.

If there were a method to avoid the need to collect training data and rebuild models, it would be a significant advancement. The concept of transfer learning was introduced precisely for such situations, where researchers began exploring knowledge transfer across different task domains (Zhu *et al.*, 2022). The goal of transfer learning is to enable machines to learn from experience similar to humans, where a model trained on Task A can provide some level of knowledge for Task B, thus avoiding the need

to build a model from scratch. The integration of transfer learning with models that exhibit high predictive accuracy to establish a long-term, effective model for predicting new infectious diseases remains an important and unresolved issue.

This study hypothesizes that predictive performance can be enhanced and emerging infectious diseases can be effectively forecasted through the use of a hybrid model—an approach that combines the strengths of multiple algorithms. Hybridization refers to the integration of various models (e.g., Ridge regression, decision trees, and XGBoost) to create a unified model that leverages the advantages of each model while mitigating their weaknesses. This hybrid approach enables the model to capture a more comprehensive understanding of data patterns, ultimately improving prediction accuracy and generalization. By synthesizing predictions from multiple models and employing transfer learning and incremental learning techniques, this hybrid model can address challenges such as limited data availability and the necessity for frequent model updates, resulting in faster and more accurate predictions of emerging infectious diseases.

1.4 Research Goal

In modern infectious disease prediction, a single model often struggles to manage the complexity of the data. Therefore, multi-model fusion techniques are particularly important for enhancing prediction accuracy. Stacking is one such model fusion method that primarily combines the results of various sub-models through a meta-learning algorithm. By training this meta-learning algorithm, the prediction outcomes of several distinct sub-models are integrated into a single cohesive module, generating the final prediction results and thereby reducing generalization error. Blending is similar to stacking in that blending does not use cross-validation to train the base model, thus simplifying the aggregation process and reducing the risk of overfitting while maintaining the model complexity. The main advantage of blending models is their flexibility and robustness to adapt to datasets with different characteristics, thus better coping with the complex and variable nature of infectious diseases.

This study aims to propose a blended model with ensemble of Ridge Regression ,Decission Tree, and XGBOOST models to predict COVID-19 cases. The proposed blended model is then enhanced with transfer learning and incremental learning to enable effective adaptation to a new target task, which is monkeypox case prediction. The goal is to establish a sustainable blended ensemble model for predicting emerging infectious diseases.

1.4.1 Research Objectives

The objectives of the research are as follows:

i. To identify base models from deep learning and machine learning for prediction of COVID-19 cases using a multi-objective optimization approach.

ii. To propose a blended multi-model ensemble from identified based models for more accurate COVID-19 cases prediction.

iii. To enhance the blended ensemble model with transfer learning and incremental learning for better adaptation to early prediction of emerging infectious diseases.

1.5 Research Scope

The scope of the research described below encompasses the following objectives:

i. New COVID-19 cases data for 2020 -- 2023 and monkeypox outbreak data for the United States from May 2022 -- February 2024 were obtained from the Mexican COVID-19 dataset provided by our World of Data. These datasets were cleaned and preprocessed to ensure data accuracy and completeness.

ii. Prediction of new cases of COVID-19 in Mexico via deep learning models (FNN, TCN, CNN, LSTM) and machine learning models (RF, Ridge, DT, XGBOOST).

iii. Multiobjective optimization model evaluation metrics (accuracy, generalisation, computational efficiency) are used to compare and select models that perform well.

iv. The model with excellent performance are selected by using the multiobjective optimization method is used as the base model to construct a blended ensemble model to predict new Mexican COVID-19 cases ,which further improves the prediction performance.

v. The blending model combined with transfer learning and incremental learning techniques for predicting the number of monkeypox cases in the United States, rapidly adapts to the changing nature of early and continuous updating of emerging infectious diseases, and explores real-time, dynamic predictions.

1.6 Significance of the Study

This study holds several significance:

(a) ***Improving the accuracy of infectious disease prediction:*** By combining the Ridge, DT, and XGBOOST models and utilizing the blending model to construct comprehensive models, the prediction accuracy of infectious diseases such as COVID-19 and monkeypox can be increased. This contributes to providing more accurate prediction results, aiding governments and public health institutions in formulating appropriate prevention and control strategies to address disease outbreaks and transmission effectively.

(b) ***Addressing the limitations of existing methods for predicting infectious diseases:*** Most models adopt a single-objective optimization approach, which focuses mainly on improving the accuracy of prediction while ignoring other key factors, such as the model's ability to generalzse, its computational efficiency and the feasibility of practical applications. This single-objective

pursuit may lead to limited application of the model in specific contexts and the inability to fully meet the complex needs of the public health domain. However, in the context of infectious disease prediction, the introduction of multiobjective optimization allows the model to simultaneously balance prediction accuracy, computational efficiency and the ability to generalize to new data, thus improving the overall performance and usefulness of the model.

(c) ***Exploring the application of transfer learning incremental learning in infectious disease prediction:*** Transfer learning is a technique that applies knowledge learned in one domain to another. Incremental learning allows models to be updated instantly as new data are received, which is critical for systems that require real-time response. This study explores the potential application of transfer learning incremental learning in early infectious disease prediction by applying a well-constructed integrated model from COVID-19 case prediction to monkeypox case prediction via the concept of transfer learning incremental learning. This contributes to model construction for early infectious disease data scarcity and plays an important role in real-time data processing and prediction systems. This ensures fast and accurate decision making in a constantly updating and changing data environment.

(d) ***Providing time-efficient forecasting methods:*** Accurate forecasting of the incidence of infectious diseases is essential for public health management. However, traditional prediction methods often require large amounts of data and complex model building processes. This study provides a time-saving and efficient method for infectious disease prediction via the blending model and the transfer learning incremental learning technique. It can construct reliable prediction models at a faster speed and provide a timely reference for public health decision-making.

In summary, the significance of this study lies in its ability to improve the accuracy of infectious disease prediction, address the limitations of existing prediction methods, explore the application of transfer learning in incremental learning for infectious disease prediction, and provide a time-saving and efficient prediction approach. These contributions will help enhance public health management and response to disease outbreaks, reduce the waste of personnel and resources, and improve the effectiveness

and precision of prevention and control measures. Furthermore, this study can provide new methods and insights for academic research in related fields, driving advancements in the field of infectious disease prediction. Ultimately, the findings of this research have the potential to provide strong support for infectious disease prevention and control efforts in practical applications.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In the previous chapter, an overview of the study was provided. This chapter reviews the literature and state-of-the-art model techniques related to infectious disease prediction. Existing gaps and shortcomings in the literature will be identified, and their strengths and weaknesses will be discussed. The chapter begins by introducing mainstream infectious diseases identified in recent years, along with their prediction techniques and applications. The challenges faced by infectious disease prediction research are then explored, emphasizing the gaps and issues within this field. Various models and techniques for infectious disease prediction, including their advantages and limitations, have been described. Finally, the chapter highlights the limitations of the literature and provides directions and cues for further research.

2.2 Infectious Diseases

The spread of infectious diseases has had a profound impact on human history. From ancient times to the present day, various infectious diseases have constantly threatened human health and survival. Diseases such as the plague, smallpox, cholera, tuberculosis, malaria, HIV/AIDS, Ebola, dengue fever, and SARS have circulated widely, causing significant harm to human society (Ruan *et al.*, 2020).

In recent years, outbreaks of emerging infectious diseases such as COVID-19 and monkeypox have had a significant impacts on people's lives and health. As early as 430 BC, Ancient Greece experienced a devastating plague in Athens, leading to the deaths of nearly half of the city's population and having a profound impact on Athenian civilization. Between 165 and 180 AD, the Antonine Plague, believed to be

smallpox, was recorded as one of the earliest global pandemics, causing significant loss of life and economic damage (McConnell *et al.*, 2018; Babuna *et al.*, 2020). In the 14th century, Europe was hit by the devastating Black Death, resulting in tens of millions of deaths. In the 16th and 17th centuries, a massive smallpox epidemic swept through the Americas, leading to the deaths of 50 million people. Cholera outbreaks occurred multiple times globally in the 19th century, causing countless deaths. During World War I, malaria spread among the armies of countries such as Britain, France, and Germany, resulting in the deaths of a significant number of soldiers (Karema *et al.*, 2020).

These historical events demonstrate the significant and far-reaching impact of infectious diseases on human society. Therefore, predicting and controlling the incidence of infectious diseases is crucial for protecting human health and life. Over the past few decades, global climate change and its impact on temperature-dependent infectious diseases have been significant. Research has shown that climate change can alter the transmission patterns and geographical distributions of infectious diseases, leading to the expansion of certain disease ranges or their emergence in new areas.

In recent decades, humanity has experienced several major outbreaks of infectious diseases. In 2003, China experienced outbreaks of severe acute respiratory syndrome (SARS) and HIV, followed by the H1N1 influenza pandemic in 2009, and middle east respiratory syndrome (MERS) which infected tens of thousands of people from 2012 -- 2015. These outbreaks have posed serious threats and challenges to human society.

Indeed, humans have coexisted with various viruses, bacteria, and parasites throughout history. Infectious diseases are an undeniable part of human societal development. Over the past century, infectious diseases have caused an estimated 1.68 billion deaths, surpassing the death toll of wars by more than tenfold (Lashley and Durham, 2007). Furthermore, infectious diseases have triggered a range of issues such as social unrest, economic regression, and famine, severely impacting the development of human society.

Therefore, predicting and controlling outbreaks of infectious diseases are crucial for maintaining the stability and healthy development of human society. The significance of this study lies in the application of ensemble learning and transfer learning incremental learning techniques, which provide a more accurate and efficient prediction method for different infectious diseases. This approach will help in taking early preventive measures, reducing the harm caused by disease, and safeguarding people's lives and social stability. Meeting the challenges of future infectious disease control is of great theoretical and practical importance.

2.2.1 COVID-19

In December 2019, a respiratory infectious disease caused by a novel coronavirus was first identified in Wuhan, China, and quickly spread within the city. On February 11, 2020, the International Committee on Taxonomy of Viruses (ICTV) officially named the virus "SARS-CoV-2" (Severe Acute Respiratory Syndrome Coronavirus 2), and the disease caused by the virus was named "Coronavirus Disease 2019" (COVID-19) by the World Health Organization (WHO).

The main symptoms of COVID-19 patients include fever, dry cough, and fatigue. Severe cases can rapidly progress to acute respiratory distress syndrome, septic shock, metabolic acidosis, coagulation dysfunction, organ failure, and even death. Some infected individuals may experience mild initial symptoms or even be asymptomatic, with recovery occurring after approximately one week, especially young adults, immunocompromised individuals, and those who mistakenly attribute their symptoms to the common cold.

The incubation period of the virus is typically 1 to 14 days, with most patients experiencing an incubation period of 3 to 7 days, although it can occasionally exceed 20 days. Studies have shown that patients can be infectious during the incubation period. The main routes of COVID-19 transmission include direct transmission, contact transmission, and aerosol transmission. Direct transmission refers to susceptible individuals inhaling respiratory droplets or aerosols directly

emitted by infected individuals, such as through coughing, sneezing, or speaking. Contact transmission occurs when individuals touch surfaces contaminated with droplets and then touch their mouth, nose, eyes, or other mucous membranes, leading to infection. Aerosol transmission refers to the inhalation of infectious aerosols suspended in the air that contain droplet nuclei from infected individuals.

2.2.2 Monkey Pox

Monkey pox is an infectious disease caused by the monkey pox virus, which is similar to the smallpox virus but generally milder (Jarman *et al.*, 2022). Although smallpox has been eradicated globally since 1980, monkey pox infections continue to occur in Central and West African countries. Monkey pox can be transmitted through zoonotic contact and typically occurs in areas near tropical rain forests where animals carrying the virus are present. Various animals such as squirrels, Gambian rats, dormices, and different types of monkeys can potentially carry the monkey pox virus.

The symptoms of monkey pox include fever, widespread characteristic skin rash, and swollen lymph nodes. Differential diagnosis is required to distinguish it from other diseases such as chickenpox, measles, bacterial skin infections, scabies, syphilis, and drug-related allergies. The incubation period of monkey pox is typically 5 -- 21 days. The disease progresses through a febrile stage and a stage characterized by the eruption of skin lesions, which can last for 2 to 4 weeks.

According to data from the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC) in the United States, monkey pox outbreaks have occurred globally. As reported, as of August 2022, more than 70 countries have reported approximately 26,800 cases of monkey pox, with over 7,100 cases reported in the United States.

Research and surveillance on monkey pox are highly important for promptly implementing control measures and preventing the spread of this disease. Understand-

ing the transmission routes, characteristic symptoms, and epidemiological trends of the virus can aid in the development of targeted prevention and control strategies. This includes promoting personal protective measures, enhancing animal surveillance, and implementing appropriate hygiene measures to minimize the risk of monkey pox transmission.

2.3 Feature selection methods

Feature selection is a key step in machine learning and data mining that aims to identify the most important features for model prediction from a large number of features. Its main goal is to improve model performance, reduce overfitting, and decrease computational complexity. Feature selection is usually divided into three main categories(see Fig2.1):

(a) ***Filtering:*** independently of the model, the relationship between the features and the target variable is assessed by statistical tests (e.g., correlation analysis, chi-square test, etc.), and the most relevant features are selected.

(b) ***Wraparound:*** relies on a specific model and uses methods such as recursive feature elimination (RFE) to optimise model performance by training and evaluating the model with iterative feature selection.

(c) ***Embedding method:*** The Embedding method combines feature selection and model training by using the learning process of the model to assess the importance of features. In XGBoost, for example, the importance of features is usually assessed by calculating the frequency and contribution of feature splits in the tree model.

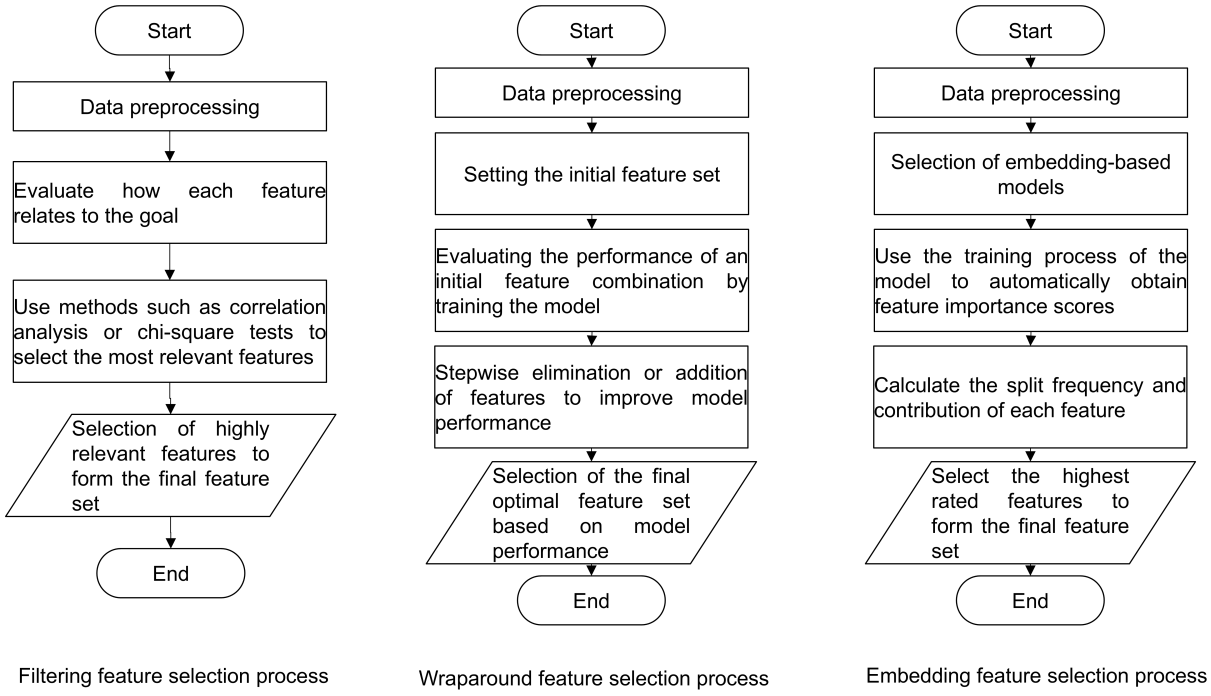


Figure 2.1 Flowchart for feature selection

Feature selection is an integral part of data preprocessing, as it not only improves the efficiency of the model but also helps to understand the important factors behind the data. With effective feature selection, models are better able to generalise to new data.

2.4 Infectious Disease Prediction Methods

COVID-19's global pandemic has had a significant influence on people, and infectious disease prediction has received much attention, Researchers have proposed a variety of methodologies to anticipate infectious illnesses, such as time series models, mathematical models, machine learning models, deep learning models, Hybrid models, The subsections that follow outline these strategies in depth and examine their merits and shortcomings.

2.4.1 Mathematical Models

With the development of human beings, communication and cooperation between different cities is becoming increasingly rapid, and the convenience of technology and transportation has brought people closer to each other, however, this has accelerated the spread of infectious diseases. Researchers have begun to study the patterns and transmission routes of infectious diseases. More than 100 years ago, mathematical modeling became a mainstream technique for addressing the spread of infectious diseases (Guan *et al.*, 2022).

2.4.1.1 SIR

Kermack and McKendrick (1927) first proposed the warehouse model (SIR model). Since the establishment of the model, it has been widely used by later generations and has been continuously enriched and developed. Further understanding of the model has provided a strong theoretical basis. The so-called SIR warehouse

model divides the total population of a certain type of infectious disease area into N, which are divided into three categories according to their conditions. When **S** is susceptible, **I** is infected, and **R** is cleared (Shulgin *et al.*, 1998):

***S**usceptible persons (Susceptibles):* which means the number of people who are not infected at time t during the epidemic period but are likely to be infected, generally denoted as S(t);

***I**nfected persons (Infectives):* the number of people infected at time t during the epidemic period,who have developed the corresponding disease, and have the ability to transmit, generally recorded as I(t);

***R**emoved (Removed):* represents the total number of people who have been separated, cured, or unfortunately died at time t during the epidemic (see Fig 2.2).

Figure 2.2 SIR model diagram

Notably, however, subsequent infections following recovery are not considered here. The SIR is predicated on three assumptions:

(a) There is no change in the total population because natural death and birth are not taken into account.

(b) According to this hypothesis, if a vulnerable person comes into contact with an infected person, the study will almost certainly become infected and gain infectious power. After a certain amount of time, an infected person's infectious potential is proportional to the total number of susceptible persons in the community.

(c) The number of people infected and recovered at any one moment is related to the number of people infected.

The SIR model has an adequate theoretical basis. By solving differential equations and using available data, it can fit the curve well, on the basis of which measures to suppress infectious diseases can be analyzed. However, the SIR model is too simple to allow for a more detailed classification of the population. For example, medical isolation is not considered. In real life, the isolation of patients and suspected patients is a very good prevention and control method to control epidemics. The model is not robust enough. The model is sensitive to the determination of the initial values of each parameter, and the solution of its differential equations is more difficult (Cooper *et al.*, 2022).

2.4.1.2 SEIR

The SIR model cannot account for infectious characteristics if an infectious illness has an incubation period. At this time, the SEIR model evolved, which was built on the SIR model by introducing the incubator, **E**. The entire population had been classified into four groups at this stage. s, e, i, r. SEIR is likewise based on the three assumptions. of the SIR model. On the basis of the first three assumptions, criteria were added: susceptible persons who came into contact with the patient would not become sick, but would become carriers of the disease's pathogen and were categorized into group **E**. The population was classified as susceptible (S), infected (I), exposed (E), and recovered by the SEIR model (R) (see Fig 2.3).

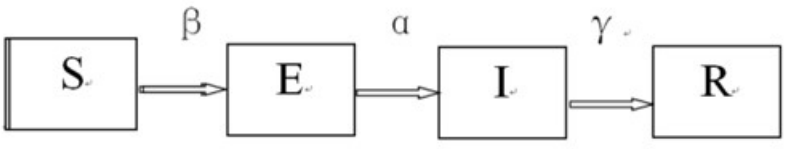


Figure 2.3 SEIR model diagram

Unlike the SIR model, which assumes that contact is infectious, the SEIR model assumes that only a subset of the susceptible population who comes into contact with an infected person is infectious. This property enables a longer disease transmission cycle (Corsi *et al.*, 2021).

Aspects of the epidemic prediction process that are more specific, such as the extent of government control, the rate and extent of population movement, the population's psychological quality, distinct populations' physical traits, and the intensity of patients' symptoms, are not taken into account by the SEIR model. All of these variables influence the disease infection rate, transmission period duration, and number of contacts of affected patients.

SEIR ignores the numerous derived linkages between heterogeneous populations in real-world contexts. For example, whether latent individuals may infect vulnerable populations, whether mortality rates in different groups are comparable, and whether recovered individuals can be infected twice. The model can be improved further to address these flaws (Estiri *et al.*, 2021a).

2.4.1.3 Mathematical Model applications

Kucharski *et al.* (2020) to estimate changes in transmission between January 2020 and February 2020, a stochastic transmission model was integrated with data on Wuhan 2019 coronavirus disease (2019 coronavirus disease) cases and foreign cases originating in Wuhan, and revealed that, in late January 2020, transmission of 2019 coronavirus disease in Wuhan may have declined, which coincides with the introduction of travel control measures.

Atangana and İğret Araz (2020) conducted an exhaustive statistical analysis using data collected from Turkey and South Africa between March 11, 2020 and May 3, 2020 and March 5,2020 and May 3, 2020, respectively, with the aim of understanding why there were fewer deaths and infections in South Africa than in Turkey.

Boulant *et al.* (2020) presented a SEAIR model with an additional input of a customized risk prediction score Each individual is categorized on the basis of his or her current disease condition (moderate or severe symptoms) and the projected level of risk (low or high). This concept results in a fourfold extension of the ODE model in the classic SEAIR.

Liu *et al.* (2020a) used data from China's National Health Council, to examine the evolution of COVID-19 in Wuhan city and assess the efficacy of intervention efforts. A four-stage modified susceptibility-exposure-infection-removal (SEIR) model is proposed. The model takes into account many influencing factors, including the Spring Festival (Spring Festival), city closure, and the construction of Fangcaodi Shelter Hospital. A new approach is also proposed to address the anomalous data from February 12-13 due to the change in diagnostic criteria. The results show that the four-stage model is effective in predicting the peak, magnitude and duration of COVID-19.

Liu *et al.* (2020b) real-time COVID-19 data and population migration figures were obtained from publicly available sources. SEIR (susceptibility, exposure, infection, recovery) synthetic neural network models (NNs) were created to mimic disease patterns in Wuhan, Beijing, Shanghai, and Guangzhou. On the basis of the number of confirmed cases, the Wuhan epidemic is expected to peak in late March and terminate in late April 2020. The epidemics in Beijing, Shanghai, and Guangzhou are projected to be over by the end of March, with medical services returning to normal by mid-April. According to these findings, China's COVID-19 pandemic has been effectively suppressed, and medical services throughout the country are slated to resume normal operations in April.

Su *et al.* (2020) researchers developed an infection model based on reported data from January 24 to February 23, 2020, to forecast the predicted number of infections in four high-risk metropolitan areas and obtain a better understanding of COVID-19 transmission patterns.The R0 values in Beijing were forecasted to be 2.91, 2.78 in Shanghai, 2.02 in Guangzhou, and 1.75 in Shenzhen. These findings suggest that public health actions will reduce the likelihood of COVID-19 transmission, that tighter control and preventive measures are effective in preventing its spread, and that

preventative awareness should be reinforced when business and social activities resume before the outbreak ends.

Deressa and Duressa (2021) constructed a SEAIR epidemic model with Atangana-Baleanu fractional order derivatives. Toufic and Atangana's numerical approach is used to estimate the model's solution. Numerical simulations for various fractional orders reveal that when the fractional order decreases, the spread of the pandemic slows.

Efimov and Ushirobira (2021) analysed the epidemiological history of COVID-19 in eight different nations utilizing publicly accessible data from France, Italy, Spain, Germany, Brazil, Russia, Novel York State (USA), and China via a new version of the well-known epidemiomathematical SEIR model. The model developed was then used to predict the propagation of the SARS-CoV-2 virus under various confinement settings.

Li *et al.* (2021) on the basis of data given by the Wenzhou Health Commission, studied the developmental features of the outbreak and changed the susceptible-exposed-infected-removed (SEIR) model in three ways. The sensitivity of control methods in slowing the spread of the epidemic was determined by simulating the number of infections under various scenarios. Furthermore, the problems associated with epidemic rebound were investigated by simulating a second outbreak, and prevention and control advice was presented. The findings indicate that the improved SEIR model accurately predicts the spread of COVID-19 in Wenzhou.

Liu *et al.* (2021) used neural networks to construct a nonlinear, adaptive dynamic coefficient infectious illness prediction model on the basis of epidemic data. Control measures are assessed, and simulations of the impact of different controls and control measure intensities are run. The adaptive SEIR model was shown to be in good agreement with the actual epidemic curve when the forecast findings for the U.S. COVID-19 epidemic data were compared with those of the conventional SEIR model, SEAIRD model, and adaptive SEIR model. This suggests that the model performs well in terms of prediction.

Lu *et al.* (2022) developed a nonlocal SIHRDP (S-susceptible, I-infected (infected but not hospitalized), H-hospitalized, R-recovered, D-dead, and P-isolated) long-term memory epidemiological model to define the multiwave peak of COVID-19 transmission. The disease-free equilibrium point was found to be globally asymptotically stable by R01, according to the findings of research conducted in Hunan, China. To assess the validity of the proposed model in modeling the multipeak scenarios, parameter identification and short-term prediction were performed on real data from France, India, the United States, and Argentina.

Mahanty *et al.* (2020) on the basis of data received from the Johns Hopkins University repository between January 30, 2020 and June 4, 2020 to explore the global prevalence of coronaviruses, an exponential model and two nonlinear growth models (Gompertz and Verhulst) were used (SIR). COVID-19 patients in India, Pakistan, Italy, Germany, and Brazil are now infected, and those in Myanmar are projected to be (236170, 88998, 234066, 184922, 645057, and 235) and (486357, 218864, 240545 193727, 1211567, and 309), respectively, on the basis of short-term projections for June 5, 2020 and June 30, 2020, The results indicate that the model may be used for long-term prediction.

Sivaraman *et al.* (2022) Exo-SIR, an extension of the popular SIR model and several versions of that model, was tested on real datasets of COVID-19 and Ebola. The model is unique in that it captures both external and endogenous virus transmission. The Exo-SIR model outperformed the SIR model in predicting the peak period. The findings indicate that the Exo-SIR model will assist governments in planning policy measures in the event of a pandemic.

2.4.2 Statistical models

Statistical models are a set of numbers for statistical indicators that are sorted chronologically. Predicting a time series involves analysing the time series and making analogies or extensions on the basis of the methods, directions, and trends reflected in the time series to estimate the levels that may be reached over time or over a number

of years in the future. Time series analysis has the advantage of predicting morbidity (Song *et al.*, 2021). Time series models are advantageous in covering or controlling for a variety of artificial and natural triggers of disease in the time domain (Ma *et al.*, 2020).

2.4.2.1 ARIMA Model

The ARIMA model is built on the autoregressive model (AR) and the sliding average model (MA) with a difference concept, and its model theory interprets the forecast object's data series over time as a random series, and then uses a specific mathematical model to approximate this series (Cao *et al.*, 2020). The differential integrated moving average autoregressive model is the full name for the ARIMA model, which is the integrated moving average autoregressive model. ARIMA is a statistical model that is frequently used for forecasting time series (p, d, q) (Feng *et al.*, 2021). where p represents the autoregressive term, d represents the total number of smoothed data differences, and q represents the number of moving average terms. This is the model:

$$(1-\sum_{i=1}^{p}\phi_i L^i)(1-L)^d X_t = (1+\sum_{i=1}^{q}\theta_i L^i)\epsilon_t \tag{2.1}$$

There are three parameters p, d, and q in the above equation, and their meanings are as follows:

(a) The number of autoregressive terms, denoted by p, represents the number of lags in the time series data utilized in the forecasting model.

(b) The amount of differences necessary for the time series data to be stable is given by d, which indicates the order of the differences.

(c) q is the number of moving average terms in the forecasting model, which is the number of lags of the prediction error.

The previous values of the time series are used to generate the model, and the present values are utilized to forecast the future values, to develop an ARIMA model.

(a) Obtain the observed time series data.

(b) Determine whether the time series data are stationary by plotting them in accordance with the time series data. To convert a nonstationary time series to a stationary time series, the d-order difference operation is needed.

(c) The autocorrelation coefficient ACF and partial autocorrelation coefficient PACF for the aforementioned stationary time series are computed, and the ideal phase p and order q are determined by assessing the autocorrelation and partial autocorrelation plots.

(d) The ARIMA model is built via the previously calculated values of d, q, and p, and it is subsequently evaluated.

Notably, a time series, becomes stable if there is no systematic change in its mean (no trend), no systematic change in variance, and cyclical variation is strictly eliminated.

However, ARIMA as a linear model, generally has disadvantages, such as unsatisfactory long-term forecasting results and an inability to capture the nonlinearity of the data. It is difficult to achieve satisfactory results via linear model forecasting methods. Moreover, the goal of model forecasting is not limited to the next moment, but focuses on longer-term forecasting results. Therefore, linear model-based forecasting methods need further improvement (Zhan *et al.*, 2020; Zhou *et al.*, 2020).

2.4.2.2 Statistical Model applications

Yang *et al.* (2020a) used data from Hubei, China, during a coronavirus blockage. ARIMA time series models were developed for new cases and fatalities. These models were used in Italy, which has the same population as Hubei and is under lockdown, to forecast the outbreak over the following ten days, and the study provides a theoretical basis for the progression of the outbreak in various nations.

Singh *et al.* (2020) the model was trained on observed cases from January 22 -- March 31, 2020, and it was validated on case data from April 1 -- April 17, 2020, retrieved from the official websites of the Malaysian Ministry of Health (MOH) and Johns Hopkins University. The ARIMA model predicted daily COVID-19 instances properly. With a mean absolute percentage error (MAPE) of 16.01 and a Bayesian information criterion (BIC) of 4.170, the ARIMA (0,1,0) model generated the best fit to the observed data. According to these findings, the ARIMA model with optimum covariate selection is the best option for monitoring COVID-19 cases. This study shows how ARIMA models with carefully selected variables may be used to monitor and forecast changes in COVID-19 cases in Malaysia.

Tan *et al.* (2022) COVID-19 A seasonal autoregressive integrated moving average (SARIMA) model using 593 data points and smoothed case and covariate time series data was used to generate a 28-day forecast of the third wave of COVID-19 case trends in Malaysia, with model training and validation performed on the official website. From January 22, 2020 to September 5, 2021, for model training and validation, daily COVID-19 case data were employed. The findings proved that the SARIMA model, which was built with 593 data points, smoothed data, and sensitive factors, was capable of accurately predicting COVID-19 case trends.

Awan and Aslam (2020) auto ARIMA forecasts the number of confirmed cases in four main European nations in the following 10 days via the R package "prediction." The authors discovered that applying Auto ARIMA to the samples correctly predicted verified coronavirus cases over the next ten days.

Aronu *et al.* (2020) between February 28 and March 31, the autoregressive integrated moving average (ARIMA) projection technique was employed to investigate the survival rate of COVID-19 patients in Nigeria, 2020 and June 30, 2020, utilizing secondary data from the Nigerian Center for Disease Control (NCDC) daily publications/reports. The mean daily survival rate of COVID-19 patients was projected to be 27.5%, and the results demonstrated that ARIMA (0, 1, 1) was appropriate for forecasting the survival of Nigerian COVID-19 patients during the observation period.

Diaz Perez *et al.* (2020) researchers have examined two statistical models: the autoregressive integrated moving average (ARIMA) model and the Gompertz functional growth model, with data on verified, nonasymptomatic cases, and confirmed fatalities between February 21 and May 19, 2020. Both models demonstrated promising adjustment errors (R-2> 0.99), with the ARIMA model performing better for infections and the Gompertz model performing better for mortality.

Swaraj *et al.* (2021) described a hybrid model that combines an autoregressive integrated moving average model (ARIMA) and a nonlinear autoregressive neural network (NAR). To extract linear correlations from the data, the ARIMA model was utilized, and the NAR neural network was used to replicate nonlinear ARIMA residuals. On the basis of the performance assessment criteria, the hybrid model exhibited considerably lower RMSE (16.23%), MAE (37.89%), and MAPE (39.53%) values in daily observed cases than did the single ARIMA model. The figures were much lower (39.53%).

Sahai *et al.* (2020) from February 15 to June 30, 2020, researchers collected daily time series data from an online database of total infection cases for the top five countries in the United States, Brazil, India, Russia, and Spain, to compare the first 18 days of July, and then computed ARIMA model parameters via the Hannan and Rissanen technique. Using MAD and MAPE, the predictions were compared with the actual data and determined to be within an acceptable range of prediction accuracy. While Russia and Spain have hit tipping points in the disease's spread, the United States, Brazil, and India continue to see exponential curves.

2.4.3 Machine Learning Model And Deep Learning Models

Machine learning methods include deep learning,and the algorithm techniques, and methodologies for machine learning are very rich and complex. Machine learning algorithms work better than simple statistical models when datasets show complex properties, and machine learning can learn important information from large datasets. Machine learning and deep learning can be categorized by training

methods: supervised learning, unsupervised learning, semisupervised learning, and reinforcement learning (Banyal *et al.*, 2021).

Unsupervised learning is a type of self-organized learning that creates models from unlabeled datasets. Unsupervised learning can be approached in two ways. The first is clustering, where similar data are collected in homogeneous groups. It is achieved by applying one of the many existing clustering algorithms. The second is dimensionality reduction, which involves the reduction of features in high-dimensional data. The aim is to extract new features and find the best linear transformation to represent the maximum number of data points by ensuring the minimum loss of information (Meraihi *et al.*, 2022).

Semisupervised learning is a strategy that combines the ideas of supervised and unsupervised learning. Semisupervised learning approaches make use of both labelled and unlabelled training data, requiring the creative use of supervised and unsupervised learning methodologies to address specific difficulties (Saba Raoof and Durai, 2022).

Reinforcement learning algorithms are slightly different from supervised and unsupervised learning algorithms. This approach allows the algorithm to learn from its own mistakes. To learn how to make the right decision, the AI program is faced directly with a choice. Thus, it learns how to make the right decision. If it is wrong, it is "punished". Good decisions are "rewarded". Artificial intelligence does everything it can to optimize its decisions to obtain increasing rewards (Bai *et al.*, 2023).

2.4.3.1 Supervised Learning

The idea of supervised learning is to guide the algorithm in the form of learning, which is based on prelabelled examples of expected results. Supervised learning relies heavily on manual supervision. Algorithms under the category of supervised learning learn the training data and the corresponding output from a mapping between two variables, and use this mapping in data never seen before. Classification and regression are the two main learning algorithms for supervised learning (Ren, 2021).

Temporal Convolutional Network (TCN) is a deep learning model for sequence data, especially for time series prediction and sequence generation tasks. The core idea of TCN is to use convolutional neural networks (CNNs) to process sequence data, which overcomes the limitations of traditional recurrent neural networks (RNNs) in modelling long sequences. TCN maintains the order of the time series by ensuring that the model can only access past information through causal convolution. Compared to RNNs, TCNs have better parallel computational capabilities and can effectively handle long distance dependency problems. In addition, TCNs use techniques such as residual concatenation and layer normalisation to improve training efficiency and model stability. TCNs have performed well in the fields of speech recognition, natural language processing, and financial forecasting, and are becoming a popular choice for sequence modelling tasks due to their flexibility and efficiency. By combining convolutional operations and time series properties, TCN provides a powerful tool for processing complex sequence data.

Feedforward Neural Network (FNN) is one of the most basic artificial neural network structures, widely used in classification and regression tasks. FNN consists of an input layer, a hidden layer, and an output layer, and information propagates unidirectionally through the network, from the input layer to the output layer, with no feedback connections.The hierarchical structure of FNN makes it easy to understand and implement. Each node (neurone) processes input data through weighting and activation functions.Commonly used activation functions in FNNs include ReLU, Sigmoid, and Tanh. The activation functions introduce nonlinearities that allow the network to capture complex patterns. fNNs can be used in a variety of fields, such as image recognition, natural language processing, and financial forecasting, and are suitable for both linear and nonlinear problems.

Convolutional Neural Network (CNN) is a deep learning model especially suited for image processing and computer vision tasks. CNN is inspired by the biological vision system and is able to extract image features automatically, reducing the need for manual feature engineering.CNN is able to automatically learn important features in an image, such as edges, texture, and shapes. CNN has excellent performances in the domains of image classification, target detection, and semantic

segmentation, and has become one of the standard models in computer vision tasks. CNN has become one of the standard models in computer vision tasks.

Long Short-Term Memory Network (LSTM) is a special type of Recurrent Neural Network (RNN) designed to solve the problem of gradient vanishing and explosion faced by traditional RNNs when dealing with long sequential data.LSTM effectively captures long-distance dependencies through the introduction of memory units and three gating mechanisms.LSTM is capable of effectively learning long-term dependencies in a sequence, which makes it suitable for time series prediction and tasks such as natural language processing and speech recognition.

Random Forest (RF) is an integrated learning algorithm mainly used for classification and regression tasks. It improves the accuracy and robustness of models by constructing multiple decision trees and combining their predictions. Random Forest performs well in many tasks, especially when dealing with large-scale data. Widely used in finance, healthcare, and market analysis, Random Forest is favoured for its simplicity and good performance.

Decision Tree (DT) is a supervised learning model for classification and regression, and its basic structure is a tree diagram for making decisions based on features. Each node represents a test of a feature, branches represent test results, and leaf nodes represent final decision results or predicted values. The decision tree model is intuitive and easy to understand, and the decision-making process can be shown in a tree diagram for easy understanding and interpretation. Decision trees are able to capture complex non-linear relationships between features and target variables.

XGBoost (Extreme Gradient Boosting) is an efficient enhancement learning algorithm mainly used for classification and regression tasks. It is based on the integration of decision trees and improves the overall performance of the model by incrementally building multiple weak classifiers (usually decision trees).XGBoost uses parallel computation, which greatly improves training speed and performs well when dealing with large-scale data.Due to its powerful performance and flexibility, XGBoost has become one of the very popular models in modern machine learning.

Ridge regression is an extension of linear regression that employs L2 regularisation to address multicollinearity. By adding the sum of squares of the weights to the loss function, Ridge regression effectively reduces the complexity of the model and prevents overfitting. With regularisation, Ridge regression reduces the sensitivity of the model to the training data and improves the stability of the prediction.

Linear regression (LR) is a basic statistical and machine learning method for modelling linear relationships between variables. The goal is to predict the target variable by fitting a linear equation. The meaning of the model parameters is intuitive, and it is easy to understand and explain the effect of the characteristics on the target variable. It is computationally fast and suitable for large-scale datasets.

2.4.3.2 Machine Learning And Deep Learning applications

Khan *et al.* (2021) the model was developed utilizing confirmed COVID-19 cases from 146 countries. To predict death in COVID-19 patients, decision trees (DT), logistic regression (LR), random forest (RF), extreme gradient boosting (XGBoost), and K-nearest neighbors (KNN) and DL models (with six ReLU layers and output layers with sigmoid activation) were utilized as ML techniques. The decreased feature sets of the ML and DL models were compared. A comparison analysis was carried out. The suggested DL model generated the best results, with an accuracy of 0.97. compared with previous studies, the experimental findings highlight the importance of the proposed model in the literature of reduced feature sets.

Aldhyani *et al.* (2021) on the basis of real-time World Health Organization (WHO) data. To identify coronavirus, a deep learning system and a Holt trend model are being created. The number of confirmed cases and deaths is forecasted via long short-term memory (LSTM) and Holt trend algorithms. The proposed model was evaluated in three countries and correctly predicted the number of confirmed cases each with 99.94%, 99.94%, and 99.91% accuracy. According to these findings, the LSTM model predicted coronavirus cases more correctly.

Goyal and Singh (2021) using two publicly available chest X-ray image datasets, researchers created a novel model for diagnosing lung illnesses such as pneumonia and COVID-19 from patients' chest X-ray photos. The system includes dataset collection, picture quality enhancement, adaptive and accurate ROI estimation, feature extraction, and illness prediction. Soft computing approaches such as artificial neural networks (ANN), support vector machines (SVM), K-nearest neighbors (KNN), integrated classifiers, and deep learning classifiers are used for classification. Deep learning architectures based on recurrent neural networks (RNN) with long short term memory (LSTM) were designed for correctly identifying lung illnesses, and the results show that the proposed models outperform existing state-of-the-art approaches in terms of resilience and efficiency.

Adnan *et al.* (2022) used a dataset maintained and updated by Johns Hopkins University for the analysis, modeling, and prediction of COVID-19, as well as a deep learning algorithm called the artificial neural network (ANN) and several ML algorithms such as support vector machine (SVM), polynomial regression, and Bayesian ridge regression (BRR) modeling. The results show that the BRR approach delivers more accurate COVID-19 estimates over the following ten days. Infectious illness modeling can assist governments in taking appropriate precautions and making timely choices.

Pishgar *et al.* (2022) developed a deep learning model to predict death in COVID-19 patients, which was updated every 6 hours for the first 72 hours after admission The deep learning/process mining model extracts temporal information associated with variables and integrates demographic and clinical data to anticipate mortality, comparing the model's performance to published and self-developed conventional machine learning models that do not use time as a variable.The results On imbalanced datasets, the proposed process mining/deep learning model outperforms the comparator model for virtually all time periods and has a good AUROC of more than 80%. The results reveal that the suggested process mining/deep learning model outperforms existing machine learning approaches that do not consider temporal information; as a result, temporal information should be added into the model to better anticipate outcomes.

Abdulaal *et al.* (2020) the researchers built and tested an artificial neural network using data from 398 admitted patients to generate patient-specific estimations of mortality risk upon admission to help with early clinical care decisions. ANN examine a set of patient data such as demographics, comorbidities, smoking history, and presenting symptoms to predict current inpatient Patient-specific mortality risk during hospitalization. It predicted patient-specific mortality 86.25% of the time, with a sensitivity of 87.50% (95% CI 61.65%-98.45%) and a specificity of 85.94% (95% CI 74.98%-93.36%). The positive predictive value was 60.87% (95% confidence interval: 45.23%-74.56%), whereas the negative predictive value was 96.49% (95% confidence interval: 88.23%-99.02%). The area under the operating characteristic curve of the subject was 90.12%. The findings illustrate deep learning technology's early usefulness in a rapidly changing epidemic.

Arora *et al.* (2020) To forecast the number of new coronavirus positive cases, deep learning models based on recurrent neural network (RNN) versions of long and short-term memory (LSTM), deep LSTM, convolutional LSTM, and bidirectional LSTM were applied to the Indian dataset (COVID-19). With a daily prediction error of less than 3% and a weekly prediction error of less than 8%, the suggested technique demonstrated exceptional short-term forecast accuracy. This research can be used as a model for other countries to predict COVID-19 incidence at the state or national level.

Kumar *et al.* (2021) researchers have used real-world data to develop two learning algorithms to predict COVID-19: There are two forms of learning: deep learning and reinforcement learning. A model was built utilizing recurrent neural networks (RNN), specifically a modified long short-term memory (MLSTM) model, to estimate the number of newly infected people, losses, and cures in the following days. In terms of the error rate, the suggested method outperforms the long short-term memory (LSTM) model and the machine learning model logistic regression (LR) and is likely to predict the outcome of the present COVID-19 pandemic.

Sayed *et al.* (2021) developed a machine learning model that is based on X-ray pictures to predict the sudden shift risk of COVID-19 patient populations. First, features are extracted via the CheXNet deep which combines principal component analysis (PCA) and sequential backwards elimination (RFE), two different methods

to select the most important features, followed by six machine learning techniques The XGBoost classifier outperforms the other methods by combining (PCA + RFE) features, achieving 97% accuracy, 98% accuracy, 95% recall, 96% f1-score, and 100% ROC-AUC.

Albeshri (2021) proposed a COVID-19 case prediction method based on deep learning and dynamic weighting (DLDW) using various important factors via deep learning to predict new cases, assigning two weights to data instances on the basis of feature importance and time on the basis of dynamic weighting. Data that are older are given less weight, and vice versa. The characteristics that impacted the pace of new case formation over time were determined via feature selection. The prediction accuracy of the DLDW approach was 80.39%, which was 6.54%, 9.15%, and 7.19% greater than those of the other three classifiers, Deep learning (DL), random forest (RF), and gradient boosting machines are examples of machine learning (GBM). This study implies that population lockdown, vaccination, and self-awareness are useful techniques for controlling and managing the COVID-19 pandemic.

Ayoobi *et al.* (2021) six distinct deep learning algorithms were explored via World Health Organization data, To anticipate the number of new cases and deaths in Australia and Iran, unidirectional extensions for each technique were attempted. Three deep learning techniques, LSTM, convolutional LSTM, and GRU, were used to predict new cases of COVID-19 and death time series, as well as their bidirectional extensions, were tested concurrently. The results reveal that the bidirectional model results in fewer errors than the other models do, demonstrating its superiority.

Ardabili *et al.* (2020) studied machine learning and soft computing models to anticipate COVID-19 outbreaks as an alternatives to susceptible infection recovery (SIR) and susceptible exposure infection removal (SEIR) models, and two of the extensively researched machine learning models showed promising results (i.e., multilayer perceptron, MLP; and adaptive Network-based Fuzzy Inference System, ANFIS).

Guadiana-Alvarez *et al.* (2022) constructed a COVID-19 mortality risk calculator in Spain via a deep learning (DL) model and a dataset provided by the HM Hospital in Madrid. A K-nearest neighbor-based synthetic minority method of oversampling (SMOTE) and data interpolation methodology were used. The study included 1503 critically ill COVID-19 patients with a median age of 70 years, 927 (61.7%) men and 576 (38.3%) women, and the results suggest that the proposed approach is the best for forecasting the likelihood of death in COVID-19 patients.

Meem *et al.* (2022) demonstrated a deep learning, machine learning, and convolutional neural network-based approach for identifying COVID-19 positive and normal individuals via chest X-ray images. TensorFlow was used to design and train the neural network, while Scikit-learn was used for end-to-end machine learning, constructing Dense Net, Dropout, and Maxpooling2D models with various deep learning characteristics. After training and testing X-ray images, the proposed approach has a classification accuracy of 96.43% and a validation accuracy of 98.33%.

Chew *et al.* (2021) to evaluate daily COVID-19 time series records as well as a substantial chunk of COVID-19-related Twitter data, researchers have developed ODANN, a hybrid deep learning model utilizing a combination of neural networks (NNs) with an analytical approach and natural language processing. (NLP) feature extraction methods. ODANN outperforms other standard time series models and earlier studies by simulating the growth rate of the global number of confirmed COVID-19 cases using the recommended G-parameter, demonstrating its competitive advantage in predicting and forecasting the global number of COVID-19 cases.

Zreiq *et al.* (2022) long short-term memory, recurrent neural networks (RNN), gated recurrent units (GRU), and deep learning (DL) neural networks were used to mimic the climates, cultures, populations and health systems of four nations (Saudi Arabia, Egypt, Italy, and India): (LSTM). The findings demonstrate that basic structured RNN algorithms outperform deep learning methods in predicting new infections on a daily basis, and that deep learning methods have immense potential for disease modeling and may be employed effectively even with small datasets.

Ramanuja *et al.* (2022) using a publicly accessible dataset from Johns Hopkins University, suggested a supervised LSTM model and its modifications to predict infection cases in India. Experiments with multiple models and window hyperparameters were carried out, and the model was demonstrated to predict infection rates at one week, two weeks, three weeks, and one month.

2.4.4 Hybrid Model

The hybrid model in the field of machine learning and deep learning refers to combining two or more different types of models or algorithms to improve predictive performance, enhance generalization, or solve complex problems that are difficult to handle with a single model. This approach takes advantage of the strengths of each model in the hope of achieving better results than a single model does.

In Tandem hybrid modelling, the outputs of one model are used as inputs to another, forming a continuous processing chain. Each model plays a role in a different stage of data processing. This approach allows data with different characteristics to be processed step-by-step, taking full advantage of the strengths of each model.

In parallel hybrid modeling, different models process the same data in parallel and eventually the outputs of these models are combined to obtain the final prediction. This approach can combine the strong points of multiple models to increase the robustness and accuracy of the prediction.

2.4.4.1 Hybrid Model applications

Zheng *et al.* (2020a) presented a mixed artificial intelligence algorithm for the prediction of new coronaviruses. Because the classic epidemic model assumes that all coronavirus-infected people have the same infection rate, the improved susceptible infectious agent (ISI) model assumes a range of infection rates to investigate transmission patterns and trends. Compared with to the standard epidemic

model, the suggested hybrid AI model, which combines NLP modules and long short-term memory (LSTM) networks with ISI, may dramatically reduce prediction result inaccuracy.

Jin *et al.* (2022) The TCN, GRU, and DBN are used as hybrid model parameters in a TCN-GRU-DBN-Q-SVM model that was suggested for COVID-19 infection prediction and is based on a temporally convolutional network (TCN), rectified linear unit (GRU), deep belief network (DBN), Q-learning, and vector support machine (SVM) models.

Pinter *et al.* (2020) projected COVID-19 using Hungarian data to demonstrate the efficacy of our hybrid machine learning technique. To forecast the time series of infected persons and mortality rates, a hybrid machine learning technique based on an adaptive network-based fuzzy inference system (ANFIS) and a multilayer perceptron - competition algorithm (MLP-ICA) is developed. The model's accuracy has been proven, and the study serves as a first benchmark to showcase machine learning's potential for future research.

Ala'raj *et al.* (2021) used public data to investigate the characteristics of the COVID-19 pandemic to develop a dynamic mixed model for confirmation rates based on SEIRD and automated parameter selection with two components: an ARIMA model in conjunction with a modified SEIRD dynamic model The parameters of the SEIRD model were fitted to historical values of infected, and deceased people divided by the ascertainment rate, and the residuals of the first model for infected, recovered, and deceased people were corrected with an ARIMA model, which analyses input data in real time and provides long- and short-term predictions with confidence intervals. The COVID-19 Tracking Project's US COVID-19 statistical dataset was utilized to test and verify the model, which contains five regularly used metrics to assess the model's predictive power: the MAE, MSE, MLSE, normalized MAE, and normalized MSE. The study showed that the model can generate accurate forecasts for people who have been sick, have recovered, or have died, and that the program's output can be used by the government, business, and legislators.

Saqib (2021) proposed a hybrid machine learning model based on a public dataset provided by Johns Hopkins University until May 11, 2020, which not only has good prediction accuracy but also considers prediction uncertainty. The model uses a mixture of Bayesian ridge regression and n-degree polynomials to estimate the value of the dependent variable via a probability distribution and L-2 (Ridge) regularization to overcome the overfitting problem.

Liao *et al.* (2021) COVID-19 a prediction model based on moment SIRVD was created via deep learning. The model fuses deep learning techniques such as LSTM and other temporal prediction methods with statistical models of infections to predict metrics in mathematical models of infectious diseases, and analyses COVID-19 data for seven countries, including India, Argentina, Brazil, Korea, Russia, the United Kingdom, France, Germany, and Italy, from January 15 to May 27, 2021. The results show that not only does the prediction model outperform pure deep learning techniques in single-day forecasting by 50%, but it also adapts to short- and medium-term forecasting, making the forecast more interpretable and robust.

Zheng *et al.* (2020b) SEIR and RNN on a graph structure were integrated to create a hybrid spatiotemporal model for training and prediction accuracy and efficiency. For the graph structure, two characteristics were introduced: node features (local spatiotemporal infection patterns) and edge features (geographic neighborhood effects). An RNN model was created to capture the neighborhood effect and normalize the landscape of the loss function, ensuring that the local minima are legitimate and resilient for prediction, and this hybrid model (IeRNN) enhanced the prediction accuracy of fresh case data from the U.S. state of COVID-19.

Castillo Ossa *et al.* (2021) propose a hybrid model that combines the population dynamics of a differential equation SIR model with recursive neural network-based implications to produce self-explanatory results for coefficients that fluctuate with restriction measures and can be further refined by expert rules to capture expected changes in these measures.

Zhuang *et al.* (2021) for dynamically displaying and improving opinion event categorization, an LDA-ARMA deep neural network was presented. To extract information about the topic of the comment, latent Dirichlet allocation (LDA) is utilized. Then, an autoregressive moving average model (ARMA) is utilized to conduct multidimensional sentiment analysis and evolution prediction on large-scale text data linked to COVID-19 submitted on Sina Weibo by Wuhan and other countries' Internet users. Wuhan netizens were determined to be more worried about the progress of the COVID-19 problem. Netizens in other regions of the nation were more concerned about overall COVID-19 prevention and control, were more enthusiastic and hopeful about government and non-governmental organization support, and were more emotionally disturbed. This study demonstrates how this approach may be used to investigate large-scale online public opinion change.

Zandavi *et al.* (2021) is a unique hybrid model that combines an artificial recurrent neural network with long short-term memory (LSTM) with a dynamic behavioral model. This model considers a variety of parameters to improve the accuracy of case and death estimates for the top ten most afflicted nations at the time of the study, and the findings show that when data restrictions are included, the hybrid model outperforms the LSTM model.

Sah *et al.* (2022) to develop a COVID-19 predictive analytics method, Prophet, ARIMA, and stacked LSTM-GRU models were used to predict the number of confirmed and active cases in the Indian dataset, and the predicted results were contrasted via recurrent neural networks (RNN), gated recurrent units (GRU), long short-term memory (long short term), linear regression, polynomial regression, autoregressive integrated moving average (ARIMA), and Prophet. The predictions of the overlay LSTM-GRU model were found to be more reliable and had better prediction outcomes than those of the present models.

Muñoz *et al.* (2022) the proposed system employs a SIR model and an artificial recurrent neural network with long short-term memory (LSTM). It can forecast pandemics 4-8 months in advance, allowing the Panamanian government to manage them and slow their spread. Owing to the incorporation of an expert system, it is possible to introduce new variables into the model as soon as the study are known.

The system has the ability not only to provide a clear picture of the current status of the pandemic, but also to predict its evolution.

Zivkovic *et al.* (2021) used X-ray imaging data from COVID-19 patients with different disease severities. A predictive hybrid model was developed to predict COVID-19 disease by applying a hybrid machine learning approach, where features were extracted with the help of the GLCM algorithm and then, with the help of the SVM and CNN algorithms to predict COVID-19 disease. The prediction of COVID-19 disease was performed by combining the SVM and CNN algorithms.

2.5 The Challenge Of Infectious Disease Prediction

Infectious diseases, especially emerging infectious diseases, pose a serious threat to human life. Predicting infectious diseases is very important for the management of infectious diseases and prevention of pandemics. The arrival of new coronavirus epidemics has brought widespread attention to infectious disease prediction. Many techniques have been proven to be accurate in predicting infectious diseases. However, the transmission behavior of COVID-19 has a high degree of uncertainty and complexity. Although many prediction models have been proposed by researchers, each model has its own set of constraints and advantages for specific settings. These models are either limited by their generalizability and scalability or lack predictive data (Wang *et al.*, 2022a). No model to date has been able to predict it with complete accuracy (Santra and Dutta, 2022).

2.5.1 Difficulty In Modelling

One of the difficulties in building accurate models for COVID-19 prediction is the parameter values required for the models to ensure unbiasedness and reliability. More complex models can be used with more complex biological and epidemiological information data, but these complex models involve the estimation of more parameters. This estimation process can be time consuming and lead to a great deal of uncertainty

in model predictions (Roda *et al.*, 2020; Ferrández *et al.*, 2019). Constructing predictive models that consider all the characteristics of infectious diseases is a challenging problem (Kim and Ahn, 2021).

2.5.2 Data Source Difficulties

Accurate data are not available to build models. Most data are inaccurate for political and other reasons, and models trained on such faulty data will not provide accurate predictions. People take precautions and prevention activities actually reduce transmission rates, which must be accounted for in modelling. Capturing the impact of such preventive measures is challenging. This is because many nondeterministic factors are involved. These factors include cultural changes between regions. Attitude change. Social alienation and segregation are widely used preventive measures. However, the effectiveness of these measures is difficult to quantify because of the difficulty of capturing the flow of individuals. With the potential for unidentified infected individuals to transmit to others, capturing this information in models is a daunting task.

2.6 Limitations

In previous studies, researchers have proposed various models to analyse and predict infectious diseases, and with the development of machine learning, an increasing number of studies have proven its validity, however, these models have limitations (see Table 2.1).

Table 2.1 Limitations of various models

Factors	Objective	Limitation
Mathematical models	Capturing the spread of disease with a simple differential equation	The model is based purely on old, existing statistics and does not capture sudden changes in propagation and cannot handle more dynamic situations.
Statistical models	Mathematical derivation of the closed nature of disease transmission	The assumed probability distribution may not be valid for all cases.
Machine learning and deep learning models	Provide generic models that can cover a wide range of scenarios	The models require a large amount of training data and consume a lot of training time.
Hybrid models	Take advantage of hybrid models to deliver better performance	It is very complex as the model consists of a mixture of many models.

This study employs an ensemble learning blending architecture by combining Ridge, DT, and XGBOOST models to predict the number of COVID-19 cases. Subsequently, transfer learning and incremental learning techniques are utilized to apply the ensemble model, which performs well in the source task (COVID-19 case prediction), to a new target task (monkeypox case prediction).

2.7 Opportunities in Advanced Learning Techniques for Infectious Disease Prediction

The prediction performance of a single model is limited, and it is difficult to meet complex and changing infectious disease prediction requirements. The ensemble learning technique has been widely used in infectious disease prediction. These methods can often improve the accuracy and robustness of prediction by integrating

multiple models.Transfer Learning has been increasingly used in infectious disease prediction, especially in the case of data scarcity, and can effectively address the lack of data and improve model performance by transferring the knowledge of models from other diseases or regions to the new task.Incremental Learning techniques can update models as new data are added without the need to train them from scratch, which is particularly suitable for the dynamically changing data environment in infectious disease prediction. These techniques provide new perspectives for the early prediction of emerging infectious diseases. This section reveals research opportunities in these areas by analysing existing research in Ensemble Learning, Transfer Learning and Incremental Learning for infectious disease prediction.

2.7.1 Ensemble Learning

Ensemble learning is a type of machine learning (Tao *et al.*, 2021). It is a very strong machine learning approach whose core idea is to produce base learners on specific criteria and then use an integration strategy to integrate the prediction outcomes of these base learners to form the final result. Integration learning, as opposed to a single machine learning algorithm, integrates numerous machine learning algorithms into a single powerful learner to enhance prediction outcomes. Integration learning has evolved over time to allow for more model generalization. Integration learning may be divided into the following general categories:

2.7.1.1 Bagging

The basic idea behind bagging is to provide a group of independent observations with the same size and distribution as the original data, as well as to provide pooled forecasters that outperform a single predictor on the basis of the original data. Bagging adds two phases to the original model. The first method involves creating bagging samples and sending each sample to the basic model; the second method involves combining multiple predictor predictions. Bagging samples can be created with or without substitution. As most majority voting is used for

classification issues, combining the basic predictors' outputs may make a difference, whereas averaging approaches are used for regression issues with pooled outputs (Ganaie and Hu, 2021).

Ensemble models based on tagging outperform single multilayer predictors (Ha *et al.*, 2005). However, because deep learning models have a long training period, training on various datasets to optimize numerous deep models is not a practical solution (see Fig 2.4).

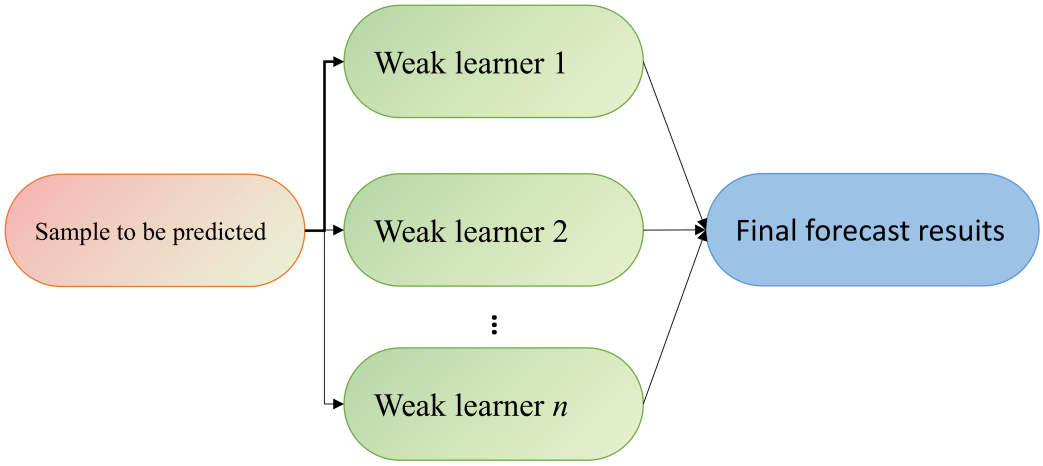


Figure 2.4 Schematic diagram of the bagging technique

2.7.1.2 Boosting

Boosting algorithms are used in ensemble models to transform a weak learning model into a more wider learning model. Using majority voting methods in classification tasks and linear combinations of weak learners in regression problems resulted in better predictions than a single weak learner. Boosting, also known as forward-stage additive models, was initially created to improve the performance of classification trees. Recently, it has been utilized to improve the performance of deep learning models (see Fig 2.5).

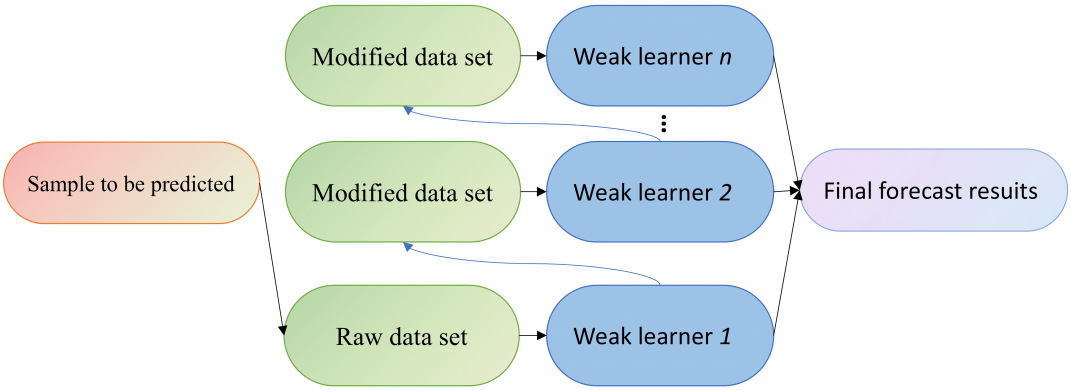


Figure 2.5 Diagram of the boosting technique

2.7.1.3 Stacking

Stacking is a bias reduction approach and one of the model integration strategies. Stacking, unlike bagging and boosting, is primarily concerned with combining the outcomes of all submodels via meta-learner algorithms. Unlike bagging and boosting, Stacking may integrate multiple types of submodels. Stacking can reduce generalization errors by training a meta-learning algorithm to combine the predictions of numerous distinct submodels into an integration module, which generates the final predictions (see Fig 2.6).

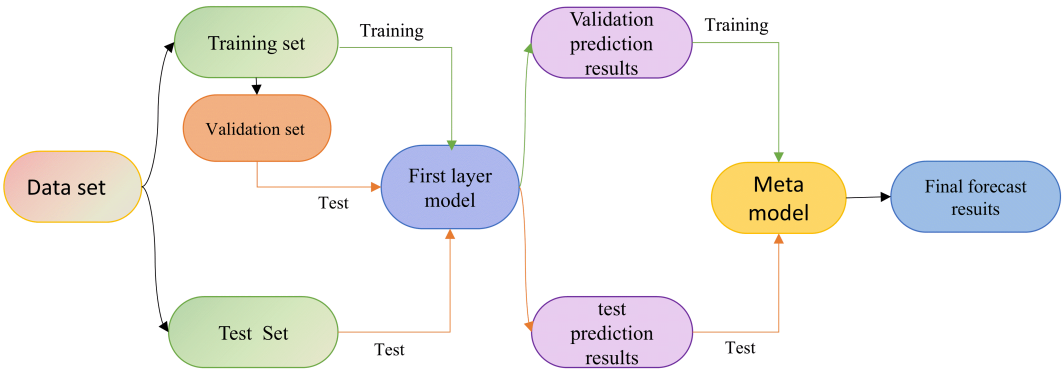


Figure 2.6 Schematic diagram of stacking technology

2.7.1.4 Voting

Voting is an ensemble technique that typically involves combining predictions from multiple models to make a final decision, and is often used in classification tasks. In voting, different models, each potentially using a different algorithm, make predictions independently. These predictions are then combined via a rule such as majority voting (hard voting) or averaging of probabilistic predictions (soft voting). Hard voting counts the votes of each classifier in the ensemble and selects the class with the majority of votes, whereas soft voting computes the average probability assigned to each class and picks the class with the highest average probability. This method can be particularly effective when the individual models are substantially diverse, thereby reducing the likelihood of correlated errors and increasing the robustness of the final prediction (see Fig 2.7).

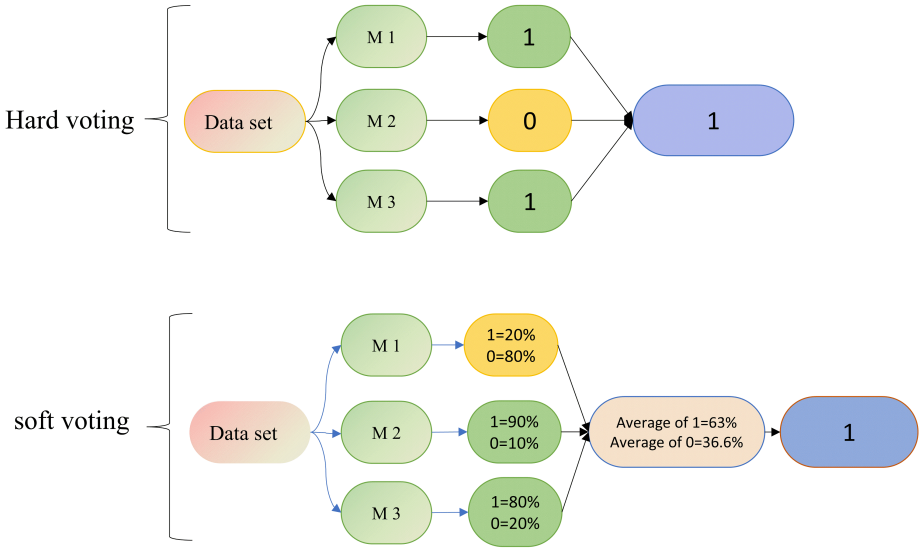


Figure 2.7 Schematic diagram of the voting technique

2.7.1.5 Blending

Blending, similar to stacking, is a model ensemble technique where multiple predictions are combined to produce a final prediction, but the method differs in how the models are integrated. In blending, instead of using a meta-learner as in stacking, a holdout set is used to train a second-level model. The first-level models are trained on a portion of the dataset, and their predictions for the holdout set are used as input features for the second-level model. This final model then learns how to best combine the predictions from the first-level models to improve accuracy. Blending avoids the risk of overfitting on the validation set by keeping the holdout set strictly separate from the training process for the first-level models, potentially providing more generalizable results than stacking does (see Fig 2.8).

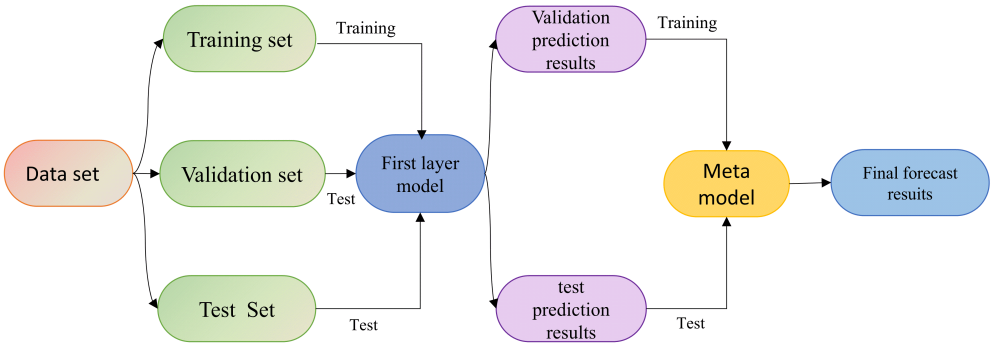


Figure 2.8 Schematic diagram of the blending technique

Blending simplifies the ensemble process by not using cross-validation to train the base models, which reduces the risk of overfitting while maintaining model complexity. This is particularly important in small datasets where cross-validation can further split the training set into smaller segments, potentially leading to underfitting due to insufficient training. This is often the case with datasets for infectious diseases, especially in the early stages of an outbreak (Dissanayake *et al.*, 2023). The simplified ensemble process of blending is crucial for rapidly adapting to changing epidemic data.On the other hand, the cross-validation model training method used in Stacking generally requires larger and more complex datasets to fit the model. This can potentially result in overfitting and poor generalization performance. Additionally, it increases the training time of the model, which could be a limiting factor in public health emergency management where rapid response is necessary.

This study discovered that the Blending method has demonstrated its superiority in other areas of time-series data research, such as risk assessment in finance and consumer behavior prediction in e-commerce (van de Pas *et al.*, 2023). Comparing these studies to our current work on blending, while these studies typically focus only on the model's accuracy or generalizability, our research evaluates the blending model from multiple dimensions—including accuracy, generalizability, and computational efficiency (Zeng *et al.*, 2021; Li *et al.*, 2023c; Hasan *et al.*, 2024b). This multidimensional assessment allows the model to excel in accuracy, generalizability, and computational efficiency, enabling it to quickly adapt to new data and emergent

situations. This capability is crucial for real-time epidemic monitoring and prediction, where rapid data adaptation and response are essential (Minor *et al.*, 2023) (see Table 2.2).

Table 2.2 Comparison of ensemble learning methods

Categories	Objectives	Data	Classifier	Integration methods	Application
Bagging	Variance	Random-based resampling	Homogeneous uncorrelated sub-models	Voting method for classification problems, mean value method for regression problems	Model instability and high variance
Boosting	Deviation	De-correlation based on stepwise misclassification	Homogeneous weak sub-model	Weighted majority voting method	Enhanced Accuracy Captures Complex Patterns
Voting	Reduction of misclassification risk	Diverse data representations	Homogeneous or heterogeneous models	Majority voting for classification (hard voting), average probability voting for regression (soft voting)	Improve robustness and accuracy
Stacking	Deviations and room differences	A wide range of	Heterogeneous hadron models	Meta-Learner	For diverse models and complex datasets
Blending	Reduction of overfitting and error variance	Split data into a training set and a validation set (holdout set)	Heterogeneous models	Use outputs of first-level models as inputs to a second-level model to make final predictions	Suitable for diverse models and time-critical scenarios

The simplest integration module outputs results by voting on the predictions of multiple submodels, which usually yields better forecast accuracy than a single model does. Stacking,blending on the other hand, relies on further learning of the differences

between the different sub-models and then better integrating the different submodels through a second stage algorithm to obtain the best possible prediction results.

2.7.2 Transfer Learning

Transfer learning is a machine learning approach that allows models to transfer what the study have learned on one task to another related task (Caballero *et al.*, 2023; Ma *et al.*, 2024). This approach is particularly useful when labelled data are scarce, or when the dataset for a task is insufficient to train a robust model on its own (Ogunpola *et al.*, 2024). With transfer learning, learning efficiency and prediction performance can be significantly improved (Qin *et al.*, 2022) (see Fig 2.9).

(a) ***Parameter migration:*** Directly reuses some or all of the model parameters in the source task (Li *et al.*, 2024c; Cai *et al.*, 2022b). The most common method is to reuse pretrained network layers in deep learning (Gul *et al.*, 2024).

(b) ***Feature Transfer:*** This uses the feature representation learned from the source task for the target task (Hebbar *et al.*, 2021). This approach typically involves extracting the feature extraction layer of the source model and then using it as input to the target task.

(c) ***Fine-tuning:*** The model is continuously trained on the basis of the source task to adapt to the target task. This typically includes retraining or tuning the top layer of the pretrained model (Peng *et al.*, 2023).

(d) ***Zero-shot learning:*** Transfer learning is used to perform tasks in the absence of a target task sample. This requires the model to be able to capture broad enough knowledge to reason within a new domain (Liu and Ozay, 2023).

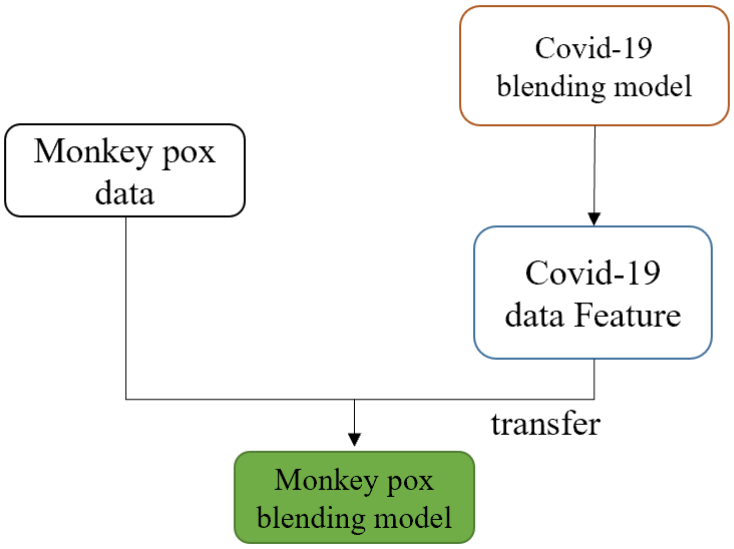


Figure 2.9 Schematic diagram of transfer learning techniques

2.7.3 Incremental Learning

Incremental learning, also known as online learning, is a data stream-oriented approach to machine learning that aims to continuously update a model to accommodate new data without the need to retrain the model from scratch. This learning strategy is ideal for scenarios that need real-time data streams or growing data sets, such as cybersecurity, recommender systems, and real-time transaction monitoring (Keya *et al.*, 2023).

(a) ***Online Incremental Learning:*** Step-by-step updates are a mechanism for a model to be updated as soon as it receives each new data point. This method is particularly suitable for data streaming applications, where data arrives gradually or continuously, and where real-time performance is needed. The main advantage of a single-step update is that it reacts quickly to changes in new data, keeping the model always up-to-date (Yu *et al.*, 2024).

(b) ***Batch Incremental Learning:*** Batch update refers to the accumulation of a batch of data and then the update of the model. This approach allows the model

to learn from more data and may more steadily capture the overall trend of the data, rather than being influenced by individual data points or noise. Batch updates are especially useful when large amounts of data need to be processed or when data arrives quickly, and can conserve computational resources by reducing the frequency of updates (Nguyen *et al.*, 2023) (see Fig 2.10).

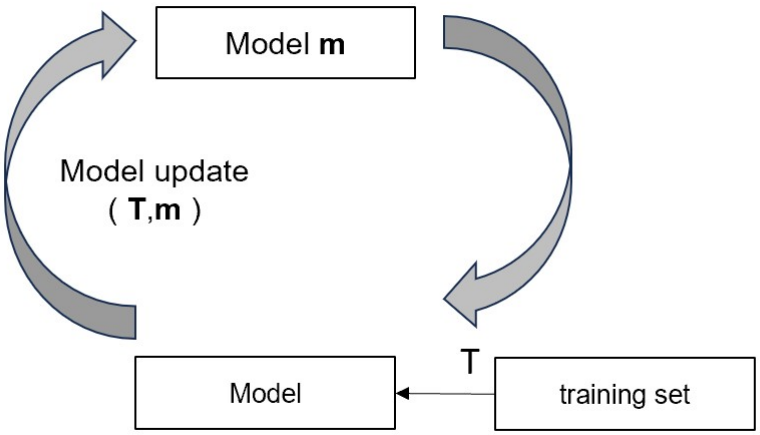


Figure 2.10 Schematic diagram of incremental learning techniques

This study employ blending techniques by combining Ridge, DT, and XGBOOST models to predict the number of COVID-19 cases. This study subsequently apply transfer learning and incremental learning techniques to adapt the well-performing blending model from the source task (COVID-19 case prediction) to a new target task (monkey pox case prediction). The goal is to establish a sustainable model for predicting emerging infectious diseases.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This section discusses the research methodology, which is divided into three main phases, each of which contains research activities undertaken to achieve the research objectives. The input dataset for the model predictions is then discussed. The chapter also discusses the model evaluation metrics. The chapter concludes with a summary of the chapter.

3.2 Data Collection And Processing

An important initial task of this study is data collection, which involves datasets on COVID-19 and monkey pox infections. Mexico is one of the countries with the highest number of COVID-19 infections, and the United States is one of the countries with the highest number of monkey pox infections. Therefore, examining the spread of COVID-19 and monkey pox in Mexico and the United States is highly important. This study uses COVID-19 and monkey pox infection data provided by our World in Data as datasets (see Fig3.1,Fig3.2). The datasets include fundamental information such as country, date, new cases, 7-day moving average, and historical cases (see Fig3.3). The COVID-19 dataset covers the period from January 2020 to May 2023, whereas the monkey pox dataset covers the period from May 2022 to May 2024 (see table 3.1).

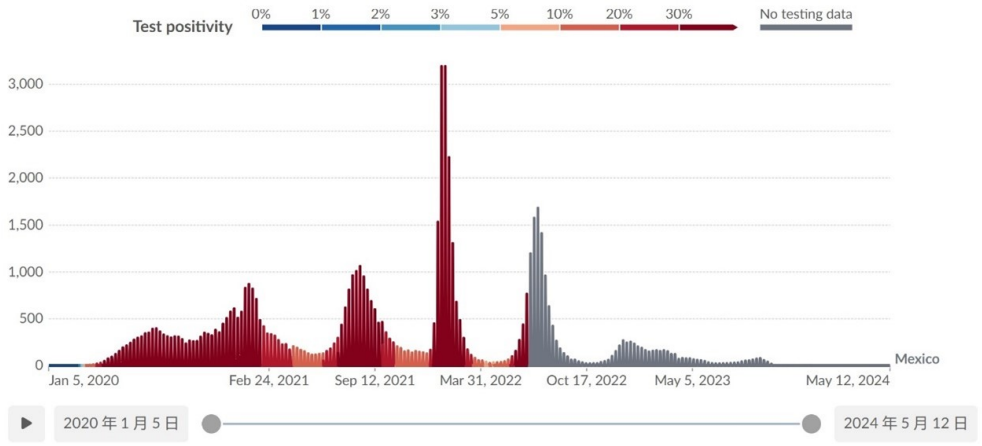


Figure 3.1 Daily trends in the number of COVID-19 cases reported by Mexico to Our World of Data.

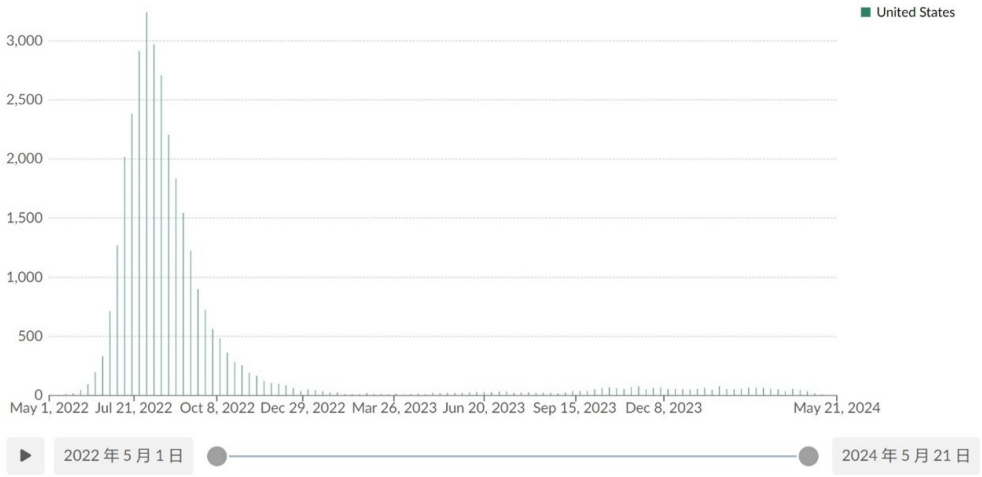


Figure 3.2 Data reporting daily trends in monkey pox cases in the United States

Table 3.1 Relevant data for the study

Data	Country	Source	time scale
COVID-19	Mexican	Our World in Data	2020/4/1 - 2023/3/30
COVID-19	Iran	Our World in Data	2020/4/1 - 2022/12/31
COVID-19	Indonesia	Our World in Data	2020/4/1 - 2023/1/2
COVID-19	Chile	Our World in Data	2020/3/23 - 2023/3/31
Monkey pox	US	Our World in Data	2022/5/29 - 2024/2/27

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	date	new_cases	total_case	new_cases	total_deat	new_deatl	new_deatl	total_case	new_cases	new_cases	total_deat	new_deatl	new_deatl
2	2020/4/1	345	2819	227.571	66	14	8.429	22.109	2.706	1.785	0.518	0.11	0.066
3	2020/4/2	311	3130	240.571	92	26	11.714	24.548	2.439	1.887	0.722	0.204	0.092
4	2020/4/3	382	3512	268.714	114	22	14.571	27.544	2.996	2.107	0.894	0.173	0.114
5	2020/4/4	347	3859	286.286	140	26	17.143	30.266	2.721	2.245	1.098	0.204	0.134
6	2020/4/5	392	4251	304.857	174	34	20.857	33.34	3.074	2.391	1.365	0.267	0.164
7	2020/4/6	250	4501	313.286	214	40	24.429	35.301	1.961	2.457	1.678	0.314	0.192
8	2020/4/7	259	4760	326.571	247	33	27.857	37.332	2.031	2.561	1.937	0.259	0.218
9	2020/4/8	631	5391	367.429	305	58	34.143	42.281	4.949	2.882	2.392	0.455	0.268
10	2020/4/9	515	5906	396.571	367	62	39.286	46.32	4.039	3.11	2.878	0.486	0.308
11	2020/4/10	581	6487	425	416	49	43.143	50.877	4.557	3.333	3.263	0.384	0.338
12	2020/4/11	483	6970	444.429	478	62	48.286	54.665	3.788	3.486	3.749	0.486	0.379
13	2020/4/12	572	7542	470.143	556	78	54.571	59.151	4.486	3.687	4.361	0.612	0.428
14	2020/4/13	464	8006	500.714	637	81	60.429	62.79	3.639	3.927	4.996	0.635	0.474
15	2020/4/14	479	8485	532.143	725	88	68.286	66.547	3.757	4.174	5.686	0.69	0.536
16	2020/4/15	953	9438	578.143	827	102	74.571	74.021	7.474	4.534	6.486	0.8	0.585
17	2020/4/16	936	10374	638.286	941	114	82	81.362	7.341	5.006	7.38	0.894	0.643
18	2020/4/17	985	11359	696	1056	115	91.429	89.087	7.725	5.459	8.282	0.902	0.717
19	2020/4/18	996	12355	769.286	1163	107	97.857	96.899	7.812	6.033	9.121	0.839	0.767
20	2020/4/19	1145	13500	851.143	1305	142	107	105.879	8.98	6.675	10.235	1.114	0.839
21	2020/4/20	815	14315	901.286	1442	137	115	112.271	6.392	7.069	11.309	1.074	0.902
22	2020/4/21	794	15109	946.286	1623	181	128.286	118.498	6.227	7.422	12.729	1.42	1.006
23	2020/4/22	1521	16630	1027.43	1834	211	143.857	130.427	11.929	8.058	14.384	1.655	1.128
24	2020/4/23	1471	18101	1103.86	2049	215	158.286	141.964	11.537	8.657	16.07	1.686	1.241
25	2020/4/24	1423	19524	1166.43	2265	216	172.714	153.124	11.16	9.148	17.764	1.694	1.355
26	2020/4/25	1506	21030	1239.29	2495	230	190.286	164.936	11.811	9.72	19.568	1.804	1.492
27	2020/4/26	1734	22764	1323.43	2758	263	207.571	178.535	13.6	10.379	21.631	2.063	1.628
28	2020/4/27	1164	23928	1373.29	3048	290	229.429	187.665	9.129	10.771	23.905	2.274	1.799

Figure 3.3 A portion of the data properties

3.3 Research Methodology

The study was designed and implemented in three main phases, as shown in Figure 3.4.

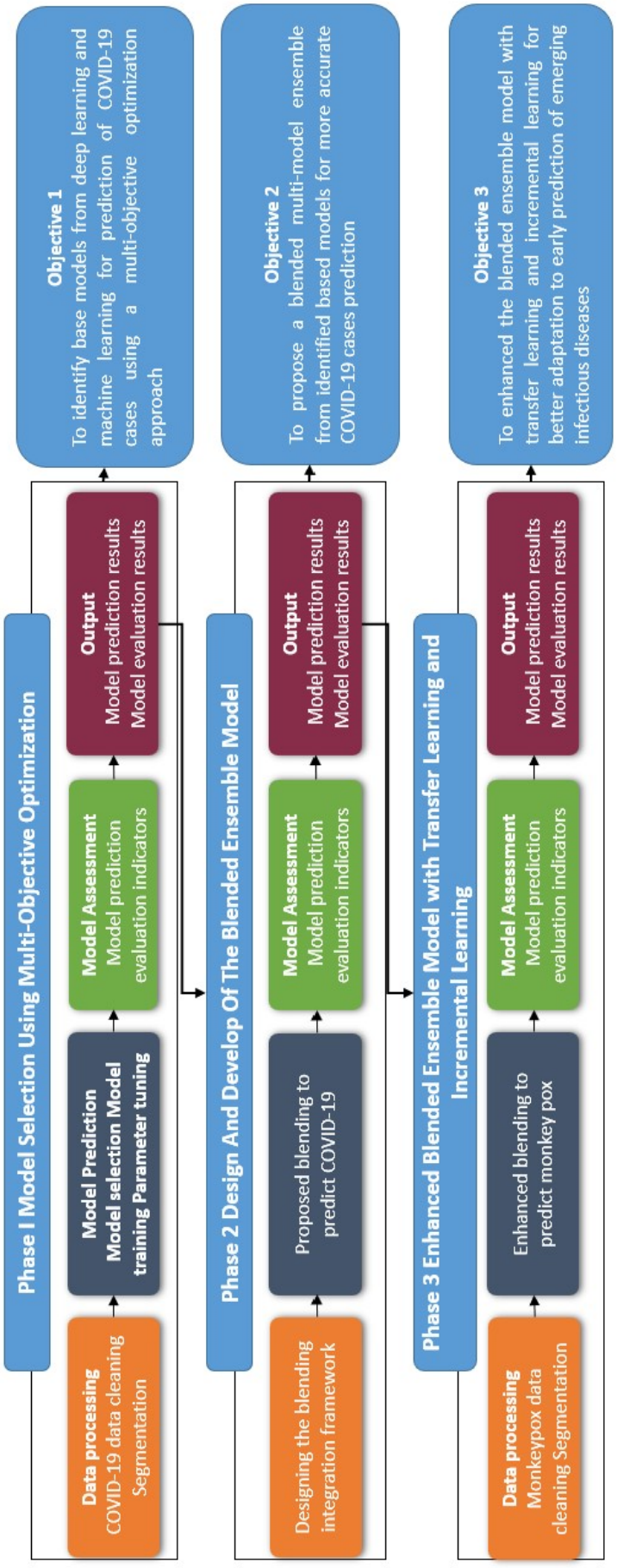


Figure 3.4 Research Methodology Framework

3.3.1 Phase I Model Selection Using Multi-Objective Optimization

In the first stage, the goal of the research model is to select the most appropriate model using multi-objective optimization. The main steps involved in this phase are as follows:

(a) ***Data Processing:*** Clean and segment the COVID-19 data first to ensure that the data quality and structure are suitable for subsequent model training.

(b) ***Model Prediction:*** A variety of models (including deep learning models such as FNN, TCN, CNN, and LSTM, and machine learning models such as RF, ridge regression, decision tree, and XGBoost) are selected for preliminary model training and parameter tuning.

(c) ***Model Selection:*** Apply a multi-objective optimization algorithm to select the best model. In this step the model is evaluated on the basis of its predictive performance and other evaluation metrics such as accuracy, run time, and model complexity.

(d) ***Model Evaluation:*** Evaluates the predictions of the selected model, using specific model evaluation metrics to measure its performance.

(e) ***Output:*** The prediction and evaluation results of the model are output to provide a basis for the next stage of model improvement.

Through these steps, the main goal of the first phase is to establish an effective model selection mechanism that provides a solid foundation for disease prediction.

3.3.2 Phase 2 Design And Develop Of The Blended Ensemble Model

In the second phase, the main goal of the research model is to improve the prediction performance of the model. The key steps in this phase include the following:

(a) ***Design a Blended Ensemble Model:*** Build an ensemble learning model that combines the predictive advantages of multiple models, such as linear

regression as a metamodel, combined with models such as ridge regression, decision tree, and XGBoost. This approach aims to leverage the strengths of each model to improve the overall prediction accuracy by fusing the prediction results of different models.

(b) ***Implementation Of The Blended Ensemble Model:*** implement the designed model to COVID-19 prediction. In this way, the information of different models can be synthesized more effectively, improving the accuracy and reliability of predictions.

(c) ***Model Evaluation:*** After the fusion model is applied, model evaluation metrics are used to evaluate the prediction results. These metrics include, but are not limited to, accuracy, rmse, mae, etc., to obtain a complete picture of how the model performs in real-world applications.

(d) ***Output:*** Outputs the prediction and evaluation results of the fusion model. These results demonstrate not only the performance of a single model, but also the performance gains that can be achieved through model fusion.

Through the above steps, the second phase aims to improve the overall performance of the prediction model through advanced model fusion technology, to more accurately predict the development trend of COVID-19.

3.3.3 Phase 3 Enhanced Blended Ensemble Model with Transfer Learning and Incremental Learning

In the third phase, the research model aims to achieve efficient prediction of emerging infectious diseases. This phase focuses on applying developed and validated blended ensemble model enhanced with transfer learning and incremental learning for the prediction of emerging infectious diseases such as monkey pox. Here are the key steps in Phase 3:

(a) ***Data Processing:*** Cleaning and segmentation of monkey pox data. This step ensures the quality and applicability of the data, providing the basis for accurate predictions.

(b) ***Enhanced Blended Ensemble :*** Apply the blended ensemble developed in Phase 2 to the prediction of monkey pox. By leveraging existing model fusion techniques, it is possible to predict the trend and development of emerging diseases such as monkey pox more accurately.

(c) ***Technology Integration:*** During this phase, techniques such as transfer learning and incremental learning are explored to optimize the model's ability to predict long-term and early-stage infectious diseases. These technologies help models adapt quickly to new disease data and environmental changes, improving the timeliness and accuracy of predictions.

(d) ***Model Evaluation:*** A variety of model evaluation metrics are used to evaluate predictions for monkey pox. This includes analysing the model's accuracy, false positive rate, and prediction stability to ensure that the model remains efficient in new use cases.

(e) ***Output:*** Output model prediction and evaluation results on monkey pox to provide a scientific basis for public health decision-making.

Through these steps, Phase 3 not only strengthens the ability to predict infectious diseases of current concern (such as monkeypox), but also provides an effective set of predictive tools for infectious diseases that may arise in the future, thereby improving the public health response to emerging infectious diseases.

3.4 Proposed Blended Ensemble

In the "Blending Building" process, first, a set of models that exhibit good predictive performance needs to be selected. These models may include deep learning models (e.g., CNNs, and LSTMs) and traditional machine learning models (e.g., random forests, XGBoost, and ridge regression). When choosing a model, its ability to adapt to different types of data and how it performed in previous experiments should be considered. The design of the fusion strategy is the core of building an effective fusion model. In this study, a meta-learner approach was used to integrate the prediction results of each base model by using linear regression as a metamodel. A performance evaluation is conducted by deploying the fusion model on an independent test set. The

evaluation metrics encompass both accuracy and computational efficiency, ensuring a thorough appraisal of the model's predictive capabilities. Furthermore, assessments of model robustness and generalizability are executed to gauge its applicability in real-world contexts (see Fig 3.5).

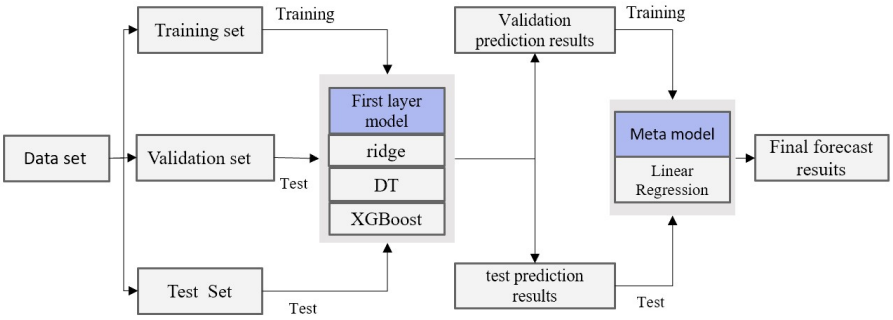


Figure 3.5 Proposed blending model

3.5 Tools and Platforms

In this study, the first phase, the project relied on the following core libraries: scikit-learn, TensorFlow/Keras, and DEAP, which played key roles in data processing, machine learning modeling, deep learning model training, and multi-objective optimization, respectively.Scikit-learn is a Python library widely used in machine learning and data mining, which provides a rich set of tools for data preprocessing, model training and evaluation, and more. The following modules are mainly used in this phase:

(a) Preprocessing module: includes and for data standardization. The scaling of the data is not affected by outliers, but the data is normalized to a distribution with a mean of 0 and a variance of 1, which improves the stability and convergence speed of the model.

(b) Model selection module: used to divide the dataset into a training set and a test set to help evaluate the generalization ability of the model. A hyperparameter

grid search is performed to find the best combination of parameters through cross-validation to optimize the performance of the model.

(c) Evaluation Indicators: and Used to evaluate the prediction effect of the model, measuring the squared mean and absolute mean of the prediction error, respectively.

TensorFlow is a widely used deep learning framework, and Keras is its high-level API that facilitates building and training neural network models. In the first phase, TensorFlow/Keras was used to build a deep learning model. DEAP is a Python library for evolutionary computation and multi-objective optimization, supporting the implementation of genetic algorithms. In the first stage, DEAP is used to achieve multi-objective optimization, using evolutionary algorithms to find the optimal solution to balance multiple evaluation metrics.

Scikit-learn is a widely used Python library for machine learning and data mining that offers a comprehensive set of tools to simplify and optimize the machine learning process. In the second phase, the project relies heavily on the scikit-learn (sklearn) library for tasks such as data processing, feature extraction, model stacking and blending, evaluation, and segmentation of time series data.

In the third phase, the project relies significantly on the scikit-learn (sklearn) library to facilitate the essential steps of data preprocessing, feature selection, transfer learning, model construction, and evaluation.

3.6 Evaluation Indicators

In this study, model evaluation metrics, including the MAE, RMSE, Computational Efficiency, Generalization. MAE is the most basic evaluation metric and reflects the accuracy of the assessment, whereas other metrics are usually used as a reference to compare strengths and weaknesses (Zhang *et al.*, 2020). The RMSE is a measure of the deviation between the predicted and actual values of a model (Guo

et al., 2022). Therefore, the RMSE, and MAE were used as evaluation metrics. The RMSE is expressed as:

$$RMSE = \sqrt{\frac{\sum(\hat{y}_t - y_t)^2}{n}} \tag{3.1}$$

where$\hat{\boldsymbol{y}}\boldsymbol{t}$ is the predicted value of the model,*yt* is the actual observed value, and *n* is the number of observation time points. The range of the RMSE value is (0, +), the smaller the error, the smaller the value, that is, the more approximate the perfect model is. When the RMSE is equal to 0, the actual value and the predicted value are completely consistent, and the model prediction effect is the most perfect (Zhang *et al.*, 2022).

The MAE reflects the actual error of the forecast, and its formula is as follows:

$$MAE = \frac{1}{n}\sum|\hat{y}_t - y_t| \tag{3.2}$$

The MAE value ranges from (0, +), the smaller the error, the smaller the value, i.e. the closer to a perfect model. When MAE is equal to 0, the actual value matches the predicted value perfectly and the model predicts the most perfect effect.

Generalization refers to the model's ability to perform well on unseen data, such as data from a test set. In machine learning, good generalization means that the model can effectively apply to new, untrained data, not just perform well on the training data.

(a) **Evaluation Method:** Similar to accuracy, Root Mean Squared Error (RMSE) is used as a metric to assess generalization. By calculating RMSE on an independent test set, the model's adaptability to new data can be evaluated. The test set should be completely independent of the training process to ensure the objectivity and fairness of the evaluation results.

Computational efficiency refers to the time and resources needed for model training and prediction. In practical applications, efficient computational performance is crucial, especially when dealing with large datasets or operating in resource-limited environments (Chi *et al.*, 2024).

(a) **Evaluation Method:** The total time from the start of training to its completion is measured. This includes the training times for the base models as well as the meta-model. This can be achieved using timing functions available in programming environments, such as the ‘time‘ module in Python. A model with optimal performance should be able to complete training and provide predictions within a reasonable timeframe.

3.7 Chapter Summary

This chapter provides an overview of the methodologies adopted in this study. Initially, the overall research design of the study is presented, followed by a description of the data sources and the research model. The research is divided into three main phases: the first phase involves establishing model predictions and comparing the predictive performance of each model. In the second phase, efforts are made to enhance model prediction performance. The third phase focuses on the efficient prediction of emerging infectious diseases. Finally, this chapter describes the tools and platforms used, as well as the evaluation metrics.

CHAPTER 4

IDENTIFICATION OF BASE MODELS USING MULTI-OBJECTIVE OPTIMISATION

4.1 Introduction

Over the past few decades, with the acceleration of globalization and the increase in population mobility, the outbreak and spread of infectious diseases have become a major challenge for global public health (Fialho *et al.*, 2023; Hernández-Giottonini *et al.*, 2023; Mirzania *et al.*, 2022). From the Severe Acute Respiratory Syndrome (SARS) outbreak in 2003, to the H1N1 influenza pandemic in 2009, and the recent COVID-19 pandemic, outbreaks of infectious diseases have not only had a tremendous impact on human health but have also posed unprecedented challenges to the world economy and social stability (Cao *et al.*, 2022; Gao *et al.*, 2022; Li *et al.*, 2023a; Yang *et al.*, 2022b). Therefore, effective infectious disease prediction models are crucial for the prevention and control of epidemic outbreaks. the study enable public health decision-makers to take proactive measures and mitigate the negative impacts of the epidemic (Dixon *et al.*, 2022; Lv *et al.*, 2021; Zhao *et al.*, 2022).

However, despite significant advances in this field in recent years, existing infectious disease prediction models still have some non-negligible limitations (Hu *et al.*, 2023; Li *et al.*, 2020; Liao *et al.*, 2022; Tian *et al.*, 2021). Among them, most models have adopted single-objective optimization approaches, focusing primarily on enhancing prediction accuracy while neglecting other critical factors such as the model's generalizability, computational efficiency, and feasibility in practical applications (Khoo *et al.*, 2024; Tsai *et al.*, 2021; Xia *et al.*, 2022; Ye *et al.*, 2020; Akbulut *et al.*, 2023). This pursuit of a single objective may lead to limitations in the application of the model under specific circumstances, failing to fully meet the complex demands of the public health sector (Akbulut *et al.*, 2023; Sassano *et al.*, 2022).

In response to these challenges, Multi-Objective Optimization (MOO) offers a new solution. Multi-objective optimization is a method designed to simultaneously optimize multiple conflicting objectives, capable of generating a set of optimal solutions that achieve the best trade-off among the objectives (i.e., Pareto optimal solutions) (Le Fouest and Mulleners, 2024; Mohammed *et al.*, 2023). In the context of infectious disease prediction, incorporating multi-objective optimization allows models to simultaneously consider prediction accuracy, computational efficiency, and generalizability to new data, thereby enhancing the overall performance of the models (Feng and Zhang, 2023; Liu *et al.*, 2024). Additionally, multi-objective optimization has been proven to be an effective method for improving decision quality in other fields, such as engineering design, resource allocation, and environmental management (Huang *et al.*, 2024; Wang *et al.*, 2023d).

This study aims to explore and empirically demonstrate the application of multi-objective optimization methods in selecting infectious disease prediction models, addressing the challenges faced by traditional single-objective optimization methods. By comprehensively considering various aspects of model performance, this research aims to enhance prediction accuracy, focusing on the model's generalizability and feasibility in practical applications. This approach provides a more comprehensive and reliable scientific basis for public health decision-making.

The main contributions of this study are as follows: First, This study propose an infectious disease prediction model that incorporates multi-objective optimization. This approach achieves a balance among multiple performance indicators and enhances the overall predictive capability of the model. Secondly, through empirical research, This study showcase the effectiveness of multi-objective optimization methods in enhancing the accuracy and stability of infectious disease predictions. Finally, the findings of this study offer new tools and insights for researchers in the field of infectious disease prediction and for public health decision-makers. This contributes to the scientific rigor and effectiveness of epidemic response strategies.

In summary, this study not only emphasizes the importance and application prospects of multi-objective optimization in infectious disease prediction but also highlights the urgency and significance of the research. Through this study, This

study aim to contribute to the development of infectious disease prediction models and provide stronger scientific support for global public health security.

4.2 Data Processing

This study utilized the Mexican COVID-19 time series dataset provided by Our World in Data (Karlinsky and Kobak, 2021). This dataset covers the period from April 1, 2020, to March 31, 2023, and includes various key indicators such as daily new confirmed cases (new_cases), total confirmed cases (total_cases), daily new deaths (new_deaths), total deaths (total_deaths), along with smoothed data and ratios calculated per million people. Such datasets are widely used in epidemiological research due to their completeness and accuracy. Based on the correlation analysis of the dataset, This study selected indicators that are highly correlated with the daily new confirmed cases (new_cases) as the main variables (see Figure 4.1). This figure displays a matrix where both rows and columns are labeled with the same set of features. The cells of the matrix are colored based on the correlation coefficient between two features. Red indicates a positive correlation, while blue represents a negative correlation. These indicators have significant predictive value in forecasting models, as identified by Husnayain *et al.* (2021); Mathieu *et al.* (2020); Sharma *et al.* (2022); Wang *et al.* (2022b); Zhang *et al.* (2023b). Given the presence of missing values in the original dataset, a median imputation method was used to fill these gaps. Due to the unique characteristics of infectious disease time-series data, outliers were retained to preserve the integrity and authenticity of the epidemiological trends.

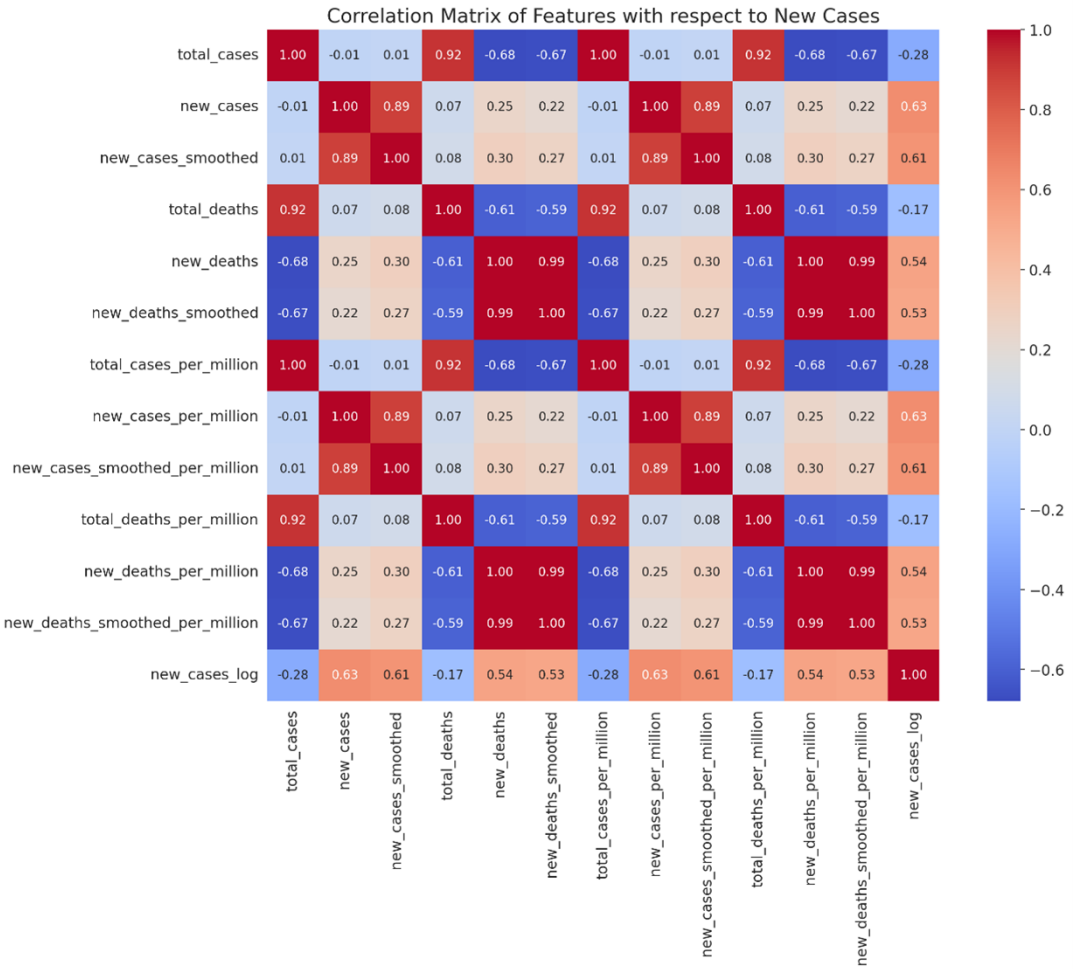


Figure 4.1 Correlation Matrix Heatmap of New Cases.

To further analyze and enhance model performance, This study performed a logarithmic transformation on the daily new confirmed cases (new_cases_log). Logarithmic transformation is a commonly used numerical processing technique that is effective in reducing the skewness of non-normally distributed data, as mentioned by Vukašinović *et al.* (2023); West (2022). This study also conducted thorough data cleaning. For the data processing of each model, This study followed the same steps. Initially, This study used RobustScaler for data scaling, a normalization method known for its robustness to outliers, as noted by Changyong *et al.* (2014); West (2022). Furthermore, This study standardized the data to ensure the feature values conform to a distribution with a mean of 0 and a standard deviation of 1, as recommended by Oka (2021); Wang *et al.* (2023c). In dividing the training and test sets, This study used

an 80% and 20% ratio, a practice supported by Joseph (2022), who argue that this split effectively maintains the sequentiality and temporal coherence of time series data.

4.3 Model Selection

This study selected several machine learning and deep learning models that are widely used in the field of time series prediction, based on their performance in existing literature and successful application in similar problems, as indicated by Ahmed *et al.* (2022); Lim and Zohren (2021) (see Fig4.2).

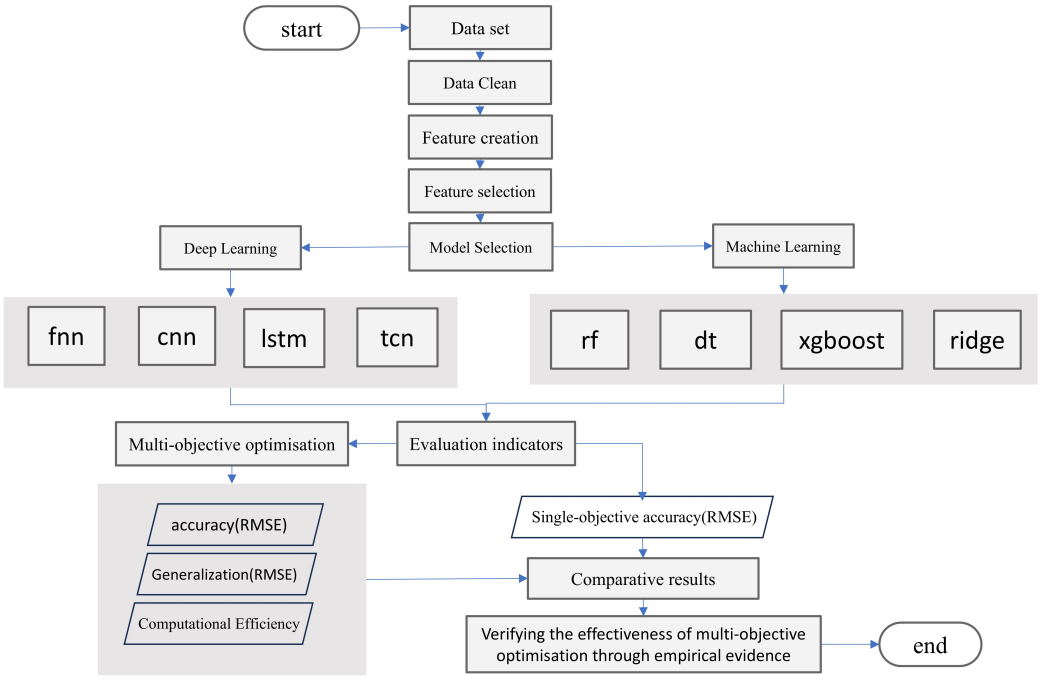


Figure 4.2 Methodology Flowchart of Identification of Base Model

4.3.1 Deep Learning Model

4.3.1.1 Feedforward Neural Networks (FNN)

FNN is a basic form of an artificial neural network, widely used in various machine learning tasks. FNN features include being simple and efficient, easy to implement, and debug (see Table 4.1).

Table 4.1 FNN model structure and parameters

Parameter type	descriptions
Model Structure and Parameter Settings	Input Layer: Set according to the dimension of the feature data Fully Connected Layers: Use Dense layers with adjustable neuron numbers, employing the ReLU activation function.Dropout Layer: Incorporate a Dropout layer with a fixed ratio of 0.2 to reduce overfitting. Output Layer: Use a single-neuron Dense layer for predicting the target variable
Hyperparameter Tuning	Learning Rate Options: 0.001, 0.01, 0.1 Neuron Number Options: 32, 64, 128 Batch Size Options: 16, 32, 64
Search Process	Iterate through various combinations, selecting the one that minimizes RMSE on the training set
Training Configuration	Utilize the Adam optimizer, with a training duration of 100 epochs, setting the batch size according to the optimal parameters.
Optimal Parameters	Learning rate of 0.001, 128 neurons, and a batch size of 16

4.3.1.2 Convolutional Neural Networks (CNN)

CNN is a neural network model widely used in the field of deep learning and is particularly effective at processing image data. It effectively recognizes patterns and features in images by mimicking the workings of the human visual system. Although originally designed for image analysis, Convolutional Neural Networks (CNNs) have also been successfully applied to the processing of time series data. It shows potential for processing a wide variety of sequence data by learning local temporal features within the data (see Table 4.2).

Table 4.2 CNN model structure and parameters

Parameter type	descriptions
Model Architecture	Input Layer: Set input shape according to the feature dimensions of the time series data. Convolutional Layer: Use a single convolutional layer with adjustable numbers of filters and kernel sizes. Pooling Layer: Add a MaxPooling layer with a fixed pooling size of 2. Flatten Layer: Data outputted from the convolutional and pooling layers are transformed into one dimension through the Flatten layer. Fully Connected Layer: Add a Dense layer with adjustable unit numbers, employing the ReLU activation function.Output Layer: A single-neuron Dense layer for predicting the target variable.
Training Configuration	Optimizer: Use the Adam optimizer. Learning Rate: Selected based on the results of a Keras Tuner search. Training Epochs: Train for 50 epochs on the training set. Batch Size: Set to 32
Parameter Search with Keras Tuner	Define a model function using Keras Tuner's hyperparameters (hp) to define the model structure and search space
Configuration	Use RandomSearch, targeting validation loss, with a maximum of 5 trials and 3 epochs per trial Search: Conduct training for 10 epochs
Optimal Parameters	conv_1_filter: 112, conv_1_kernel: 3, dense_1_units: 96, learning_rate: 0.001

4.3.1.3 Long Short-Term Memory Networks (LSTM)

Long Short-Term Memory (LSTM) is a deep learning model designed to address the long-term dependency issue in traditional recurrent neural networks (RNNs). LSTM effectively captures long-term relationships in time-series data by introducing special structural units, enabling it to demonstrate excellent performance in processing complex sequence data with temporal extensibility (see Table 4.3).

Table 4.3 LSTM model structure and parameters

Parameter type	descriptions
Model Architecture	Input Layer: Set input shape according to the feature dimensions of the time series data. LSTM Layer: Utilize LSTM units with adjustable numbers. Dropout Layer: Added after the LSTM layer, with an adjustable ratio. Output Layer: A single-neuron Dense layer for prediction
Hyperparameter Search	Search for combinations of different unit numbers, dropout ratios, and learning rates.Select the combination that minimizes RMSE on the test set.Parameter Combination Options: Unit numbers (50, 100, 150), Dropout ratios (0.2, 0.3, 0.4), Learning rates (0.001, 0.0005, 0.0001).
Training Configuration:Optimizer	Use the Adam optimizer.Learning Rate: Selected based on search results. Training Epochs: Train for 100 epochs on the training set.Batch Size: Set to 32.
Optimal Parameters	Unit Number: 100, Dropout Ratio: 0.3, Learning Rate: 0.001

4.3.1.4 Temporal Convolutional Network (TCN)

Temporal Convolutional Networks (TCN) offer a unique architecture well-suited for sequential input, especially in complex clinical decision support settings that involve time-series data. TCN gained popularity for its state-of-the-art performance across various applications (see Table 4.4).

Table 4.4 TCN model structure and parameters

Parameter type	descriptions
Model Architecture	Input Layer: Set according to the time series feature dimensions of the training data.TCN Layer: Utilize Temporal Convolutional Network layers to process time series data, with parameters including the number of filters, kernel size, number of stacks, and dilation.Flatten Layer: Data outputted from the TCN layer is transformed into one dimension through the Flatten layer. Output Layer: A Dense layer using a linear activation function for predicting the target variable
Parameter Setting and Manual Parameter Search	Manually iterate through various parameter combinations, including different numbers of filters, kernel sizes, stack numbers, and dilation options.Select the combination that minimizes RMSE on the validation set. Parameter Combination Options: Number of filters (32, 64, 128), kernel size (2, 3), number of stacks (1, 2), dilation options ([1, 2, 4, 8] and [1, 2, 4, 8, 16])
Training Configuration	Optimizer: Use the Adam optimizer. Learning Rate: Set to 0.002. Training Epochs: Train for 50 epochs on the training set.Batch Size: Set to 16
Optimal Parameters	Number of filters: 64, Kernel Size: 3, Number of Stacks: 1, Dilation: [1, 2, 4, 8]

4.3.2 Machine Learning Model

4.3.2.1 Random Forest (RF)

Random Forest (RF) is a widely used ensemble technique that utilizes a multitude of decision-tree classifiers. It operates on various sub-samples of a dataset with random subsets of features for node splits. This method enhances predictive accuracy and controls overfitting by using majority voting for classification problems or averaging for regression problems. Random Forest (RF) is particularly effective because of its ability to process large amounts of data with high accuracy (see Table 4.5).

Table 4.5 RF model structure and parameters

Parameter type	descriptions
Hyperparameter Search	Parameter Grid: n_estimators (50, 100, 150) , max_depth (None, 10, 20, 30), min_samples_split (2, 5, 10), min_samples_leaf (1, 2, 4). Conduct parameter search using GridSearchCV, combined with 5-fold cross-validation, and the evaluation criterion being negative mean squared error
Optimal Parameter	max_depth None, min_samples_leaf 1, min_samples_split 2, n_estimators 50.

4.3.2.2 DecisionTree (DT)

The Decision Tree Regressor is a type of decision tree used for regression tasks. This model is renowned for its interpretability and effectiveness in capturing non-linear relationships in data. Decision tree regressors can handle both categorical

and continuous input and output variables, making them versatile for a wide range of regression problems (see Table 4.6).

Table 4.6 Decision Tree Regressor model structure and parameters

Parameter type	descriptions
Hyperparameter Search	Parameter Grid: criterion ('squared_error', 'friedman_mse','absolute_error'), splitter ('best', 'random'), max_depth (None, 10, 20, 30, 40, 50), min_samples_split (2, 5, 10), min_samples_leaf (1, 2, 4).Conduct hyperparameter search using GridSearchCV, combined with 5-fold cross-validation
Optimal Parameter	criterion 'absolute_error', max_depth 10, min_samples_leaf 1, min_samples_split 2, splitter 'best'.

4.3.2.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a highly efficient and effective open-source implementation of the gradient boosting algorithm. It is particularly popular for its computational efficiency and strong performance in structured or tabular datasets for classification and regression predictive modeling problems (see Table 4.7).

Table 4.7 XGBoost model structure and parameters

Parameter type	descriptions
Hyperparameter Search	Parameter Grid: n_estimators (50, 100, 200) , max_depth (None, 10, 20, 30), learning_rate (0.01, 0.1, 0.2). Conduct hyperparameter search using Randomized SearchCV, combined with 5-fold cross-validation and 50 iterations
Optimal Parameter	n_estimators 200, max_depth None, learning_rate 0.1.

4.3.2.4 Ridge Regression

Ridge regression, also known as Tikhonov regularization, is a method used to estimate the coefficients of multiple regression models in situations where linearly independent variables are highly correlated. It introduces a penalty term to the loss function: the squared magnitude of the coefficient multiplied by the regularization parameter. This approach is particularly useful in mitigating the problem of multicollinearity in linear regression models, thereby enhancing the model's prediction accuracy and interpretability (see Table 4.8).

Table 4.8 Ridge model structure and parameters

Parameter type	descriptions
Hyperparameter Search	defaults
Optimal Parameter	defaults

4.4 Multi-Objective Optimization

This study focus on enhancing the performance of infectious disease prediction models through multi-objective optimization methods. Specifically, This study aim to optimize the model's performance in three aspects simultaneously: the root mean square error (RMSE) of accuracy, the RMSE of generalizability, and computational efficiency (model training time). These three objectives are often conflicting; for example, enhancing accuracy and generalizability may reduce computational efficiency, leading to an increase in model training time (Cui *et al.*, 2022; Du *et al.*, 2024). Therefore, the aim of this study is to find the optimal trade-off among these three objectives (see Table 4.9).

Table 4.9 Model Prediction Performance Metrics

Name	**accuracy (RMSE)**	**Generalization (RMSE)**	**Computational Efficiency**
FNN	152.025	4532.84	26.19 seconds
TCN	231.014	2477.45	16.65 seconds
LSTM	170.442	10307.7	35.16 seconds
CNN	1144.36	5455.59	10.16 seconds
RF	18.5805	1251.51	0.43 seconds
DT	24.4889	8.1421	0.07 seconds
XGBOOST	33.0709	13.0672	0.38 seconds
Ridge	11.5274	114.852	0.03 seconds

Performance metrics such as accuracy RMSE, generalizability RMSE, and computational efficiency (model training time) were acquired through the utilization of the chosen model, as delineated in Table 1. In this investigation, the NSGA-II (Non-dominated Sorting Genetic Algorithm II) within the Genetic Algorithm (GA) model was opted for as the multi-objective optimization algorithm. The selection of NSGA-II was based on its efficacy in managing multi-objective optimization predicaments, particularly in upholding solution diversity and pinpointing the Pareto front (Hu *et al.*,

2022; Liu *et al.*, 2023a; Padilla-García *et al.*, 2023). Furthermore, the non-dominated sorting and crowding distance mechanisms of NSGA-II empower it to efficiently recognize a collection of optimal solutions in extensive search spaces. This capability is especially vital for determining the optimal trade-offs among diverse performance indicators in infectious disease prediction models (Bolla *et al.*, 2023; Entezari *et al.*, 2023; Li *et al.*, 2022).

The model selection methodology employed in this research is executed through the NSGA-II algorithm, utilizing the DEAP (Distributed Evolutionary Algorithms in Python) library. The model comprises the subsequent essential stages:

(a) ***Population Initialization:*** The population is initialized by randomly selecting algorithm parameters or model configurations.

(b) ***Fitness Evaluation:*** The evaluation function calculates the accuracy RMSE, generalizability RMSE, and computational efficiency (model training time) for each individual (i.e., model configuration) in the population.

(c) ***Genetic Operations:*** Crossover (cxPassThrough) and mutation (mutate) operations are applied to generate new individuals, exploring the solution space.

(d) ***Selection Mechanism:*** The next generation of the population is selected using the selection mechanism (select) in the NSGA-II algorithm, based on non-dominated sorting and crowding distance.

(e) ***Iterative Optimization:*** Repeat the above process until a predetermined number of iterations or other stopping conditions are reached.

The experimental configuration consists of a population size of 100 individuals evolving over 50 generations, with a crossover probability of 0.7 and a mutation probability of 0.3. The optimization process is conducted through the utilization of the 'run_algorithm' function. The outcomes are then visually depicted using the 'plot_pareto_front_with_labels' function, which showcases the trade-offs between accuracy RMSE, generalizability RMSE, and computational efficiency (specifically model training time). This graphical representation serves to elucidate the interplay

of various objectives and the efficacy of the NSGA-II algorithm in achieving a harmonious balance among them, thereby facilitating informed decision-making in the model selection process.

4.5 Result and Discussion

4.5.1 Multi-Objective Optimisation Result

The outcomes of the multi-objective optimization conducted in this research can be effectively illustrated through a comprehensive examination using Pareto front analysis. As depicted in Figure 4.3, the positioning of different models' performance in the multi-dimensional objective space is accurately represented, encompassing metrics such as the root mean square error (RMSE) of prediction accuracy, RMSE of generalizability, and computational efficiency (model training time). These parameters are used in graphical representations to assess and compare different prediction models, providing a clear visualization of how models handle the trade-offs among these conflicting objectives.

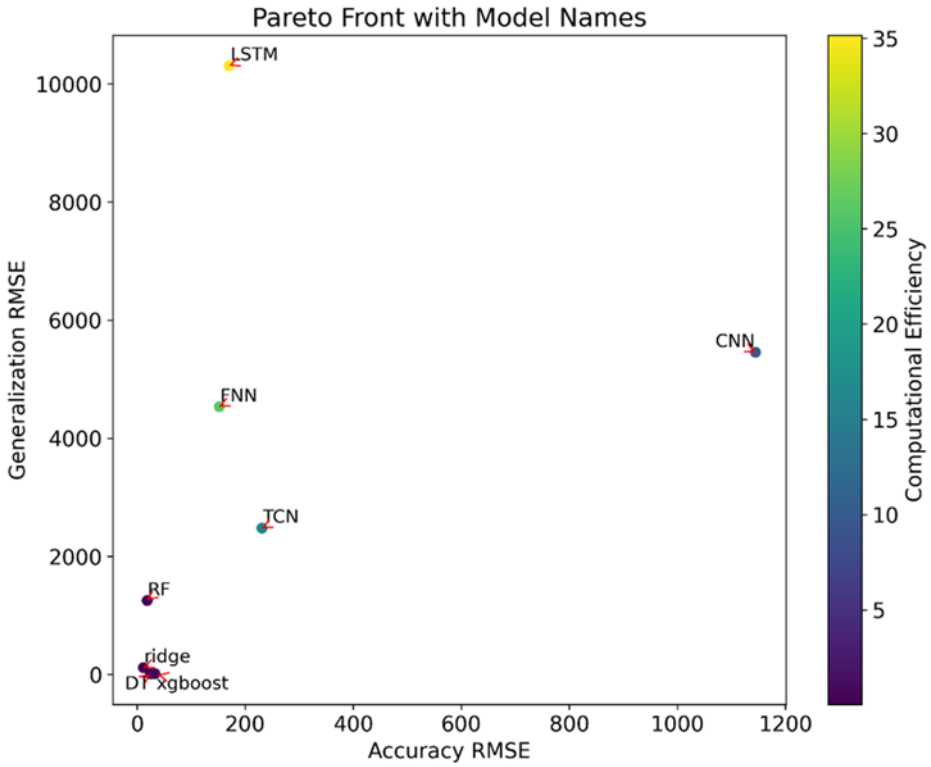


Figure 4.3 Multi-objective optimisation of the Pareto frontiers

The Pareto front analysis visualizes the trade-offs between the following metrics:

(a) ***Accuracy RMSE:*** This metric measures the precision of the model when predicting data. On the x-axis of the chart, a lower RMSE value indicates higher prediction accuracy.

(b) ***Generalization RMSE:*** This metric reflects the model's ability to generalize to new data. A lower RMSE value on the y-axis indicates better generalizability.

(c) ***Computational Efficiency (Model Training Time):*** This is represented by a color bar, where the gradient (from purple to yellow) shows the change in computational efficiency from high to low.

In the chart, different points represent different models. The position of each point is determined by its RMSE values for accuracy and generalizability, while the color indicates its computational efficiency. Key observations include:

(a) Models 'DT' and 'XGBoost' both exhibit lower RMSE values for accuracy and generalizability, while also maintaining relatively high computational efficiency (closer to purple), suggesting the study may be well-balanced models optimizing accuracy, generalizability, and computational efficiency.

(b) The 'Ridge' model shows lower values in terms of accuracy and computational efficiency but slightly lacks in generalizability compared to 'DT' and 'XGBoost'.

(c) The 'LSTM' model shows the lowest generalizability and computational efficiency, while the 'CNN' model has the lowest accuracy.

These insights from the graphical representation help evaluate and compare the overall performance of different prediction models, assisting decision-makers in selecting the most suitable model based on specific needs and constraints.

4.5.2 Comparison Of Single-Objective Optimisation Models

This study further explores the performance differences between multi-objective optimization methods and traditional single-objective optimization methods for selecting infectious disease prediction models. As shown in Fig. 4.4, This study observed the change in fitness of the single-objective optimization method during the evolution process, which includes the evolution of minimum fitness and average fitness.

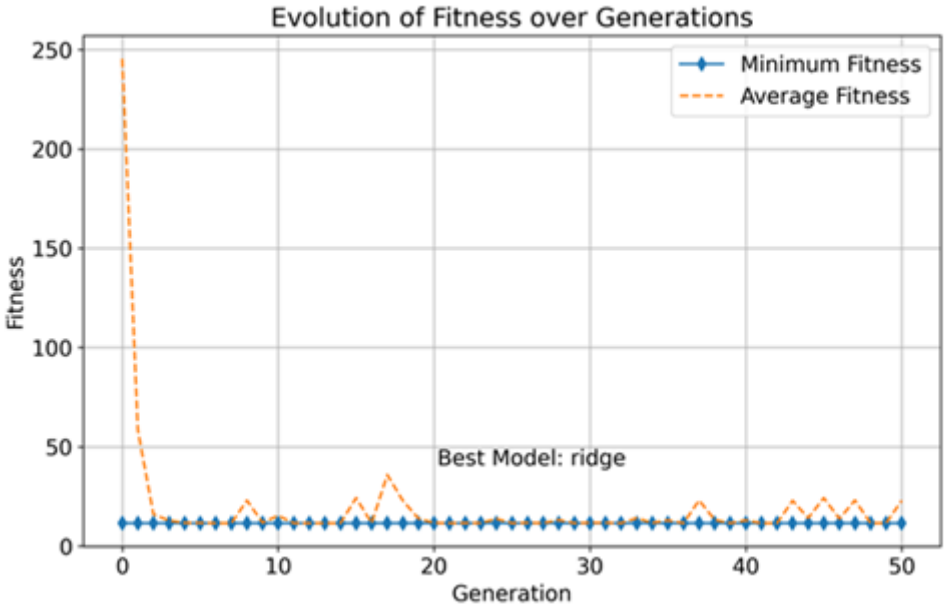


Figure 4.4 Evolution of fitness over generations

In the early stages, the minimum fitness decreases rapidly, indicating that the optimization algorithm can quickly identify and facilitate the evolution of high-performance solutions. This rapid evolutionary progress demonstrates the efficiency of the single-objective optimization approach in exploring the solution space. As generations increase, both the minimum and average fitness stabilize, demonstrating the robustness of the algorithm in consistently finding optimal solutions. The best model, "ridge," is identified after 50 generations, and it excels in all objectives, showcasing its superior overall performance.

To validate the effectiveness of the optimization process, predictions were made on the test dataset using the optimal model, ridge, derived from the single-objective optimization algorithm. The predicted values generated by the model on the test dataset closely align with the true values, indicating that the optimal model selected through single-objective optimization performs well in terms of predictive accuracy (see table 4.10). Additionally, the results of the Root Mean Square Error (RMSE) further demonstrate that the model successfully meets the optimization objective of

minimizing error, which is consistent with the outcomes of the optimization process illustrated in Figure 4.4.

Table 4.10 Single-objective optimisation verification

actual value	predicted value
5618.00	5617.94
4126.00	4130.08
3755.00	3758.24
3281.00	3283.29
2946.00	2947.27
1140.00	1139.79
1108.00	1106.53
3830.00	3828.78
3075.00	3074.75
2599.00	2598.17

4.5.3 Case Study

In order to validate the effectiveness of the method, this study utilized the infectious disease dataset of COVID-19 in Indonesia and Iran to assess the performance of the model chosen through the multi-objective optimization method against the model selected through the single-objective optimization method in a real-world scenario. Through the multi-objective optimization approach, This study selected the decision tree (DT) and gradient-boosted tree (XGBoost) as the best models. In contrast, the ridge regression model was chosen as the best model through the single-objective optimization approach.

In the actual scenario (refer to Table 4.11), the decision tree model demonstrates good performance with the lowest root mean square error (RMSE), indicating its high sensitivity to the data and strong predictive capability. Despite having a higher RMSE value compared to the decision tree model, the XGBoost model

exhibits superior predictive performance. In contrast, the RIDGE model selected by the single-objective optimization method has a relatively high RMSE value on this dataset.

Table 4.11 Comparison of model performance in predicting RMSE for COVID-19 in Iran and Indonesia

Model Name	Iran(RMSE)	Indonesia(RMSE)
DT	8.1421	6.7367
XGBoost	13.0672	13.5128
RIDGE	114.8524	16.2552

This case study demonstrates that a multi-objective optimization strategy may be better appropriate for picking a model with greater utility when dealing with a real-world challenge of predicting infectious illness data. The good performance of the decision tree model emphasizes the necessity of taking into account numerous performance measures when constructing a prediction model, rather than focusing exclusively on lowering prediction error.

4.5.4 The impact of model complexity on model performance

An increase in model complexity often improves the model's ability to fit. For example, by increasing the depth of the tree, a decision tree model can capture more features and patterns, thereby improving the prediction accuracy on the training set. However, models that are too complex tend to overfit, i.e., perform poorly on test sets or new data because the study may capture noise in the data rather than actual patterns (Barea-Sepúlveda *et al.*, 2023). Similarly, XGBoost improves prediction accuracy by integrating multiple trees, i.e., gradient boosting decision trees. Although an increase in complexity (e.g., more trees, greater depth) often improves the accuracy of the model, it can also lead to overfitting (Budholiya *et al.*, 2022).

Increasing the complexity of a model may reduce its Generalization. For decision tree models, structures that are too complex perform poorly in the face of new data because the study may have overfitted the training data. Moderate pruning and parameter tuning can help decision trees maintain good Generalization on different datasets (Kozyrev *et al.*, 2023). In the case of XGBoost, a modest increase in complexity can improve the Generalization of the model, as XGBoost employs regularization techniques to prevent overfitting. However, overly complex models may still perform worse than training data on new data (Hlongwane *et al.*, 2024).

More complex decision trees require more computational resources to train and predict. A tree structure with a large depth will increase the computation time and storage requirements, which will affect the efficiency of the model (Yang *et al.*, 2022a). Similarly, complex XGBoost models (more trees, more depth) can significantly increase computational time and resource requirements. Although XGBoost is optimized for computational efficiency, overly complex models still increase computational costs (Silvestri *et al.*, 2023).

Increasing model complexity (e.g., deeper trees, higher tree counts) often improves the accuracy of the training data, but this can lead to overfitting and reducing generality. Proper regularization and pruning techniques can help find a balance between accuracy and Generalization (Alalayah *et al.*, 2023). More complex models tend to have higher accuracy, but also require more computational resources and time. A trade-off between prediction accuracy and computational efficiency needs to be made based on the needs of the actual application scenario. For example, in applications that require real-time prediction, some accuracy may need to be sacrificed to ensure computational efficiency (Papafotis *et al.*, 2021).

4.5.5 Adaptability of different types of infectious diseases and different data characteristics

The multi-objective optimization approach has shown remarkable adaptability when processing data for different pathogen types. For example, in infections of

different genotypes or serotypes, these methods can significantly improve the accuracy of inference for different pathogen types by optimizing laboratory surveillance networks. showed that by optimizing the HFMD surveillance network, the multi-objective optimization approach can significantly reduce the mean square error of estimating serotype-specific incidence, thereby improving the performance of the surveillance network (Zhu *et al.*, 2022).

Different data features perform differently in the multi-objective optimization method. For example, when applying the multi-objective optimisation approach for COVID-19 prediction model selection, the generalisation capability of the model was incorporated, resulting in the selection of a model that not only performs well in COVID-19 prediction, but also performs well in terms of model performance in the prediction of new infectious diseases.

Some multi-objective optimization methods may not perform well when dealing with data-intensive, computationally complex problems. For example, while a multi-objective optimization approach can find a balance between different objectives, there may still be trade-offs between computational efficiency and data complexity. Showed that although the multi-objective optimization method performs well in solving complex environmental/economic power dispatch problems, its computational complexity is still a major challenge (Tan *et al.*, 2017).

4.5.6 Comparison of different algorithms

When choosing a prediction model for infectious diseases, different multi-objective optimization algorithms have their own advantages and disadvantages:

(a) ***NSGA-II (Non-Dominant Sequencing Genetic Algorithm II)*:** Good convergence and hybridization: NSGA-II is capable of generating high-quality Pareto leading-edge solutions for a wide range of optimization problems (Liu *et al.*, 2023a). Excellent performance in many complex optimization problems, such as biological learning systems, traffic data analysis, etc. Shortcoming:

In a multi-objective optimization problem, the selection pressure of NSGA-II decreases significantly as the number of targets increases, affecting the convergence ability of the algorithm (Kumari *et al.*, 2019).

(b) ***SPEA2 (Intensity Pareto Evolution II)*:** SPEA2 excels at maintaining the diversity of solutions, maintaining the diversity of the Pareto front through intensity and distance measurements. It is especially suitable for design problems that require a high diversity of solutions (Cai *et al.*, 2022a). Shortcoming: Although SPEA2 performs well in terms of diversity, it may not converge as well as NSGA-II in some issues (Babor *et al.*, 2023).

(c) ***MOEA/D (Decomposition-based Multi-Objective Evolutionary Algorithm)*:** MOEA/D optimizes multi-objective problems by decomposing them into multiple single-objective subproblems and solving them in parallel. This method excels in dealing with complex multi-objective problems, especially in high-dimensional object spaces (Liu and Ye, 2023). Shortcoming: MOEA/D may require more complex parameterization and higher computational resources for some problems (Sun, 2023).

4.5.7 The long-term validity of the model and the impact of new data

Due to its simplicity and explanatory nature, decision tree models can effectively deal with complex data features in long-term forecasts (Lange, 2023). the study can capture complex nonlinear relationships by recursively segmenting data, which can help with long-term prediction. Over time and as new data emerges, decision tree models may need to be updated frequently to maintain predictive performance (Zhang *et al.*, 2024a). These models perform mediocre in dealing with data drift because the study are not flexible enough for emerging patterns. By integrating the advantages of multiple trees, XGBoost is able to capture more complex patterns and exhibit greater robustness to long-term data changes (Li *et al.*, 2024a). Although XGBoost has shown strong adaptability to new data, it still needs to be retrained regularly to ensure that the model adapts to changing infectious disease transmission patterns.

In the face of new data, the decision tree model may be difficult to adapt to the new model due to the fixed model structure. This requires frequent model updates and retraining to maintain prediction performance. Thanks to its gradient boosting mechanism, XGBoost is better able to adapt to new data. However, with the increase of data volume and the change of characteristics, it is still necessary to regularly adjust and optimize the hyperparameters to maintain high prediction accuracy.

4.5.8 Challenges and strategies for integrating optimization models into public health decision-making systems

As new data continues to pour in, predictive models need to be updated and maintained frequently to ensure their accuracy and usefulness. This can be a challenge for public health agencies with limited resources. To do this, an automated data update and model retraining process is needed to ensure that the model can respond to new data and changes in a timely manner. At the same time, the necessary technical support and training are provided to improve the technical capacity of public health workers.

Effective deployment and use of predictive models requires cross-sectoral collaboration, including the involvement of multiple stakeholders, including governments, healthcare organizations, technology providers, and more. Lack of policy support and coordination mechanisms can lead to deployment failures. Therefore, it is necessary to establish a cross-sectoral collaboration mechanism and policy support model to ensure that all parties can work closely together in the deployment and use of the model. At the same time, clear standards and guidelines should be developed to regulate the development and application of models.

4.5.9 Advantages of multi-objective optimisation methods

This study demonstrates the significant advantages of the multi-objective optimisation approach in the prediction of infectious disease data. Firstly, the multi-objective optimisation approach significantly improves the usefulness and adaptability

of the model by considering multiple evaluation criteria for model selection, rather than based on a single metric alone (Khatun *et al.*, 2022). For example, the Decision Tree (DT) model and the XGBoost model not only perform well in terms of prediction accuracy, but also maintain high computational efficiency, showing excellent overall performance.

In addition, the multi-objective optimisation approach supports a comprehensive evaluation of the model, providing a model for decision makers to weigh various performance metrics (Zhao *et al.*, 2024). For high-risk decision support systems such as infectious disease prediction models, there is a real need to carefully balance the accuracy, generalisation ability and computational efficiency of the model in order to improve its usefulness, as seen in the case of COVID-19 outbreak data prediction in Indonesia.

A comparison of existing studies shows that although multi-objective optimization has been widely used and studied in other fields, relatively little research has been conducted on model selection for infectious disease prediction. Previous work has focused on improving a single performance metric, such as prediction accuracy, while insufficient attention has been paid to the generalization ability and computational efficiency of the model. In contrast, the methodology of this study not only considers prediction accuracy but also integrates other important performance metrics of the model, providing a more comprehensive evaluation model for infectious disease prediction.

In this study, the multi-objective optimization methods employed surpass most existing methods in their ability to find the optimal equilibrium between multiple indicators. Furthermore, through empirical studies, This study demonstrate that in real health crises, such as the COVID-19 pandemic, the utility of these methods far exceeds that of traditional single-objective optimization methods. This is particularly important in assessing the development of epidemics and guiding public health strategies.

4.5.10 Practical implications of model selection

In the context of public health, the choice of infectious disease prediction models is related to the rational allocation of resources, the timely implementation of preventive measures, and the efficiency of emergency response (Piscitelli and Miani, 2024). Models selected by multi-objective optimisation methods, such as the Decision Tree (DT) and XGBoost models that perform well in this study, provide policy makers with accurate and timely information on the development of epidemics due to the good balance between accuracy, generalisability and computational efficiency. This helps the government and public health organisations to develop more scientific and effective response strategies in the face of limited medical resources and the need to make quick decisions.

High-quality model predictions can enhance the accuracy of outbreak early warning systems, thereby guiding communities in the early stages of an outbreak to take measures to slow down the spread of the virusCai *et al.* (2023). For example, the ability of DT models to be understood and trusted by non-technical people due to their simplicity and easy-to-understand decision rules is a non-negligible advantage in outbreak management.

In real-world scenarios, models must also be able to adapt to changing data and sudden outbreak developments. In the case of this study, the DT model showed a high level of adaptability, which emphasises the importance of considering not only the accuracy of the predictions, but also the ability of the model to adapt to new data when selecting a model (Zhang *et al.*, 2023a). This flexibility in modelling is essential for monitoring outbreaks in real time and predicting their trends.

4.5.11 Limitations of the study

In this study, a multi-objective optimisation approach was used in the development and evaluation of infectious disease prediction models, and although positive results were obtained, several limitations existed:

(a) Firstly, only the COVID-19 dataset was used in this study, which limits the assessment of the generalisation ability of our models. Although the models selected by the multi-objective optimisation approach performed well on the COVID-19 dataset, This study cannot confidently predict the performance of these models under other different infectious disease conditions.

(b) Secondly, the NSGA-II algorithm was chosen as the tool for multi-objective optimisation in this study, which may have influenced the results of the optimisation. Other multi-objective optimisation algorithms may have produced different Pareto fronts and final set of models chosen, which implies that our findings were limited by the chosen algorithm. In addition, the experimental design and setup may also affect the interpretation of the results. For example, the choice of evaluation metrics may have an impact on the optimisation process and the final results.

4.6 Chapter Summary

The primary goal of this research is to create and verify a multi-objective optimisation model to improve predictive model selection for infectious disease data. By combining numerous criteria such as prediction accuracy, generalization ability, and computational efficiency, This study verify DT and XGBoost as models that perform well on the COVID-19 dataset. These models outperformed models chosen using typical single-objective optimisation methods (e.g., ridge regression) in terms of prediction accuracy, while also demonstrating a good balance of other performance indicators.

From a public health standpoint, our research emphasizes the significance of multi-objective optimisation methods in model selection for infectious illness prediction. As global health security faces new challenges, such as the COVID-19 pandemic, effective outbreak prediction models are crucial for developing public health strategies. Our findings give models that can help public health decision-makers better plan resource allocation, estimate the potential risk of epidemic waves, and implement appropriate preventative and control strategies.

Theoretically, this study shows that a multi-objective optimisation method works well when dealing with multi-dimensional performance indicators in predictive models. It offers a novel perspective that takes into account more than just one accuracy parameter in the model selection process, integrating accuracy, generalization capabilities, and computing efficiency into a holistic model. This approach stretches the bounds of classic predictive model selection theory, resulting in a solution that is more appropriate for real-world challenges.

In practice, by taking into account multi-objective optimisation of forecasting models, this work delivers more refined and balanced forecasting tools for public health decision making. These technologies, especially during global health crises like the COVID-19 outbreak, can assist public health professionals in better predicting the development of outbreaks, developing more effective interventions, and optimizing resource allocation. The practical significance of this technique is not limited to current health concerns, but might be used to any field where several indicators must be merged for optimal decision-making.

Future research will be critical to enhancing the effectiveness and usability of predictive models for infectious illnesses. As new data emerges and prediction needs expand, new algorithms and approaches will be required to handle larger datasets and more diverse prediction jobs. Furthermore, investigating ways to more effectively incorporate expert knowledge and public health practice experience into a multi-objective optimisation model would result in more accurate and useful prediction models.

CHAPTER 5

A BLENDED ENSEMBLE MODEL FOR PREDICTION OF COVID-19 CASES

5.1 Introduction

Since its initial emergence at the end of 2019, COVID-19 has rapidly spread globally, exerting unprecedented impacts on human health, the global economy, and societal activities (Bar-Or *et al.*, 2022; Hernández-Giottonini *et al.*, 2023). As of 2023, the pandemic has led to hundreds of millions of confirmed cases and millions of deaths worldwide, underscoring the importance of rapidly and accurately predicting the transmission and outcomes of infectious diseases (Krause *et al.*, 2024). In this context, developing efficient disease prediction models is crucial not only for assisting governments and health organizations in implementing more effective intervention measures but also for optimizing resource allocation to mitigate the impacts of the pandemic on society and the economy (Rodrigues *et al.*, 2023).

To effectively control the COVID-19 pandemic, precision in predicting its spread and outcomes is crucial (Lu *et al.*, 2021; Yang *et al.*, 2020b). Traditional single models face limitations when dealing with large-scale, multivariable data. the study require extensive data preprocessing and feature engineering, which can lead to the curse of dimensionality. Moreover, their performance is subpar in handling nonlinear relationships and complex patterns, often resulting in overfitting or necessitating complex kernel functions. Due to their inability to capture multiple patterns, single models fail to fully leverage data information, potentially leading to poor performance (Shen and Li, 2024; Cheuque *et al.*, 2022; Sangphukieo *et al.*, 2020). Ensemble learning methods, such as Bagging, Boosting, Voting, and Stacking, have been introduced to address these limitations by combining multiple models to enhance predictive performance (Sun *et al.*, 2024; Hasrod *et al.*, 2024). However, each ensemble method has its own set of advantages and disadvantages. Bagging,

also known as Bootstrap Aggregating, aims to reduce variance by creating multiple independent models and then combining their predictions through averaging or majority voting. This approach is particularly effective for high-variance algorithms (Anastario *et al.*, 2023). However, the performance of bagging relies on the assumption of independence among the base models, which may not always hold true in practical scenarios (Lim *et al.*, 2023). Boosting, on the other hand, focuses on correcting the mistakes made by previous models by adjusting data weights iteratively, thereby reducing bias. While boosting can be effective in improving prediction accuracy, it is susceptible to overfitting when dealing with noisy data and typically requires higher computational resources (Zaghloul *et al.*, 2021). Voting and stacking are two additional ensemble techniques that aim to enhance prediction performance by combining outputs from multiple models. Voting involves averaging the predictions from different models, assuming equal importance for each model, which may not always be appropriate in real-world scenarios (Seal *et al.*, 2023). Stacking, on the other hand, utilizes cross-validation to train base learners and incorporates a meta-model to combine their predictions, resulting in a more sophisticated ensemble approach (Jiang *et al.*, 2023). However, the stacking process is computationally intensive and may be prone to overfitting and data leakage (Pant *et al.*, 2023).

In current research on predictive modeling for infectious diseases, many scholars tend to favor the Stacking model over the Blending model (Dervishi, 2024b; Shen *et al.*, 2024b; Gupta *et al.*, 2022; Mahajan *et al.*, 2022b; Ismail *et al.*, 2023). This preference often leads to a primary focus on improving model accuracy at the expense of neglecting aspects such as model generalizability and computational efficiency. Such bias may result in several crucial areas being overlooked or inadequately explored. Specifically:

(a) ***Model Practicality (Feasibility):***Stacking models typically involve complex structures, requiring extensive parameter tuning and computational resources, which may not be practical in resource-constrained settings or scenarios requiring rapid response. For instance, in the context of sudden public health emergencies, the ability to deploy and execute models quickly is crucial. Stacking models may suffer from overfitting when combining multiple models, leading to poor performance on new or unseen data, thus affecting the

model's generalization ability and limiting its effectiveness in different or novel environments.

(b) ***Model Interpretability:***By introducing multi-layered model structures to integrate predictions from different base models, Stacking may enhance accuracy but also increase the model's opacity, making it challenging to interpret the relationships between model inputs and outputs. This lack of interpretability is particularly limiting in the healthcare domain, where explainability is a critical component of medical decision support systems. High interpretability contributes to enhancing the model's credibility, enabling healthcare professionals and policymakers to understand and trust the model's predictions, thereby increasing their willingness to adopt these technologies to guide real-world decision-making.

Blending offers a potential solution to these issues by simplifying the training process, as it does not rely on complex cross-validation procedures to train base models. This simplification reduces computational burden, making the model more practical for rapid deployment and response. In contrast, Stacking models may be challenging to interpret due to their overly complex internal structures, such as multi-layered nested models and extensive parameter tuning. By simplifying the model structure, Blending avoids this excessive complexity, facilitating a clearer understanding and review of the model's prediction and decision-making processes.

The primary innovation of this study lies in the utilization of a multi-objective optimization algorithm to select base models of different algorithm types for constructing a blending ensemble learning model. This approach has demonstrated outstanding performance across multiple metrics, including accuracy, generalization ability, and computational efficiency. By enhancing the model's generalization ability and computational efficiency while ensuring accuracy, the model becomes more practical and interpretable (see Fig 5.1).

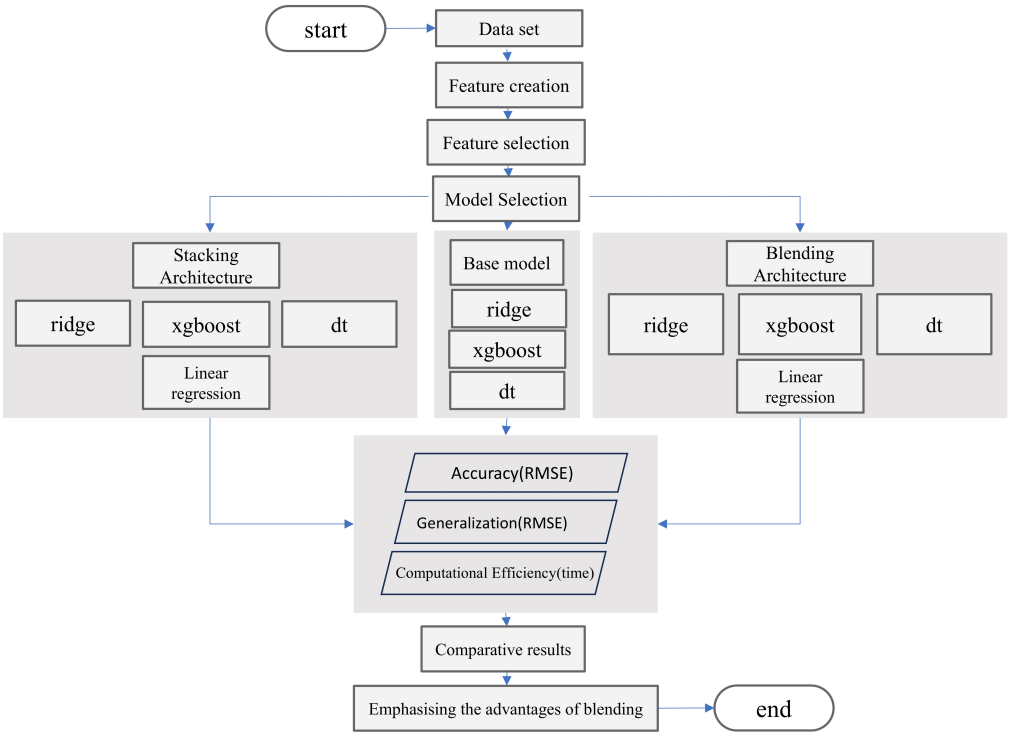


Figure 5.1 Methodology Flowchart of The Blended Ensemble Model

5.2 Data Processing

This study utilized the Mexican COVID-19 dataset provided by Our World in Data (Ding and Zhang, 2022). This dataset covers the period from April 1, 2020, to March 31, 2023, including various key indicators such as daily new cases (new_cases), total confirmed cases (total_cases), daily new deaths (new_deaths), total deaths (total_deaths), smoothed data, and rates calculated per million people. Such datasets are widely used in epidemiological research due to their completeness and accuracy (Larsen and Kraay, 2024). Given the presence of missing values in the original dataset, a median imputation method was used to fill these gaps. Due to the unique characteristics of infectious disease time-series data, outliers were retained to preserve the integrity and authenticity of the epidemiological trends.

Feature engineering is a crucial step in data processing, especially when predicting the spread of diseases. Well-designed features can significantly enhance a

model's predictive capabilities (Chen *et al.*, 2021). In this study, This study conducted a series of feature engineering processes on the Mexican COVID-19 dataset to capture the dynamic changes and trends of the disease's transmission. To understand the trend of COVID-19 case numbers over time, This study computed 7-day and 14-day moving averages. These features help the model understand recent trends in case increases or decreases.

Lag features are a common technique in time series analysis that help the model capture the autocorrelation of the sequence (Błażkiewicz, 2022). This study generated lagged features for new cases with 1-day and 7-day intervals. Differential features help the model capture the rate of change in the series, which is crucial for detecting accelerations or decelerations in disease transmission (Jiang *et al.*, 2021). the 7-day interval is closely related to the biological cycle of many infectious diseases, especially when considering the incubation period and infectious cycle of the virus. For example, the incubation period for COVID-19 is typically 2 to 14 days, with an average of about 5 days. Therefore, the 7-day time window can better capture the natural cycle of disease transmission, providing models with critical information about how the disease spreads in a shorter period. disease reporting cycles in many countries and regions are based on a one-week basis, which means that data is naturally clustered in a seven-day time frame. This reporting habit means that using the same time interval minimizes bias in data processing, making the features extracted from the raw data more realistic to the pattern of disease transmission. Additionally, This study created growth rate features, which provide another perspective on the relative changes in case numbers by displaying the percentage increase in new cases daily (Perramon-Malavez *et al.*, 2023).

Time point features: extracting the year, month, and day of the week from the date index. These features help the model identify potential seasonal patterns or weekly cyclical changes that might occur. Due to the potential generation of NA values from the calculation of sliding windows and lag features, rows containing NA values were removed after completing feature engineering to ensure the integrity of the data for model training and testing. Through the aforementioned feature engineering, This study have prepared a comprehensive and insightful set of features for the COVID-19

case prediction model. These features not only enhance the expressiveness of the data but also improve the accuracy and interpretability of the predictive model.

Feature selection is a crucial aspect of building effective predictive models, aimed at identifying the features that most significantly impact the target variable (in this case, the number of new cases, "new_cases") (Rodrigues *et al.*, 2020). This study utilizes the XGBoost regression model for feature selection. XGBoost (Extreme Gradient Boosting) is a popular Gradient Boosting Decision Tree (GBDT) algorithm. During the training process, XGBoost provides importance scores for each feature, which are based on their contributions to the model's predictive performance, such as reducing errors on the training data. XGBoost employs regularization techniques during the construction of decision trees to prevent overfitting, leading to the utilization of fewer but more informative features. By efficiently leveraging features, it helps identify those with the greatest predictive power for the target variable (Bolla *et al.*, 2023).

During the training, some features may not be selected for any splits in the trees, indicating that these features may not significantly contribute to the model's predictive ability. By observing these overlooked features, one can consider removing them from the model to simplify it (as illustrated in Table 5.1). This study excluded features with zero contribution values. Through this method, useful features were precisely selected, optimizing the model's performance (White *et al.*, 2024). These insights will guide future research to more accurately predict and manage the spread of COVID-19 and similar infectious diseases (see Table 5.1).

Table 5.1 Contribution of features assessed using the XGBoost feature importance assessment feature

Feature name	Contribution value
new_cases_per_million	0.9995
new_deaths	6.1137
new_cases_lag1	5.2678
day_of_week	4.8319
new_cases_diff	4.6435
new_cases_smoothed	4.5880
new_deaths_smoothed	4.1125
new_cases_14d_avg	3.7031
new_cases_lag7	3.4105
total_cases	3.2471
month	2.6358
new_cases_growth_rate	2.5328
total_deaths	0.0
total_cases_per_million	0.0
new_cases_smoothed_per_million	0.0
total_deaths_per_million	0.0
new_deaths_per_million	0.0
new_deaths_smoothed_per_million	0.0
new_cases_log	0.0
new_cases_7d_avg	0.0
year	0.0

For the data splitting phase, the process is completed in two steps: 60% of the data is used for model training to ensure there is sufficient data for this purpose. The remaining 40% of the temporary dataset is further divided equally into a validation set and a test set. For time series data, ‘shuffle=False‘ is set to avoid random shuffling of the data, thereby maintaining the chronological order of the data (Choi *et al.*, 2023).

5.3 Identified Base Models

Empirical research was conducted to compare a set of machine learning models in predicting the COVID-19 dataset. These models have been proven effective in time series forecasting. The NSGA-II algorithm was utilized to evaluate the models based on prediction accuracy, generalizability, and computational efficiency. Models such as Ridge Regression, Decision Tree (DT), and XGBoost, selected through the multi-objective optimization method NSGA-II, demonstrated superior performance in terms of accuracy, generalizability, and computational efficiency. The results are illustrated in Figure 4.2.

Ridge regression, also known as Tikhonov regularization, is an improved method of least squares estimation. It addresses certain issues encountered in linear regression models, particularly in cases of multicollinearity among features. By adding a regularization term to the loss function, Ridge regression stabilizes the estimation process.

In standard linear regression, our goal is to minimize the Residual Sum of Squares (RSS), which is formulated as:

$$RSS = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \tag{5.1}$$

Where *yi* is the response variable, *xi* is the predictor variable (eigenvector), and β is the regression coefficient. In Ridge regression, in addition to minimising the RSS, an L2 paradigm penalty is added to the coefficients, as:

$$Ridge = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|^2 \tag{5.2}$$

where λ is the regularisation parameter $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ is the sum of the squares of the coefficients. Ridge regression reduces the absolute size of the

coefficients through an L2 regularization term, making the model more robust to small changes in input data. This shrinkage effect is particularly beneficial because it reduces the complexity of the model and enhances its generalizability. In datasets with a large number of features, standard linear regression models are prone to overfitting. Ridge regression addresses this by penalizing the coefficients, thereby limiting model complexity and reducing the risk of overfitting.

In cases of multicollinearity, where predictors are highly correlated, the estimates of regression coefficients in standard linear regression can be highly unstable or even undefined (coefficients may become infinitely large). Ridge regression improves the numerical stability of these estimates by incorporating an L2 regularization term, which always provides a unique solution. This regularization term improves the condition number of the problem, thereby enhancing the stability of the numerical solution process, especially in calculations involving matrix inversions.

Ridge regression is an effective method for dealing with multicollinearity, particularly suitable in scenarios where variables are highly correlated or when the number of variables exceeds the number of samples (Babatunde *et al.*, 2024).

Decision trees construct a tree structure by recursively dividing the dataset into progressively smaller subsets. At each split point, the optimal feature is selected, and the feature is split based on a threshold. Split criteria are typically based on the increase in purity, such as Information Gain (ID3 algorithm), Gain Ratio (C4.5 algorithm), or Gini Impurity (CART algorithm). In regression problems, decision trees specifically utilize the minimization of Mean Squared Error (MSE) to determine the optimal split.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{5.3}$$

Here,$\hat{y}$ is the average target value of the samples at the current node. In regression problems, decision trees generally predict continuous values as the mean of the target values of the samples at each leaf node. This approach is well-suited for

handling nonlinear data, as it does not rely on any specific assumptions about data distribution (Abraham *et al.*, 2022).

XGBoost (Extreme Gradient Boosting) is an optimized distributed gradient boosting library designed to be efficient, flexible, and portable. It implements machine learning algorithms within the Gradient Boosting model by combining multiple weak prediction models (usually decision trees) to construct a robust predictive model. The core of XGBoost involves sequentially adding models, with each new model correcting the residuals of the previous one. The optimization in XGBoost comes from its utilization of a technique known as second-order gradient boosting. This method not only utilizes the first-order derivatives (gradients) but also incorporates second-order derivatives (Hessian) to minimize the loss function.

$$L(t) = \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_{i,t} f_t(x_i) + \frac{1}{2} h_{i,t} f_t^2(x_i) + \Omega(f_t) \right] \tag{5.4}$$

In this context, gi and hi represent the first and second derivatives of the loss function, respectively, Ω is the regularization term, and ft is the model added at step t. XGBoost also includes many features designed to handle modern data science challenges, such as dealing with missing values, supporting categorical features, and regularization. Due to its strong generalization capability, XGBoost is applied in various fields, including bioinformatics, particle physics, and more (Xu *et al.*, 2023).

Linear regression is one of the most fundamental regression methods, used for predicting continuous variables. It establishes a predictive model by minimizing the sum of the squares of the differences between the actual outputs and the predicted outputs. Mathematically, the linear regression model can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \tag{5.5}$$

In this formulation, y is the dependent variable, $x1$,$x2$,...,xp are the independent variables, $\beta0$,$\beta1$,...,βp are the regression coefficients, and ϵ is the error term. The advantage of linear regression lies in its simplicity and strong interpretability, offering an intuitive understanding of the relationships between the response variable and the predictor variables (Wu *et al.*, 2020).

5.4 Basic Blended Model

This study presents a blended ensemble model that combines three distinct base models: Ridge Regression, Decision Tree Regressor, and XGBoost Regressor. Each of these models utilizes different underlying algorithms, enabling them to capture unique characteristics of the data from various perspectives. This diversity in model structure forms the foundation for constructing a robust and flexible blending ensemble learning model.

5.4.1 Model Composition and Training Process

5.4.1.1 Base Models:

(a) Ridge is trained using a preprocessing pipeline that includes RobustScaler for normalization and FastICA (Independent Component Analysis) for dimensionality reduction. These steps are designed to improve the model's stability and predictive accuracy, especially in the presence of outliers or multicollinearity in the dataset.

(b) The Decision Tree is employed with its default settings, designed to capture nonlinear patterns in the data. Its ability to model complex decision boundaries allows it to effectively complement the more linear approach of Ridge Regression.

(c) XGBoost with a configuration of 100 trees, is used to model intricate interactions within the data. XGBoost is known for its high performance,

thanks to gradient boosting and regularization, which help prevent overfitting and improve model generalization.

Each of these base models is trained independently on the training set, ensuring that the study learn different patterns and features from their unique perspectives. This independence allows for an effective integration of their strengths during the blending process, promoting diversity in predictions.

5.4.1.2 Meta-Model and Blending Process:

After training the base models, their predictions are generated on the validation set. These predictions are treated as a new feature set for training the meta-model, which is a critical aspect of the blending strategy. In this study, the meta-model is implemented using Linear Regression. The use of a linear regression model allows for a straightforward and interpretable combination of the base models' predictions, with the regression coefficients providing insight into the contribution of each base model to the final output.

The meta-model is trained on the prediction results from the base models, using them as input features to learn the optimal weights. This step is essential for the blending ensemble, as it enables the meta-model to leverage the individual strengths of each base learner while compensating for their weaknesses.

5.4.1.3 Validation and Testing:

The validation set is used not only for assessing the base models but also for training the meta-model. By ensuring that the meta-model is trained on a separate validation dataset (20% of the overall training data), this method effectively mitigates the risk of overfitting. This approach also prevents data leakage, as the models are evaluated on different sets of data before making predictions on the final test set. After

the meta-model is trained, it is applied to make predictions on the test set, generating the final_predictions that represent the outcome of the blending ensemble model.

5.4.2 Advantages of the Blending Ensemble Model

5.4.2.1 Combination of Diverse Models:

The proposed blending ensemble combines models with diverse inductive biases. Ridge Regression represents a linear model, while Decision Trees and XGBoost are nonlinear models with distinct strengths in capturing different types of data relationships. The inclusion of these diverse base learners enables the model to capture a wide range of patterns, features, and interactions within the data. This heterogeneity improves the overall robustness and generalization capability of the ensemble, minimizing the likelihood of overfitting to specific data points or noise in the dataset.

5.4.2.2 Meta-Model for Optimal Integration:

The use of Linear Regression as the meta-model in the blending ensemble provides several benefits. First, it offers a transparent and interpretable model for understanding how the predictions from each base model contribute to the final output. Although Linear Regression is a simpler model, its effectiveness lies in its ability to efficiently synthesize the outputs of the base models by learning their optimal combination weights. This not only enhances the overall predictive performance but also maintains simplicity and clarity in model interpretation, making it easier to understand the contribution of each base model.

5.4.2.3 Mitigation of Overfitting via Validation Strategy:

By dividing the dataset into training (60%), validation (20%), and test (20%) sets, the blending ensemble model ensures that each model is trained and validated on distinct subsets of the data. The use of the validation set to train the meta-model helps prevent overfitting to the training data, ensuring that the meta-model generalizes effectively to unseen data. This careful use of data partitions and prediction results from the validation set helps the ensemble model maintain a strong generalization ability when applied to real-world datasets.

5.4.2.4 Enhanced Predictive Accuracy:

The blending method significantly improves the predictive accuracy by optimizing the weights assigned to each base model's predictions. This allows the meta-model to focus on the strengths of the most effective models, while minimizing the influence of models that may underperform on certain aspects of the data. The combination of predictions from Ridge Regression, Decision Tree, and XGBoost creates an ensemble that is greater than the sum of its parts, delivering superior accuracy and robustness.

When the amount of data is limited, blending can effectively mitigate the risk of overfitting, as validation only needs to be conducted on the designated validation set. If the predictions from the selected base models are similar, blending's weighted average can effectively smooth the results, thereby enhancing performance. Consequently, when the base models are robust and stable, blending can leverage their strengths without the need for complex meta-learning. Additionally, the computational overhead associated with blending is relatively low, making it suitable for resource-constrained environments. It is generally easier to interpret and is well-suited for scenarios where understanding the model's decision-making process is essential. Figure5.2 illustrates the flowchart of the ensemble learning Blending strategy applied to COVID-19 prediction (see Fig 5.2).

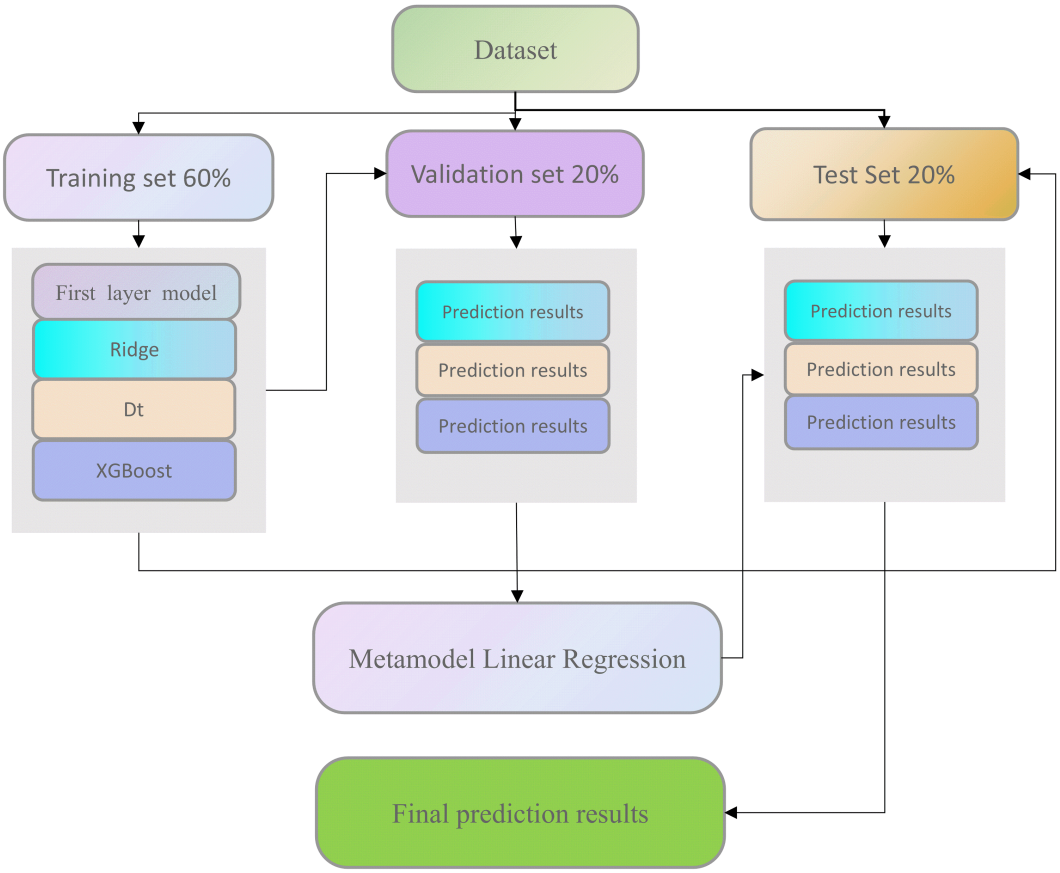


Figure 5.2 Flowchart of the Ensemble Learning Blending Strategy Applied to the COVID-19 Prediction Model

5.5 Result and Discussion

In this study, by comparing the performance of the base models with their blending ensemble method, This study have demonstrated the effectiveness of each model in terms of accuracy, generalization ability, and computational efficiency. The specific results are shown in Table 5.2, and Fig 5.3 - Fig 5.6, which displays the performance of the base models Ridge Regression, XGBoost, Decision Tree, and their blending ensemble method across different evaluation metrics.

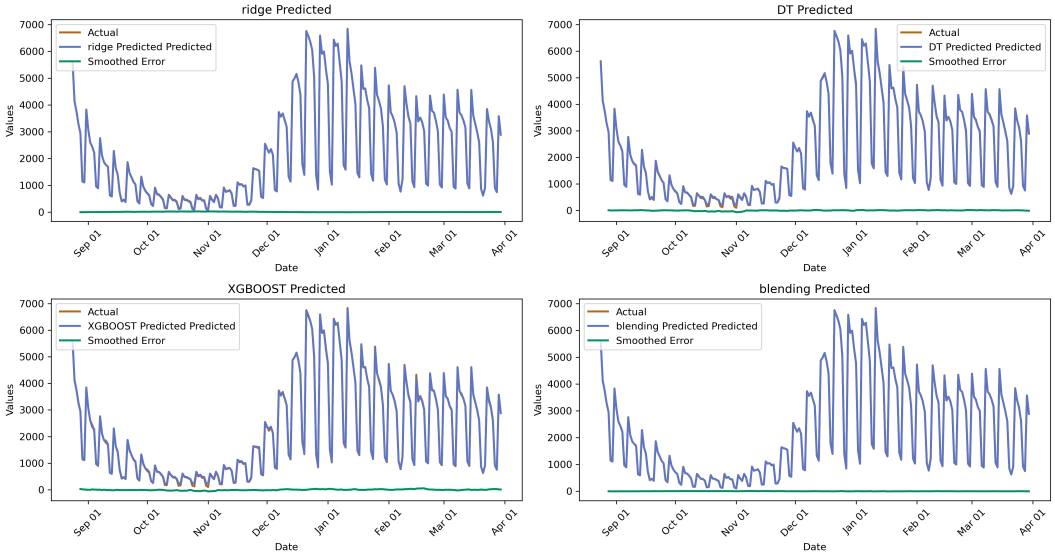


Figure 5.3 Model Accuracy (RMSE)

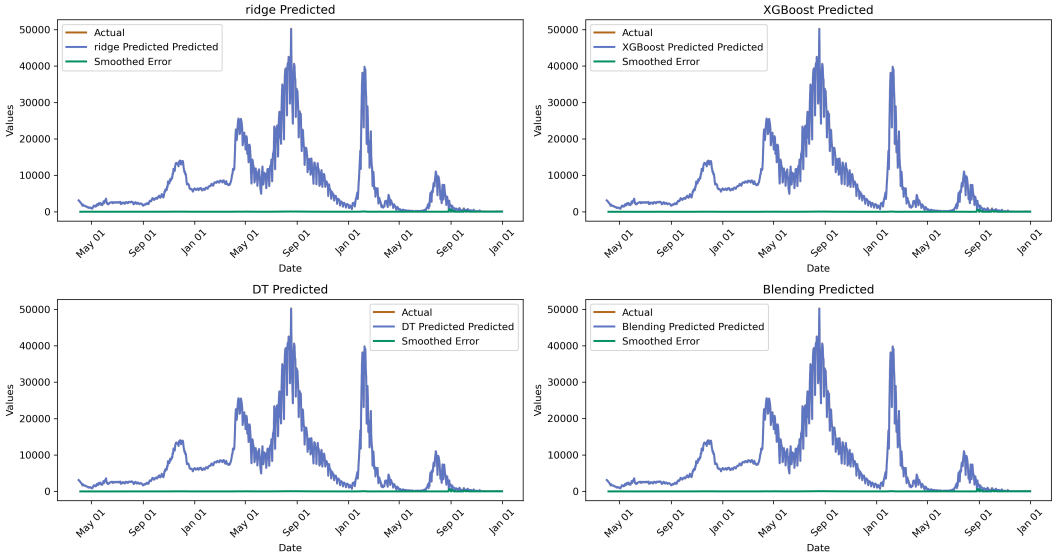


Figure 5.4 Model Generalization (RMSE Iran)

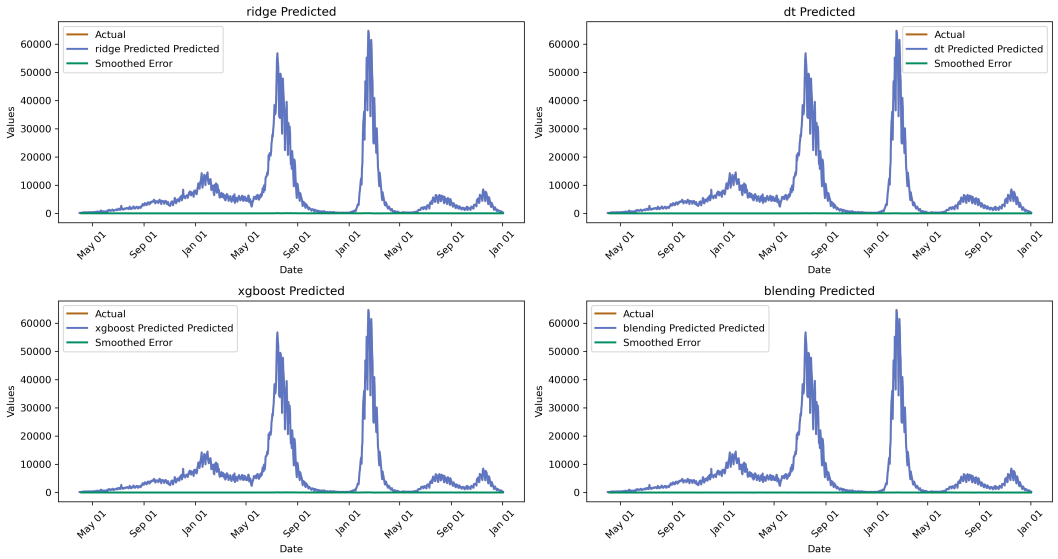


Figure 5.5 Model Generalization (RMSE Indonesia)

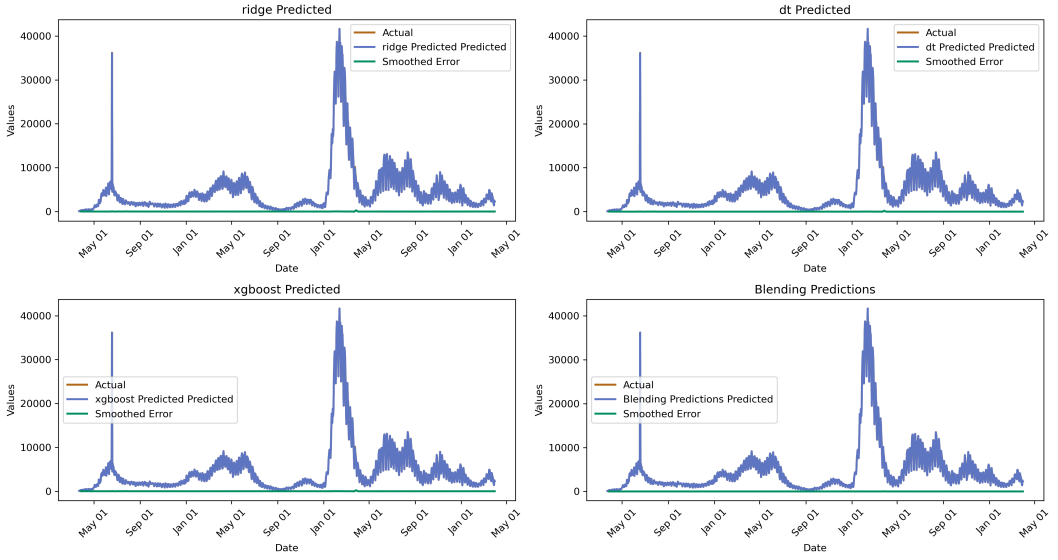


Figure 5.6 Model Generalization (RMSE Chile)

Table 5.2 Comparison of blending and its base model performance under different metrics.

Model Name	Accuracy (RMSE)	Generalization (RMSE,Iran)	Generalization (RMSE,Indonesia)	Generalization (RMSE,Chile)	Model Training Time
Ridge	11.52	114.85	16.25	23.24	0.03 seconds
XGBoost	33.07	13.06	13.51	7.98	0.36 seconds
Decision Tree	24.48	8.14	6.73	4.44	0.07 seconds
Blending	1.87	1.95	1.67	2.07	0.40 seconds

The Blending model significantly outperforms the base models across all evaluated performance metrics. It achieves an RMSE of 1.87 for accuracy, which is notably better than the best-performing base model, Ridge Regression, with an RMSE of 11.52. In terms of generalization, the Blending model also demonstrates exceptional performance, showing significantly lower Root Mean Square Error (RMSE) values on COVID-19 datasets from Iran, Indonesia, and Chile, with scores of 1.95, 1.67, and 2.07, respectively. These results underscore the blending method's ability to reduce errors and enhance stability by synthesizing predictions from different models.

In contrast, the base models, particularly Ridge and XGBoost, exhibit noticeable variability in generalization performance. The Ridge model performs poorly on the dataset from Iran, with an RMSE as high as 114.85, likely due to its sensitivity to outliers in the data. Although XGBoost shows more consistent performance in Iran and Indonesia, its accuracy still falls far short of the Blending model. The Blending ensemble method, by integrating the unique strengths of multiple predictive models, significantly enhances both the accuracy and the generalization ability of the model. This approach is particularly suitable for handling complex and highly variable epidemic data, such as that of COVID-19. Although the Blending model requires slightly more computational time than its base models, the substantial improvements in accuracy and generalization capabilities highlight its potential as a tool for epidemic prediction.

Compared to the study by Li *et al.* (2023d), this research presents a significantly innovative methodology. While both studies utilize the Blending Integrated Learning strategy, Li et al. concentrate on optimizing the combination of base models through genetic algorithms to enhance prediction accuracy and mitigate overfitting. However, the complex model structure and multi-level optimization process employed by Li et al. elevate computational costs, which may render their approach impractical in resource-limited application scenarios. In this study, a multi-objective optimization algorithm (NSGA-II) is employed to select the base model, emphasizing the balance between accuracy, generalization ability, and computational efficiency. This approach enhances the model's efficiency when handling large-scale data. It allows for the rapid deployment and application of the model in public health emergencies, particularly in complex outbreak prediction tasks such as COVID-19, demonstrating significant practicality and scalability.

This study compared the same base models (Ridge, XGBoost, Decision Tree) and a meta-model Linear Regression configured as a Stacking ensemble with the Blending ensemble model on the same datasets. Both stacking and blending methods aim to enhance prediction performance by combining the strengths of different base models, but the study exhibit significant differences in implementation and outcomes. Table 5.3 and Fig 5.7 - Fig 5.10 presents a performance comparison of these two methods in accuracy, generalization ability, and model training time.

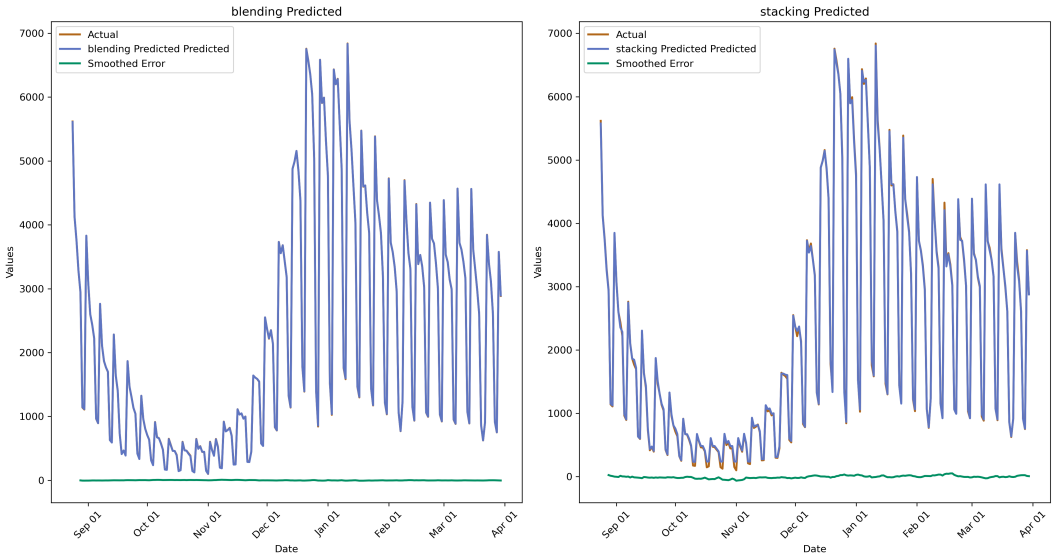


Figure 5.7 Blending And Stacking Accuracy (RMSE)

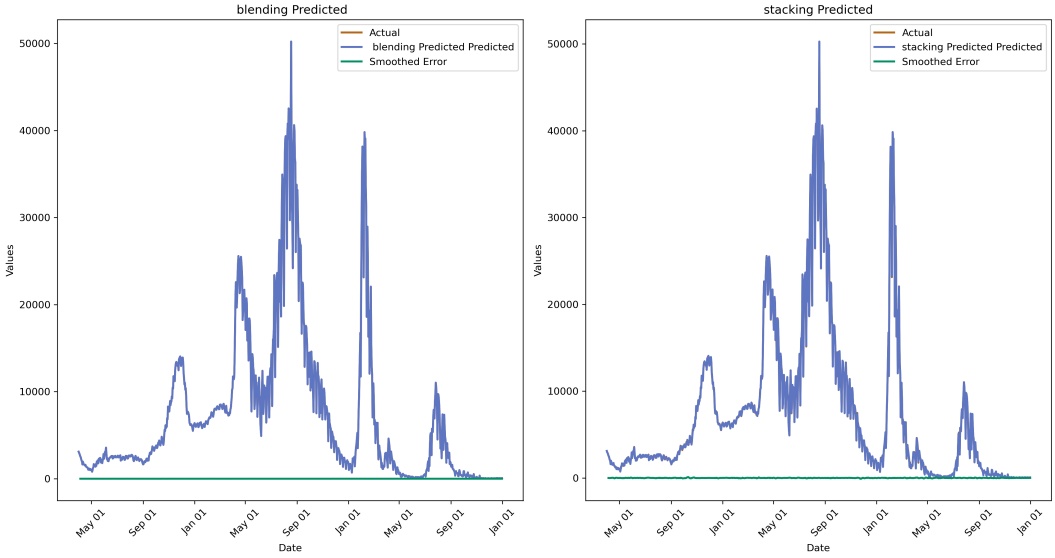


Figure 5.8 Blending And Stacking Generalization (RMSE Iran)

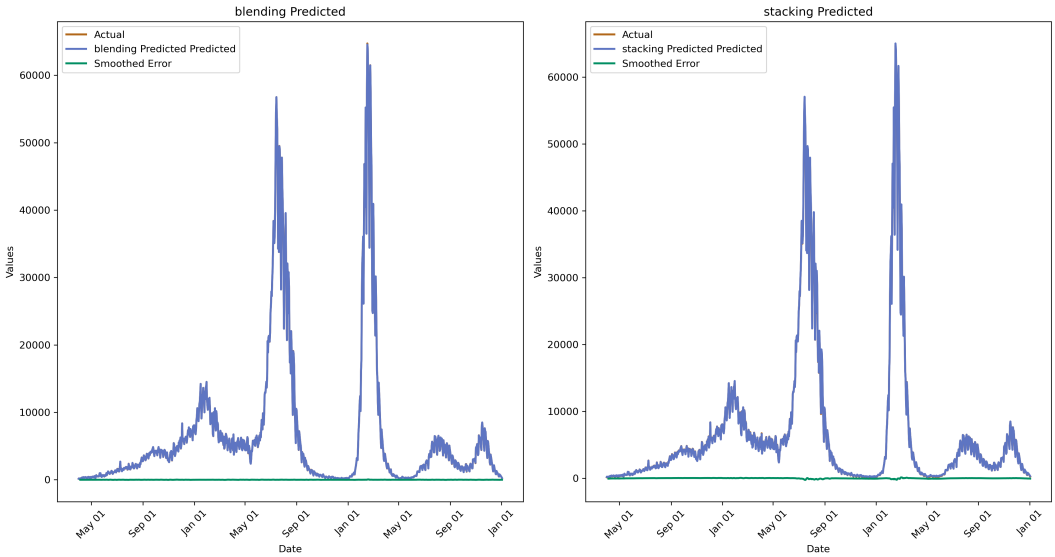


Figure 5.9 Blending And Stacking Generalization (RMSE Indonesia)

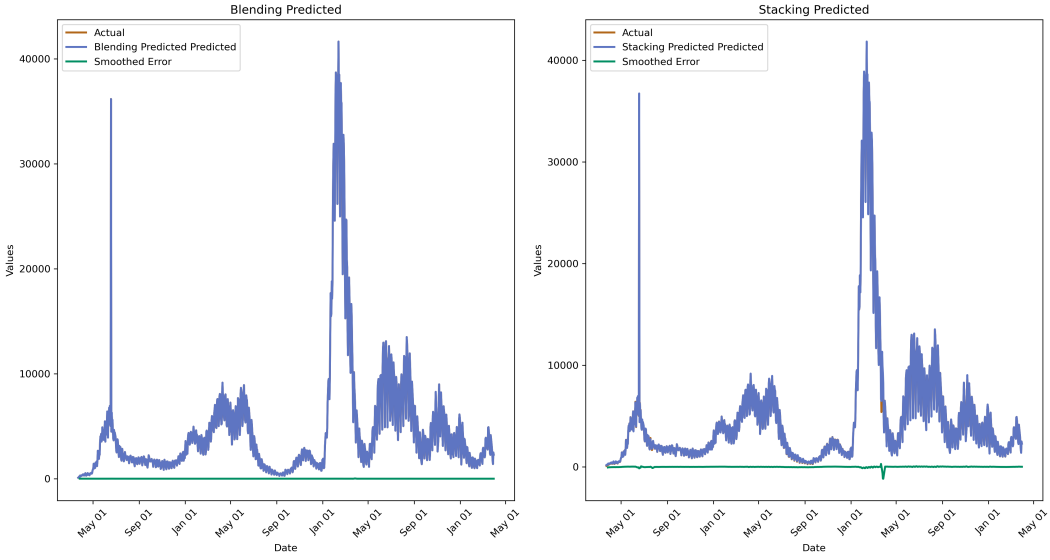


Figure 5.10 Blending And Stacking Generalization (RMSE Chile)

Table 5.3 Blending and Stacking models differ significantly in implementation and performance

Model Name	Accuracy (RMSE)	Generalization (RMSE,Iran)	Generalization (RMSE,Indonesia)	Generalization (RMSE,Chile)	Model Training Time
Stacking	73.81	66.51	72.34	126.39	1.17 seconds
Blending	1.87	1.95	1.67	2.07	0.40 seconds

From the table, it's evident that the Blending model significantly outperforms the Stacking model in terms of accuracy, with an RMSE of 1.87 compared to Stacking's 73.81. This indicates that the Blending model can predict the spread of COVID-19 more precisely, with much lower prediction error.

In terms of generalization ability, the Blending model also performs exceptionally well, achieving a maximum generalization error of 1.95 in Iran and 1.67 in Indonesia. This is much lower than the Stacking model's errors, which are very high, particularly in Chile, where it reaches 126.39. This disparity likely stems from the different ways the two models handle data. Blending optimizes performance by linearly combining predictions from base models, whereas Stacking may introduce excessive model complexity due to its deeper hierarchical structure, potentially leading to overfitting.

Regarding model training time, the Blending model is more efficient, requiring only 0.40 seconds, compared to the Stacking model, which needs 1.17 seconds. This difference reflects the optimization in the blending method during model training. By reducing the complexity of model layers and parameter tuning, blending can quickly integrate and adjust predictions from various base models.

A detailed comparison of the Blending and Stacking ensemble methods was conducted to evaluate their performance in handling COVID-19 data prediction tasks. The results show that the Blending architecture outperforms the Stacking method on

multiple key performance indicators, revealing differences in the inherent mechanisms and applicability of each method (Inyang *et al.*, 2023).

In the field of machine learning, blending and stacking are two widely used techniques that aim to enhance overall prediction accuracy by combining the predictive powers of multiple models. While both methods share similarities in application, the study significantly differ in implementation details and data utilization approaches, which may lead to variations in their performance outcomes (Hasan *et al.*, 2023).

These findings underscore the importance of selecting the appropriate ensemble technique based on the specific requirements and characteristics of the data being analyzed. While stacking may introduce complexity that leads to overfitting, especially in scenarios involving highly variant data, blending tends to provide a simpler yet effective solution for integrating diverse model predictions, making it a more favorable choice for the given COVID-19 prediction task. The study's insights into the strengths and limitations of each method can guide future research and practical implementations in epidemic modeling and other areas that require robust predictive analytics.

Blending and stacking are both methods used in ensemble learning to enhance the performance of machine learning tasks by integrating multiple different models. the study are commonly used to reduce the bias or variance of the model, thus enhancing its predictive capabilities.

Stacking involves dividing data into test and training sets and constructing multi-layer models. The first layer consists of several base models, each trained independently to generate predictions; the second layer, also known as the meta-model, then uses these predictions as inputs. A distinctive feature of Stacking is its utilization of cross-validation to train the base models, ensuring that the training data is fully utilized, thereby enhancing the model's predictive capacity. However, this method can also lead to potential issues such as data leakage and overfitting, which can result in overly optimistic prediction outcomes. Additionally, it requires more training time, which can reduce computational efficiency (Kalule *et al.*, 2023) (see Fig 5.11).

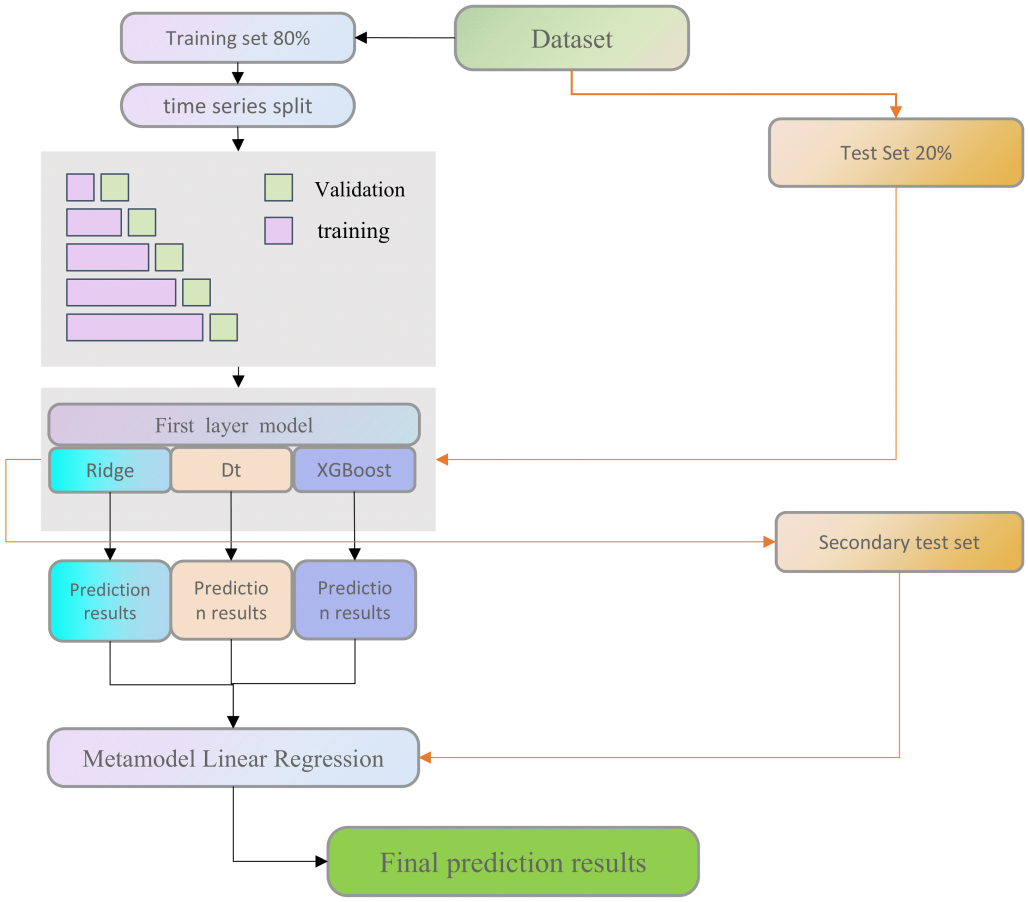


Figure 5.11 Stacking Ensemble Model Architecture

Blending is similar to stacking, but it simplifies the data partitioning process. It typically divides the dataset into three parts: one part is used to train the base models, another part (usually a separate validation set) is used to generate the training data for the meta-model, and the remaining part serves as the test set. This method reduces computational complexity by avoiding the multiple model training processes associated with cross-validation.

This study highlights the effectiveness of utilizing the Blending method as an ensemble technique. This method not only enhances prediction accuracy but also reduces model complexity by streamlining the ensemble process. Theoretically, this confirms that it is possible to maintain high-quality predictions while reducing model complexity, providing new directions for the development of future ensemble

learning methods. In ensemble learning, information leakage and overfitting are two common issues (Cheng *et al.*, 2022). The blending approach in this study effectively addresses these issues by integrating independently trained models from different algorithmic types and simplifying the meta-model process. It offers valuable empirical evidence on how to enhance generalizability while ensuring model independence. By comparing various ensemble strategies, this research illustrates how to evaluate performance across multiple objectives, including accuracy, generalizability, and computational efficiency. This multi-objective evaluation method theoretically enriches the application model of ensemble learning, particularly in resource-constrained environments.

Due to its high accuracy and robust generalizability, the Blending model is particularly well-suited for predicting trends in disease transmission. This can assist public health officials and policymakers in more accurately forecasting epidemic developments, optimizing resource allocation, and formulating more effective disease control strategies. In situations requiring rapid response, such as during the COVID-19 pandemic, the computational efficiency of the Blending model provides the potential for real-time data processing and prediction. This is crucial for early warning and immediate decision-making.

5.5.1 Potential Limitations and Challenges

Despite achieving certain theoretical and practical outcomes, this study still faces several limitations and challenges.

(a) ***Dependency on Data Quality:*** The performance of the blending model is significantly influenced by the quality and completeness of the input data. In cases where data collection is incomplete or of poor quality, the predictive performance of the model may be compromised.

(b) ***Generalization across Different Diseases:*** While the Blending model demonstrated good generalizability in this study, it remains to be seen whether this performance can be generalized across different types of

diseases and epidemiological data. The transmission mechanisms and societal impact factors of different diseases may affect the model's generalization effectiveness.

(c) ***Model Interpretability:*** Although blending improves prediction accuracy, the interpretability of ensemble models is generally lower compared to single models. Enhancing the interpretability of models to make them more transparent and trustworthy in practical applications is another important direction for future research.

5.6 Chapter Summary

In this chapter, This study employed multi-objective optimization algorithms to choose base models from various algorithm types and investigated the use of the Blending ensemble learning model for predicting COVID-19 transmission trends. This study compared this approach with the commonly used Stacking ensemble learning method in the realm of infectious disease prediction. The results clearly demonstrate that the blending model outperforms the stacking model across multiple key performance indicators, including accuracy, generalizability, and computational efficiency. Particularly in terms of generalizability, the blending model demonstrated exceptional performance with the datasets from Iran and Indonesia. Its RMSE was significantly lower than that of the Stacking model, indicating its strong adaptability and stability across different data environments.

The blending method, which integrates the independent predictions of multiple predictive models, effectively reduces the bias and variance that may occur in a single model, thereby enhancing the accuracy and reliability of predictions. Additionally, this method has shown high computational efficiency in both model training and prediction processes, making it a powerful tool for real-time disease monitoring and prediction.

Future work, considering the success of the Blending model with COVID-19 data, could involve applying this model to other infectious diseases, such as influenza or Ebola, to potentially achieve similarly favorable predictive outcomes. This

would require adjusting the model parameters and structures according to the specific transmission mechanisms and influencing factors of each disease. Further research on the application of the Blending model across various geographical and socio-economic contexts is necessary to assess its generalizability and applicability on a global scale.

CHAPTER 6

ENHANCED BLENDED ENSEMBLE MODEL WITH TRANSFER LEARNING AND INCREMENTAL LEARNING

6.1 Introduction

The outbreak of emerging infectious diseases poses a significant challenge to global public health (Olum *et al.*, 2024; Ji *et al.*, 2023; Nair *et al.*, 2023). In recent years, with the acceleration of globalization and changes in the ecological environment, the speed and scope of infectious disease transmission have significantly increased (Li *et al.*, 2024b). This not only places immense pressure on healthcare systems but also threatens global economic and social stability (El Taha *et al.*, 2022). Emerging infectious diseases, such as novel coronavirus disease 2019 (COVID-19) and monkeypox, have triggered widespread health crises in a short period, and their uncertainty and suddenness render traditional response strategies ineffective (Chaudhary *et al.*, 2023).

Early prediction plays a crucial role in public health responses (Alizargar *et al.*, 2024). Timely and accurate epidemic forecasting assists decision-makers in proactive resource allocation and the implementation of effective control measures, thereby minimizing the societal and economic impact of outbreaks. Specifically, early prediction supports the formulation of vaccination strategies, allocation of medical resources, and guidance for public health interventions (Wang *et al.*, 2020b). This forecasting entails quantitative analysis of infectious disease transmission trends and necessitates consideration of various complex factors, including pathogen biological characteristics, modes of transmission, and population mobility. Therefore, the development of efficient and reliable forecasting tools, particularly those capable of providing accurate predictions in the early stages, is a significant task in the field of public health (Soliman *et al.*, 2019).

Current predictions of early infectious diseases have some shortcomings and gaps (Popescu and Myers, 2021; Wang *et al.*, 2020a). A major issue is the reliance on single-patch models in machine learning methods, which overlook the impact of human mobility on disease transmission. Additionally, the lack of real-time data utilization and the use of small datasets hinder the accuracy of infectious disease diagnostic models (Alqaissi *et al.*, 2022). While internet and search engine data show significant potential in early prediction, the study also highlight deficiencies in current methods in terms of automatic keyword filtering and real-time updates, affecting their alert capabilities (Wang *et al.*, 2023b). Although early SIR models are useful, the study may underestimate the required resources due to limited data availability and constantly changing epidemic characteristics, emphasizing the necessity of frequent model input modifications (Liu *et al.*, 2023b). The lack of integration of epidemiological knowledge, real-time forecasting, and dynamic model updates in early prediction model construction is evident. This is crucial for improving the accuracy and effectiveness of early infectious disease predictions.

The primary objective of this study is to increase the accuracy and practicality of infectious disease prediction by leveraging existing blending models, employing transfer learning and incremental learning techniques, and incorporating the biological feature Rt. Specifically, our aim is to address the limitations of current models in terms of data scarcity and prediction generalizability. By utilizing feature transfer from COVID-19 data and dynamically updating real-time data, This study seek to improve the performance of the model in predicting emerging infectious diseases such as monkeypox.

This study makes several contributions. First, with increasing globalization and changes in the ecological environment, the frequency and scope of emerging infectious diseases are increasing, posing a significant challenge to global public health (Ji *et al.*, 2023; Lou *et al.*, 2022; Olum *et al.*, 2024). Accurate early forecasting can provide crucial decision support for public health authorities, aid in the effective allocation of resources, formulate emergency measures, and reduce the spread and impact of epidemics (Canino *et al.*, 2022; Estiri *et al.*, 2021b; Ruan *et al.*, 2023). Second, this study explores the potential of multimodel fusion, which integrates the strengths of multiple models to provide more stable and reliable forecasting results (Gong *et al.*,

2022; Wang *et al.*, 2023a). By introducing the biological feature Rt, the models can capture not only the trends of disease transmission but also the effectiveness of epidemic control measures, thereby providing a basis for the development of more scientifically informed public health strategies. Additionally, the application of transfer learning and incremental learning, especially in situations of sparse and continuously updated data, has unique advantages in addressing emerging infectious diseases (Didier *et al.*, 2024; Lenatti *et al.*, 2023). By leveraging knowledge from previous epidemics and real-time data updates, these methods can significantly improve the accuracy and adaptability of the models. This is crucial for mounting rapid responses in the early stages of an epidemic, aiding in the timely implementation of control measures to mitigate the spread of the disease (see Fig 6.1).

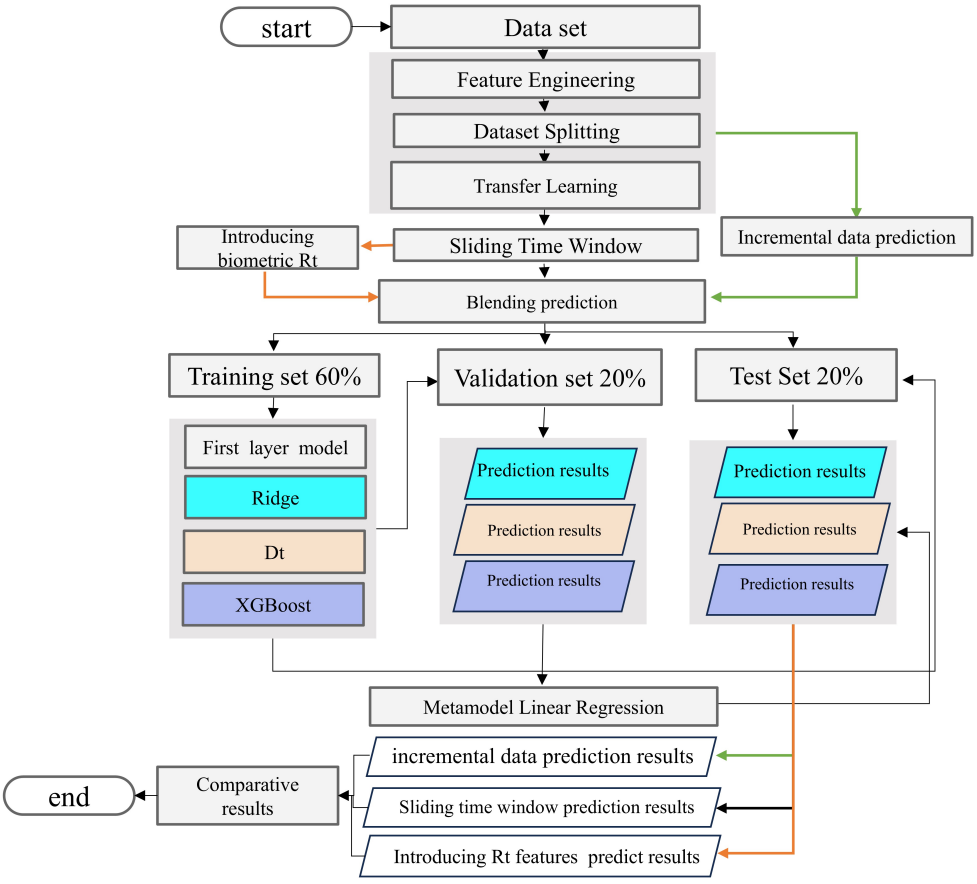


Figure 6.1 Methodology Flowchart Of The Enhanced Blended Ensemble Model

6.2 Data collection and processing

This study utilized the US Monkeypox dataset provided by our World in Data. This dataset spans from May 10, 2022, to February 27, 2024, encompassing key metrics such as daily new cases, 7-day averages, and total cases. Given the limited early infectious disease data and feature information, This study enhanced the data usability by generating new features.

6.2.1 Feature Engineering

Introducing additional dimensions of features can enhance the model's understanding of disease transmission patterns and improve its predictive ability (Liu and Su, 2024; Zhang *et al.*, 2024b). Given the time-dependent nature of infectious disease spread, incorporating lag features and smoothing features can capture the dynamic changes in time series data. The specific features are as follows: The extraction of Year, Month, and Day of the Week from dates is typically performed via date-time libraries in programming languages such as Python's datetime module. This process aids in capturing the effects of seasonal and periodic variations in models (Du *et al.*, 2022; Qiu *et al.*, 2020). New_Cases_14d_Avg is calculated as the average number of new cases in the past 14 days, achieved through the application of rolling window functions. This feature helps smooth daily fluctuations, offering a clearer view of disease transmission trends. New_Cases_Cubic represents the cube of the number of new cases, potentially assisting models in identifying nonlinear patterns in case growth. New_Cases_Diff indicates the daily increment in new cases compared with the previous day, reflecting the immediate changes in disease transmission speed. The New_Cases_Growth_Rate is calculated as the ratio of new cases on a given day to those on the previous day, providing insight into the growth rate of new cases. New_Cases_Lag1 and New_Cases_Lag7 denote the number of new cases one day and seven days prior, respectively. These lag features help models understand the short-term and medium-term dynamics of disease transmission. New_Cases_Per_Million represents the number of new cases per million people, standardizing case numbers for comparison across different regions. New_Cases_Smoothed refers to new cases

processed by moving averages or other smoothing techniques to reduce the impact of daily fluctuations (Borré *et al.*, 2023). Week_of_Year indicates the week number within a year, aiding models in capturing the periodic effects of weeks.

6.3 Dataset Splitting

The entire dataset was divided into a training set and a validation set, with the training set comprising 60% and the validation set comprising 40%. In this step, This study set shuffle=False to preserve the temporal order of the data, as time series data exhibit temporal dependencies, and shuffling the order could disrupt the model's ability to learn these temporal features accurately. The validation set was subsequently further divided into a validation set and a test set, each accounting for 20%. Similarly, shuffle=False was employed to maintain the temporal sequence, ensuring that the model has an adequate amount of data for training.

6.4 Transfer learning

This study employ transfer learning to address the challenge of limited data in the early stages of the monkeypox epidemic. Transfer learning leverages existing knowledge to address new but related problems, significantly enhancing model performance on novel tasks (Lee *et al.*, 2024). Specifically, This study utilize features from the COVID-19 dataset for feature transfer and adapt the model to suit the monkeypox data. Feature transfer, an essential method in transfer learning, aligns features from the source domain (COVID-19) with the target domain (monkeypox), enabling models trained in the source domain to be applied to the target domain. The following outlines the specific steps involved in our feature transfer process:

6.4.1 Feature Alignment

In the COVID-19 dataset, features such as 'new_cases_per_million', 'new_deaths', 'new_cases_lag1', 'day_of_week', 'new_cases_diff', 'new_cases_smoothed', 'new_deaths_smoothed', 'new_cases_14d_avg', 'new_cases_lag7', 'total_cases', 'month', and 'new_cases_growth_rate' have been identified as valuable for training the model as the study contribute to the target variable. Similarly, in the monkeypox dataset, features such as 'new_cases_cubic', 'new_cases_diff: 7-Day Average', 'new_cases_lag1', 'new_cases_lag7', 'day_of_week', 'total_cases', 'new_cases_14d_avg', 'new_cases_growth_rate', 'week_of_year', 'month', and 'year' are considered valuable for predicting the target variable. Retaining the relevant features from each dataset and aligning them for analysis is crucial. Notably, the COVID-19 dataset includes unique features such as 'new_cases_per_million', 'new_deaths', 'new_cases_smoothed', and 'new_deaths_smoothed'. To ensure full alignment of the features between the two datasets, it is necessary to create missing features in each dataset. For the COVID-19 dataset: new_cases_cubic can be derived by calculating the cube of new_cases. week_of_year can be derived from the date, and Year can also be derived from the date. For the monkeypox dataset: new_cases_per_million can be calculated using a population of 335.9 million people. Regarding new_deaths, a simple assumption is made that deaths constitute a fixed percentage of new cases. In this study, a 2% mortality rate is assumed, meaning that for every 100 new cases, there are expected to be 2 deaths. This ratio is an estimate, as early stages of novel infectious diseases often lack relevant information, typically on the basis of data from similar outbreaks. new_cases_smoothed can be created by computing the moving average of new_cases. Similarly, new_deaths_smoothed is calculated by applying a moving average to new_deaths to smooth daily fluctuations and provide a more stable trend of death cases. In this study, a 7-day moving average is employed. This methodology is commonly used for time series data to help reveal long-term trends and reduce the impact of short-term fluctuations. Figure 6.2 depicts the feature after completing feature alignment. Due to the limited availability of early monkeypox data, the feature-aligned datasets for COVID-19 and monkeypox were merged into a new dataset.

6.4.2 Feature Transfer

The pre-trained blending model, originally developed using the COVID-19 dataset, is reloaded and retrained on a newly merged dataset. The retraining process employs the same feature set utilized during the initial training of the COVID-19 model. By integrating and aligning features from distinct data sources, this approach enables the model to leverage the knowledge acquired from the COVID-19 dataset and apply it to a new dataset that includes monkeypox-related features. This method facilitates knowledge transfer and adaptation, allowing the model to generalize effectively across the expanded feature space derived from both the COVID-19 and monkeypox datasets (see Fig 6.2).

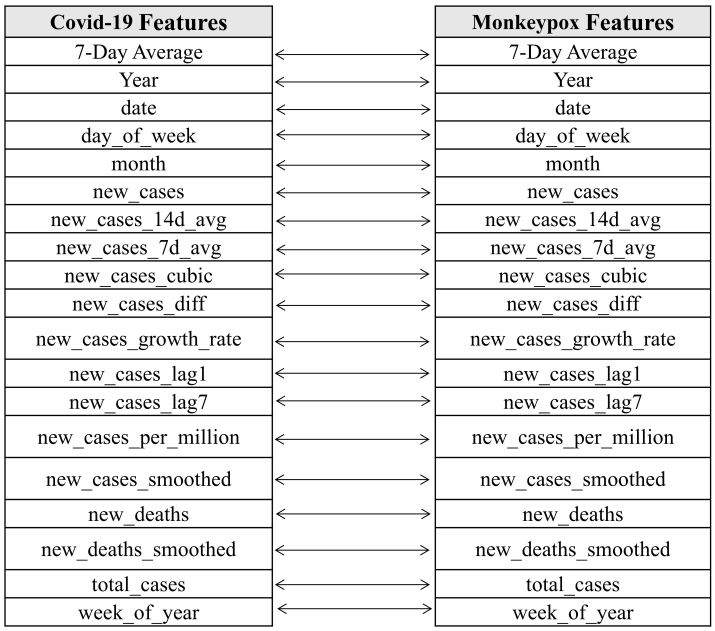


Figure 6.2 Alignment of COVID-19 and Monkeypox dataset features

6.5 Incremental learning

Incremental learning is a technique that allows models to be gradually updated as data continue to change and increase (Shyaa *et al.*, 2023). It enables models to incorporate new data without retraining the entire model, thus maintaining real-time

adaptability. The ridge regression, decision tree, XGBoost, and linear regression models used in this study do not support true incremental learning. To enable models to dynamically adapt to continuously changing new data in real time, This study have devised a set of dynamic update mechanisms to approximate the effects of incremental learning.

6.5.1 Dynamic Updating Mechanism with Sliding Time Windows

The sliding time window is a commonly used dynamic updating mechanism that involves sliding data within a fixed time window to progressively update a model (Cui *et al.*, 2021). Specifically, This study define a fixed window of 30 days in length, where the model is trained using only the data within the window for updating. As new data arrives, the window slides forward to encompass the latest data while discarding the earliest data. The absolute error time series distribution plot of the blending model on the COVID-19 dataset (Figure 6.3) clearly shows that the blending model can provide relatively accurate results for the next 7 days. The x-axis spans from September 2022 to April 2023, whereas the y-axis represents the predicted absolute errors ranging from 0-4. The label "7 days" in a red box is positioned in the bottom left corner of the chart, highlighting a period of consistently low absolute errors. Therefore, This study slide the time window forward by 7 days each time to achieve a matching performance effect on the blending model, as shown in Figure 6.4.

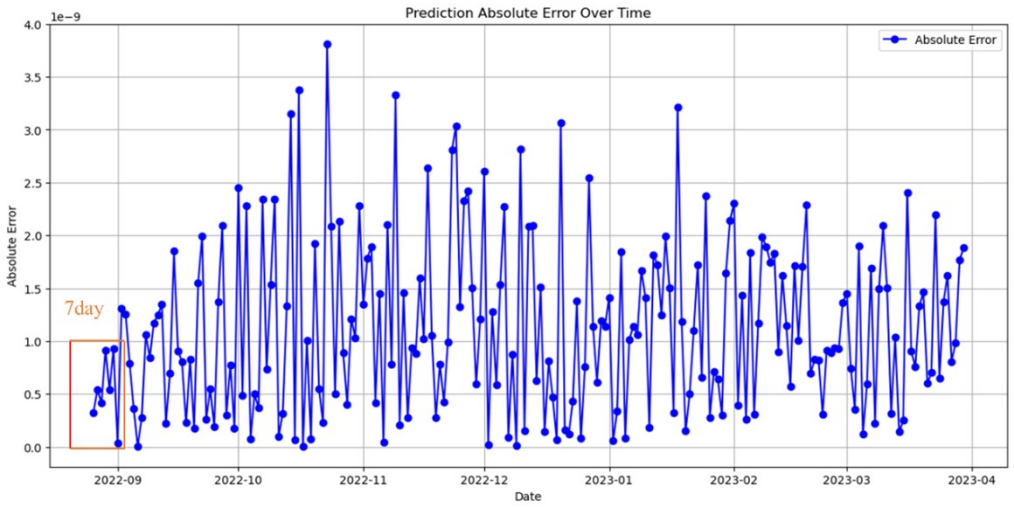


Figure 6.3 The distribution of the absolute error time series predicted by the blending model.

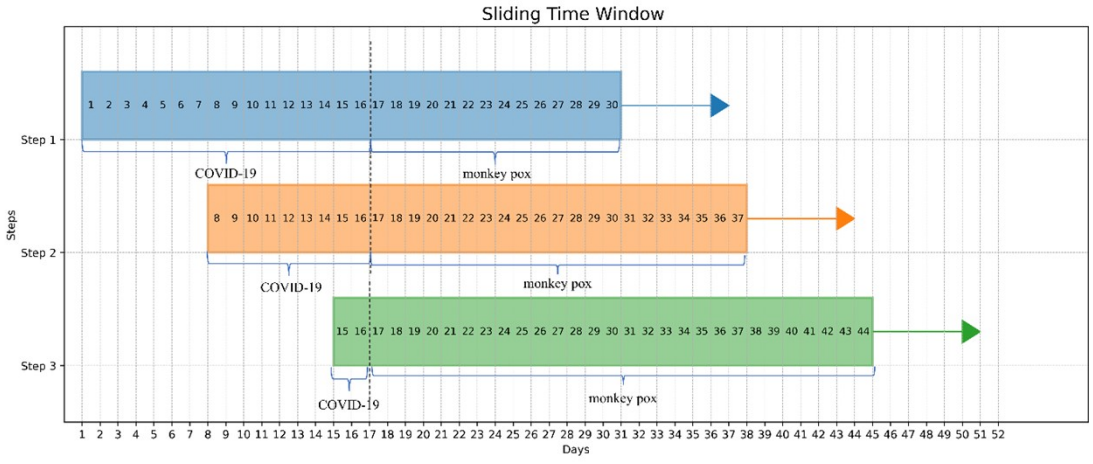


Figure 6.4 Illustration of a sliding time window.

This study hypothesized that by utilizing 14 days of monkeypox data in the very early stages of an outbreak, followed by the incorporation of 16 days of COVID-

19 data to complete a 30-day time window, with each additional week of monkeypox data replacing the earliest week of COVID-19 data, a sliding time window approach was implemented to enable real-time prediction and dynamic model updating (Soni *et al.*, 2022).

6.6 The biological feature Rt is introduced.

In epidemiology, the effective reproduction number, Rt, is a crucial metric that signifies the average number of individuals to whom an infected person can transmit the infection during a specific period under prevailing conditions (Champredon *et al.*, 2024). The dynamic fluctuations in Rt serve as indicators of the speed of transmission and the efficacy of control measures during an outbreak. Therefore, incorporating Rt as a biological characteristic is highly valuable for predicting the emergence of new infectious diseases such as monkeypox:

(a) When Rt is greater than 1, the epidemic spreads, with each infected individual on average transmitting to more than one person.

(b) When Rt equals 1, the epidemic is in a stable state, with the number of infections no longer increasing.

(c) When Rt is less than 1, it signifies a decline in the epidemic, with each infected individual on average transmitting to fewer than one person.

6.6.1 Methodology for Calculating Rt

The calculation of the effective reproduction number (Rt) is highly important in the formulation and assessment of public health policies such as social distancing measures and vaccination strategies, as it provides real-time insights into the impact of these interventions (Madewell *et al.*, 2023). Various commonly employed methods for estimating Rt include the following:

(a) ***Time series methods:*** This method uses time series data of reported case numbers to infer Rt by estimating the growth rate of infections. The growth rate is estimated by the slope of the logarithmically transformed case numbers, which is then combined with the virus's serial interval to calculate Rt.

(b) ***The Bayesian method:*** a sophisticated statistical approach, integrates uncertainty and prior knowledge (such as historical data or other epidemiological characteristics). Typically, employing Markov chain Monte Carlo (MCMC) techniques, this method estimates the probability distribution of Rt, offering confidence intervals and uncertainty assessments regarding Rt estimates.

(c) ***Real time estimation tool:***There are several readily available tools and software packages, such as EpiEstim and EpiFilter, that can be utilized for estimating Rt. These tools typically incorporate the aforementioned methods and enable researchers to input real-time case data to obtain estimates of Rt.

This study utilized time series methods to calculate Rt to understand the transmission characteristics of monkeypox and other related diseases. It is plausible to consider using a generation interval similar to that of smallpox or cowpox, which typically falls between 12 and 14 days. For the purpose of this analysis, This study may opt for an average value of 13 days as the generation interval to estimate Rt. The fundamental steps for calculating Rt are as follows:

(a) The daily growth rate, denoted as r, is calculated by comparing the number of cases on two consecutive days. If Ct represents the number of cases on day t, the daily growth rate r can be calculated via the following formula:

The daily growth rate, denoted as r,is calculated; this estimation is derived by comparing the number of cases over two consecutive days. If This study denote the number of cases on day t as Ct, then the daily growth rate r can be calculated via the following formula:

$$r_t = \ln\left(\frac{C_t}{C_{t-1}}\right) \tag{6.1}$$

Here, 'ln' denotes the natural logarithm.

(a) To calculate Rt, utilize the serial interval T in the context of secondary transmission;

Once This study have the daily growth rate, This study can calculate Rt via the estimated generation interval. The generation interval is the average time it takes for an individual to infect the next individual. The formula for Rt is as follows:

$$R_t = e^{r \times T} \tag{6.2}$$

Here, *e* represents the base of the natural logarithm, indicating that, in the absence of interventions, an infected individual will on average infect *Rt* other individuals during their infectious period.

6.7 Results and Discussion

In this section, this study present the experimental results when the blending model and transfer learning incremental learning technique are used on the monkeypox dataset. Our focus lies in evaluating the predictive performance of the models under different features and methodologies, encompassing the efficacy of time window-based feature transfer incremental learning and the performance enhancement upon the introduction of the biological feature Rt.

6.7.1 Preliminary Predictions of the Blending model

Initially, this study evaluated the progressive predictive capability of the blending model on the monkeypox dataset. With the ongoing accumulation of early monkeypox data, there was a corresponding increase in the volume of data within

the training set, enabling the observation of fluctuations in model performance. This methodology facilitated our comprehension of how the blending model performs under the circumstance of continually refreshing early data.

As shown in Table , the data were trained, validated, and tested within four time periods of 30, 37, 44, and 51 days as the study increased. The corresponding numbers of days for the training, validation, and testing sets, along with their predictive performance metrics (RMSE and MAE), are presented below (see Fig 6.5 and Table 6.1).

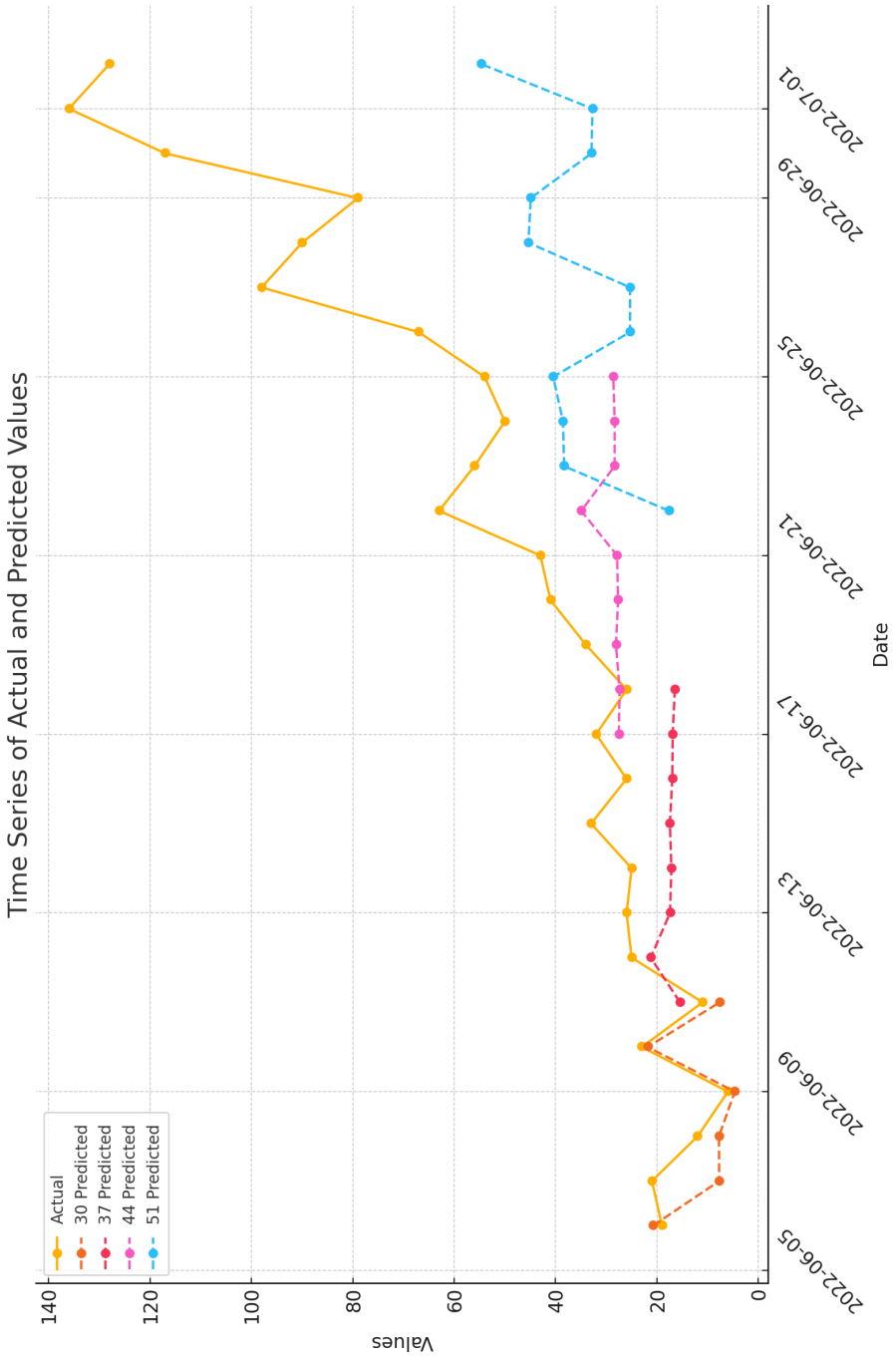


Figure 6.5 Preliminary Predictions of Blending

Table 6.1 Preliminary Predictions of Blending

Times	Number of training set	Number of validation set	Number of test set	RMSE	MAE
30 day	18 day	6 day	6 day	5.94	4.19
37 day	23 day	7 day	7 day	10.06	9.22
44 day	26 day	9 day	9 day	18.64	15.88
51 day	31 day	10 day	10 day	57.18	49.25

These findings clearly indicate that as the size of the training dataset increases, the model's prediction error significantly increases. This could be attributed to several factors:

(a) The complexity of features in the training and testing data of the model: The blending model demonstrates promising performance in training on COVID-19 data and making predictions on COVID-19 datasets. However, as the volume of monkey pox data increases, the model may struggle to capture all features and patterns during training. Consequently, the model may exhibit signs of underfitting, leading to an increase in errors on the validation and test sets.

(b) Variations in data distribution: The blending model is trained on a COVID-19 dataset, and over time, the monkey pox epidemic may exhibit varying patterns of transmission. Discrepancies in the data distribution between the training and validation sets may hinder the model's ability to accurately predict future trends.

6.7.2 Comparing the effects of blending transfer learning with incremental learning for predicting monkey pox with the incorporation of the biological feature Rt.

This study further evaluated a blending model that demonstrated robust performance in training on COVID-19 data, utilizing transfer learning and incremental learning on the monkey pox dataset. To increase the predictive accuracy of the model, This study introduced the biological feature Rt to capture the dynamic changes in epidemic spread. The model's prediction results with and without the Rt feature are presented in Table 6.2 and Fig 6.6 for varying sliding time windows.

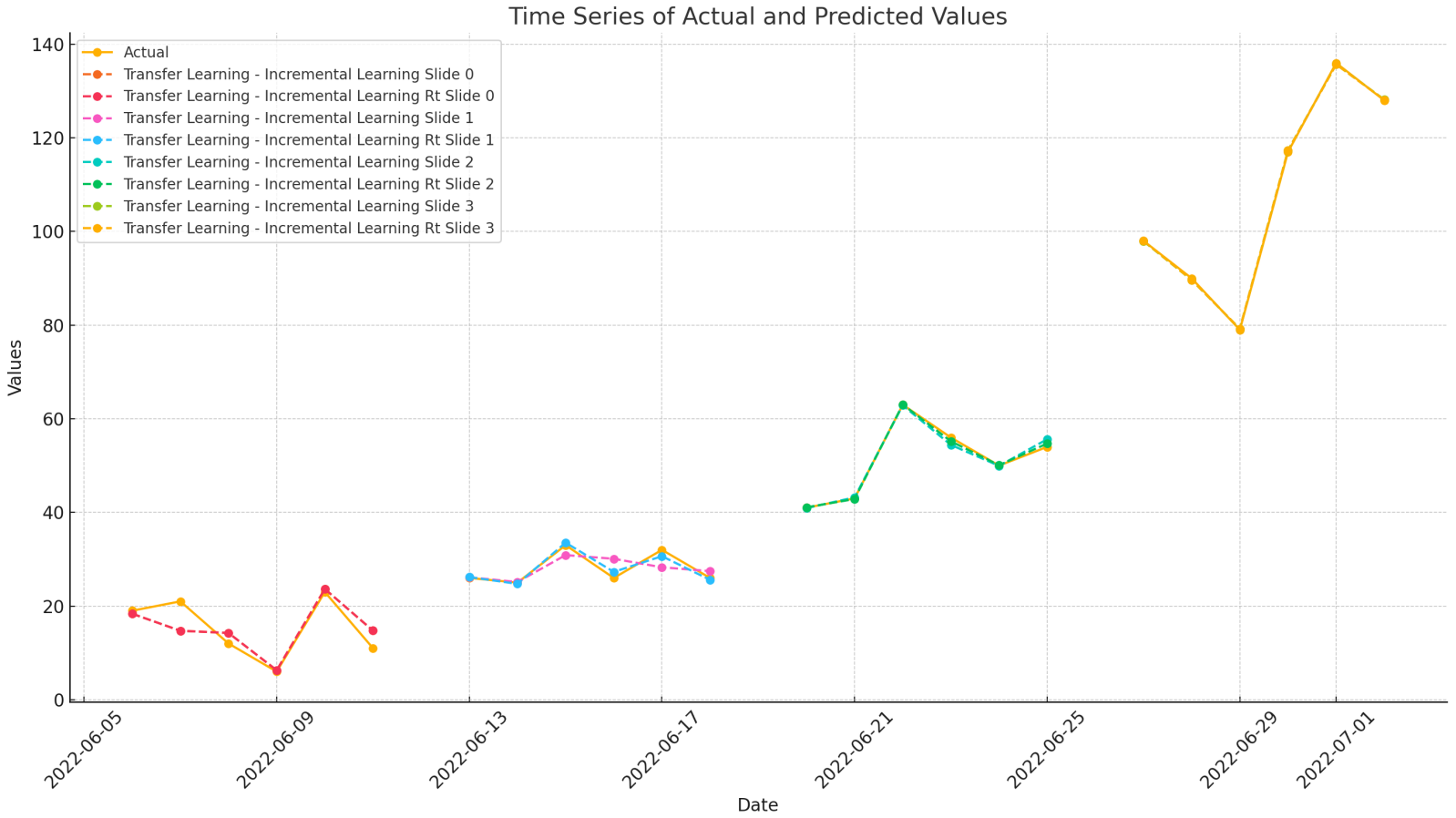


Figure 6.6 Compares the model prediction.

Table 6.2 Compares the effects of transfer learning, incremental learning, and Rt-based monkeypox prediction

Model Type	Times	RMSE	MAE
Transfer Learning Incremental Learning	Slide 0 weeks	3.14	2.3
Transfer Learning Incremental Learning Rt	Slide 0 weeks	3.14	2.3
Transfer Learning Incremental Learning	Slide 1 weeks	2.49	1.94
Transfer Learning Incremental Learning Rt	Slide 1 weeks	0.82	0.67
Transfer Learning Incremental Learning	Slide 2 weeks	0.93	0.61
Transfer Learning Incremental Learning Rt	Slide 2 weeks	0.45	0.31
Transfer Learning Incremental Learning	Slide 3 weeks	0.27	0.25
Transfer Learning Incremental Learning Rt	Slide 3 weeks	0.21	0.19

By employing transfer learning, This study utilize the pertinent features of the COVID-19 dataset to construct a model for the monkeypox data. Incremental learning facilitates the model's progressive updates to accommodate the evolving data. The findings reveal a gradual reduction in prediction errors as the time window shifts, underscoring the adaptiveness of incremental learning to the ongoing integration of new data.

6.7.3 Introducing the effect of the biological trait Rt

Among all the tested models, the inclusion of the biological feature Rt demonstrated a significant performance improvement in each sliding window phase. In particular, following the first week of sliding, the model's RMSE decreased significantly from 2.49--0.82, and the MAE decreased from 1.94--0.67. These results indicate that the Rt feature substantially enhances the model's ability to capture epidemic transmission trends. Transfer learning leverages features from previous COVID-19 datasets, resulting in notable performance gains in data-scarce scenarios. The initial findings show that at sliding week 0, the model's RMSE and MAE were 3.14 and 2.30, respectively. As the sliding window progresses, incremental learning gradually updates the model to adapt to new monkeypox data. This study observed a significant decrease in the model's RMSE and MAE with each additional week of data, highlighting the effectiveness of incremental learning in adapting to data changes. The introduction of Rt led to improved predictive performance across all time windows, particularly in the sliding windows of the first and second weeks. Rt, as a key biological feature, can reflect real-time changes in epidemic transmission; thus, its incorporation into the model significantly enhances prediction accuracy.

This study, investigate the application and effectiveness of the blending model, transfer learning, incremental learning, and biological feature Rt in predicting emerging infectious monkeypox. The experimental results demonstrate the significant advantages of these methods in addressing the challenges of early-stage emerging infectious disease data scarcity and real-time updates.

6.7.4 comparison of models before and after the ensemble enhancement

In order to compare the performance of the hybrid model with the model enhanced by transfer learning and incremental learning, the prediction performance was compared using 14 days of monkeypox data. The performance differences between models are evident from the RMSE (root mean square error) and MAE (mean absolute error) metrics in the table. The combined model of transfer learning

and incremental learning techniques exhibits excellent performance on this dataset, with an RMSE value of 3.14 and a MAE value of 2.3, which reflects the higher prediction accuracy and lower error, highlighting the effectiveness of these techniques in dealing with data scarcity cases. In contrast, the performance of the hybrid model before optimization was poor, with RMSE and MAE values of 18.48 and 16.99, respectively, indicating that the model error was large before careful tuning. This comparison highlights the potential of transfer learning and incremental learning to improve prediction accuracy and model efficiency, as well as the importance of hybrid models that need to be fully optimized before integration (see Fig 6.7 and Table 6.3).

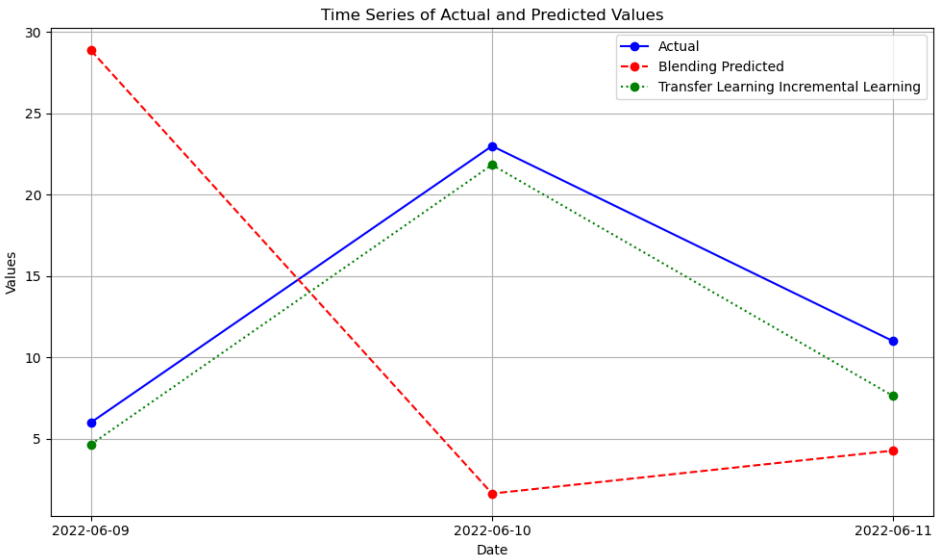


Figure 6.7 blending model before and after enhancement

Table 6.3 Comparison of blending model before and after enhancement

Model Type	Data Set	RMSE	MAE
Transfer Learning Incremental Learning	monkeypox 14 day	3.14	2.3
Blending	monkeypox 14 day	18.48	16.99

6.7.5 The Applicability of the Blending model in Predicting Emerging Infectious Diseases

The blending model enhances overall predictive performance by combining predictions from multiple base models. Specifically, when predicting the next 6 days, the model achieves low RMSE and MAE values of 5.94 and 4.19, respectively. However, with increasing data, particularly at 51 days, the forecasts for the next 10 days show significant increases in the RMSE and MAE to 57.18 and 49.25, respectively. This suggests that the blending model performs well in short-term forecasts (within 7 days). The findings indicate that the blending model is effective for short-term predictions but may require further optimization for handling long time series data. Adjusting model complexity or incorporating additional data preprocessing steps, such as feature selection or dimensionality reduction, may help improve the predictive accuracy of long time series data (Deng *et al.*, 2021; Kang *et al.*, 2024).

6.7.6 Advantages of transfer learning and incremental learning in the context of data scarcity and dynamic updates

The application of transfer learning and incremental learning in this study demonstrates their potential in addressing data scarcity and real-time dynamic updates (Narkhede *et al.*, 2021; Noordman *et al.*, 2023). The experimental results show that by transferring features from COVID-19 data to the monkeypox dataset, the model can still provide relatively accurate predictions in the early stages of data scarcity. For instance, at week 0 of the sliding window, the model's RMSE and MAE are 3.14 and 2.30, respectively, highlighting the advantage of transfer learning in leveraging existing knowledge for new tasks. Furthermore, incremental learning techniques enable the model to adapt to new data in real time, maintaining prediction accuracy. With each week of increase in the data, the model's prediction errors significantly decrease, reaching RMSE and MAE values of 0.27 and 0.25, respectively, after 3 weeks of sliding. This indicates the effectiveness of incremental learning in real-time data updates, particularly in dealing with continuously changing data distributions. These results underscore the key advantages of transfer learning and incremental learning:

the study can initiate predictions in data-scarce scenarios and continuously update and optimize the model as new data arrives. This is particularly crucial for predicting new infectious diseases, as early-stage data are often limited and dynamically changing.

6.7.7 The advantages of the biological feature Rt in the early prediction of emerging infectious diseases

The inclusion of the biological feature Rt significantly enhances the predictive performance of the model, particularly in the very early stages of an epidemic. Rt, a key epidemiological indicator, reflects the average number of individuals to whom an infected person can transmit the virus under existing conditions (Alvarez *et al.*, 2021; Li *et al.*, 2024d). The experimental data demonstrate that incorporating Rt results in a significant decrease in both the RMSE and the MAE of the model; for example, during a one-week sliding window, the RMSE decreases from 2.49 to 0.82, and the MAE decreases from 1.94 to 0.67. This indicates the effectiveness of the Rt feature in capturing the dynamics of epidemic spread, providing the model with more precise trend information. In the initial stages of a novel infectious disease outbreak, data are often limited and unstable; therefore, Rt can serve as a valuable feature, aiding the model in rapidly adapting to new circumstances and offering more accurate predictions.

6.7.8 Potential Applications of Models in Public Health Decision-making

The blending prediction model developed in this study holds significant potential for application in public health decision-making. By integrating transfer learning, incremental learning, and the biological feature Rt, This study can provide highly accurate early epidemic forecasts for emerging infectious diseases, which is crucial for resource allocation and emergency response. For example, the model can predict the growth trend of case numbers in the next 7 days, aiding health authorities in preparing medical resources in advance, formulating isolation policies, and optimizing vaccine distribution strategies. Accurate predictions can significantly reduce public

health risks, enhancing the timeliness and effectiveness of prevention and control measures. Compared with previous studies, this research has made several important advancements. First, This study introduce the blending model, which synthesizes the strengths of various models when dealing with multiple model outputs, thereby enhancing the overall predictive performance. Second, through transfer learning and incremental learning, This study successfully leveraged knowledge from COVID-19 data, significantly improving the accuracy of monkeypox epidemic prediction. Finally, the incorporation of the biological feature Rt provides profound insights into the dynamics of epidemic spread, enabling the model to more accurately capture changes in disease transmission trends. These enhancements position our model as superior to many traditional methods in terms of predictive accuracy and adaptability.

6.7.9 Insights for the development of future infectious disease prediction tools

The results of this study provide several important insights for the development of future infectious disease prediction tools. First, the integration of multiple models (such as the blending model) is an effective method for improving prediction accuracy, particularly when dealing with complex and variable data (Xie *et al.*, 2023). Second, transfer learning performs well in cases of data scarcity, indicating that leveraging existing relevant data for knowledge transfer can significantly enhance model performance (Yang *et al.*, 2024). Incremental learning enables real-time prediction and dynamic model updates. Third, the incorporation of biological features (such as Rt) can provide crucial epidemiological information to the model, aiding in capturing potential changes in outbreaks. Future research could further explore the optimization of these methods. For example, more advanced feature selection and extraction techniques should be investigated to further enhance the predictive capabilities of the models. Additionally, the development of real-time data updating and automated model tuning systems will make prediction tools more efficient and reliable in practical applications.

6.7.10 Limitations of the research

One significant limitation of this study lies in the impact of data quality and quantity on the model's performance. The data utilized originate primarily from public sources such as the monkeypox dataset provided by Our World in Data. These datasets may suffer from inconsistencies in data collection, reporting delays, or missing information, all of which can potentially affect the predictive accuracy of the model (Ciobanu-Caraus *et al.*, 2024; Du *et al.*, 2020; Ortiz-Barrios *et al.*, 2024). For example, in cases of sparse data, the model may struggle to adequately capture the characteristics of disease transmission, leading to increased prediction errors. Furthermore, the temporal span and geographical coverage of the data also influence the model's generalizability. During the early stages of an epidemic, data are typically limited and unstable, which could result in model overfitting to the restricted training data and failure to accurately forecast future trends. Therefore, the quality and quantity of data are critical factors influencing model performance, particularly in addressing rapidly evolving infectious disease outbreaks (De Salazar *et al.*, 2022; Vobugari *et al.*, 2022).

Another limitation is the generalizability of the model in predicting different diseases. While this study improved the prediction accuracy of monkeypox outbreaks by utilizing features from COVID-19 data through transfer learning, the generalizability of this approach may be limited. Different diseases have distinct transmission mechanisms and characteristics, such as routes of transmission, incubation periods, and severities of infection. Therefore, the successful application of a model for one disease does not guarantee similar performance for other diseases. Further research is needed to explore ways to enhance the model's generalization ability to better adapt to various infectious disease scenarios. This may involve developing more universal feature extraction methods or incorporating more biological and epidemiological knowledge into the model. Additionally, integrating multiple data sources, such as epidemiological survey data, genomic data, and environmental data, may help improve the model's prediction accuracy and generalizability (Yu *et al.*, 2023).

6.8 Chapter Summary

This study investigates the application of the blending model trained with COVID -19 data, combined with transfer learning and incremental learning, in the prediction of monkeypox outbreaks. The incorporation of the biological feature Rt significantly enhances the predictive accuracy of the model. The research findings indicate the following:

(a) ***Evaluation of the blending model's performance and limitations:*** The blending model proves effective in short-term forecasting but may require further optimization when handling long time series data. Adjusting model complexity or incorporating additional data preprocessing steps, such as feature selection or dimensionality reduction, could enhance the predictive accuracy of long time series data.

(b) ***The advantages of transfer learning:*** By leveraging the features of a COVID-19 dataset, This study successfully applied it to predict monkeypox outbreaks. The utilization of transfer learning significantly enhanced the model's predictive capacity in scenarios of limited data availability, underscoring the critical importance of leveraging interdisciplinary knowledge in forecasting emerging infectious diseases.

(c) ***Dynamic adaptability of incremental learning:*** Through incremental learning, models can dynamically adapt to changes in new data, maintaining a high level of predictive accuracy. This technique is particularly suitable for situations involving dynamic changes, such as in the case of epidemic outbreaks, enabling real-time forecast updates.

(d) ***Introduction of the biological feature Rt:*** The inclusion of the effective reproduction number (Rt) as a crucial indicator reflecting the dynamics of disease spread significantly enhances the predictive ability of models. Particularly in the early stages of emerging infectious diseases, incorporating Rt assists models in capturing changes in disease transmission more accurately.

This study demonstrates the potential application of a blending model that integrates transfer learning, incremental learning, and biological feature Rt in infectious

disease prediction, offering reliable technological support for public health emergency responses.

While this study yielded valuable results, there are still numerous areas that warrant further exploration. The following are some research recommendations for future investigations in the field of forecasting emerging infectious diseases:

(a) ***Improving Data Quality and Integrating Multiple Data Sources:*** Future research should focus on enhancing the quality and usability of data, particularly in situations where early epidemic data are scarce. Integrating multiple data sources, such as epidemiological, environmental, and genomic data, will aid in the construction of more comprehensive and precise predictive models.

(b) ***Real-time prediction and automated updating system:*** With the continuous influx of new data, the development of real-time prediction and automated model updating systems is becoming increasingly crucial. Such systems should be capable of automatically acquiring new data, updating model parameters, and generating prediction outcomes to provide timely information support to decision-makers.

(c) ***Interdisciplinary collaboration:*** The prediction of emerging infectious diseases involves knowledge from various fields, such as epidemiology, statistics, and data science. Future research should emphasize interdisciplinary collaboration, integrating the latest findings from different disciplines to develop more precise and practical forecasting models.

Through further research and exploration, This study can continuously enhance the accuracy and practicality of infectious disease prediction models, thus providing more robust support for global public health.

CHAPTER 7

CONCLUSION AND RECOMMENDATIONS

7.1 Research Objective Achievement

This research systematically explores the application and optimization of machine learning techniques, including multimodel ensembles, transfer learning, and incremental learning, in the field of infectious disease prediction. The results demonstrate that the proposed prediction model, which integrates multiobjective optimization and a blending ensemble model, has achieved significant progress in predicting emerging infectious diseases such as COVID-19 and monkeypox. The specific research findings are summarized as follows (see Table7.1):

(a) ***Effectiveness of multiobjective optimization in model selection*** By employing the NSGA-II multiobjective optimization algorithm, this study successfully selects decision tree (DT) and extreme gradient boosting regression (XGBoost) models for COVID-19 prediction. Compared with traditional single-objective optimization methods (such as ridge regression), the models selected via multiobjective optimization presented significant advantages in terms of accuracy, generalizability, and computational efficiency. The evaluation results on real-world datasets revealed that the RMSE of the DT and XGBoost models were significantly lower than those of the models selected via single-objective optimization, fully validating the potential advantages of multiobjective optimization in balancing multiple evaluation metrics. This research provides new theoretical support and practical methods for the selection of prediction models in public health decision support systems.

(b) ***Construction and Optimization of the Blending Ensemble model*** To further improve the performance of COVID-19 transmission prediction, this study selected base models through multiobjective optimization algorithms and constructed a blending ensemble learning model using linear regression as

the meta-model. The results showed that, compared with single models and stacking methods, the blending model exhibited superior performance in terms of accuracy, generalizability, and computational efficiency. This research demonstrates the potential of applying the blending ensemble model in the field of infectious disease prediction, ensuring both prediction accuracy and improving model interpretability and practical value, providing a reference for prediction applications in different diseases and scenarios.

(c) ***To apply multii-model fusion in emerging infectious disease prediction*** for the prediction of emerging infectious diseases, this study combined the blending model with transfer learning and incremental learning techniques to propose an innovative prediction method. By transferring features from the COVID-19 dataset to the monkeypox dataset and introducing dynamically updated incremental learning techniques, the model demonstrated excellent prediction capabilities even under data scarcity. For short-term prediction (7 days), the RMSE of the model improved by 91.41%, and the MAE improved by 89.13%. Furthermore, combining with the biological feature Rt further improved the accuracy of the model, with the RMSE and MAE improving by 1.91% and 2.17%, respectively. These research results demonstrate the significant application potential of multimodel fusion and real-time data updates in infectious disease prediction, providing theoretical support and technical guarantees for public health emergency response.

Table 7.1 Research questions and objectives

research questions	objectives	solution
Difficulty in model selection	To identify base models from deep learning and machine learning for prediction of COVID-19 cases using a multi-objective optimization approach	Identifying Underlying Models from Deep Learning and Machine Learning Techniques Using Multi-Objective Optimisation Methods to Predict COVID-19 Cases
Lack of widely applicable models	To propose a blended multi-model ensemble from identified based models for more accurate COVID-19 cases prediction.	A blending multi-model integration based on identified base models is proposed to improve the prediction accuracy of COVID-19 cases.
Scarcity of data on early infectious diseases	To enhance the blended ensemble model with transfer learning and incremental learning for better adaptation to early prediction of emerging infectious diseases.	propose transfer learning and incremental learning to blending integrated models to better suit the task of early prediction of emerging infectious diseases.

This study successfully constructed an infectious disease prediction model that integrates multiple models and methods, and improved the model's adaptability and practicality in predicting emerging infectious diseases through multitask transfer learning and incremental learning techniques. The research results not only enrich the theoretical foundation of infectious disease prediction models but also provide reliable technical support for public health emergency response. The practical application potential and theoretical significance of these results lay a solid foundation for future research.

7.2 Research Contributions

This research makes significant contributions to the field of infectious disease prediction, particularly in the application and optimization of multimodel ensemble learning models. Our key contributions are as follows:

(a) ***Pioneering application of multiobjective optimization in model selection:*** This study proposed the NSGA-II algorithm to select optimal prediction models in infectious disease forecasting, effectively balancing accuracy, generalization, and computational efficiency. This novel approach outperforms traditional single-objective methods, enriching the theoretical foundation of model selection in this domain.

(b) ***Innovative application of the blending ensemble in public health:*** This study have demonstrated the effectiveness of the blending ensemble model in improving the performance and practicality of COVID-19 prediction models. By combining base models selected through multiobjective optimization and a linear regression meta-model, our proposed model achieves substantial improvements in prediction accuracy, generalizability, and computational efficiency.

(c) ***Multitask Transfer Learning and Incremental Learning for Infectious Disease Prediction:*** This study have developed a novel model that leverages transfer learning and incremental learning to increase the prediction accuracy of models for emerging infectious diseases. By transferring knowledge

from COVID-19 datasets to monkeypox datasets and incorporating dynamic updates, our model effectively addresses data scarcity challenges, expanding the applicability of traditional prediction models and providing new theoretical and technical support for early warning of emerging diseases.

(d) ***Incorporation of the biological feature Rt and its impact on model performance:*** This study introduced the biological feature Rt into our monkeypox prediction model and demonstrated its effectiveness in improving prediction accuracy. By incorporating Rt, our model more accurately reflects the dynamics of disease transmission and achieves significant reductions in the RMSE and MAE. This innovation offers new perspectives for designing future infectious disease prediction models, emphasizing the importance of integrating biological features.

(e) ***A Comprehensive model Integrating Theory and Practice:*** This study have integrated multimodel fusion, transfer learning, and incremental learning to develop a versatile infectious disease prediction model. This model not only provides new insights into the combination of these techniques but also offers a reliable technical support for public health emergency response systems. The integration of theory and practice makes this model highly applicable to future infectious disease prediction and related fields.

In summary, this research makes several innovative contributions to the field of infectious disease prediction, including novel approaches to model selection, ensemble learning, transfer learning, and the incorporation of biological features. These contributions expand existing theoretical knowledge and provide valuable technical support for practical applications, making significant contributions to public health decision-making and emergency response.

7.3 Future Works

While this study has made significant advancements in infectious disease prediction, several limitations and avenues for future research remain. The following areas warrant further investigation:

(a) ***Model Optimization and Extension:*** Although this study has demonstrated the effectiveness of multiobjective optimization algorithms in model selection, future research can explore more sophisticated and efficient multiobjective optimization methods. For example, investigating more intelligent optimization algorithms, such as the combination of genetic algorithms and deep reinforcement learning, could further increase the efficiency and accuracy of model selection on larger and more complex datasets. Additionally, incorporating a wider variety of machine learning and deep learning models into the ensemble model could increase the robustness and flexibility of predictions.

(b) ***Cross-Domain Application of Ensemble Learning models:*** The blending ensemble learning model presented in this study has shown promise in COVID-19 and monkeypox prediction. However, future research can apply this model to predict other infectious diseases and even expand its application to other domains, such as climate change prediction and financial market analysis. By validating the blending model in diverse application domains, its generality and adaptability can be further enhanced, and customized ensemble model can be developed to meet the specific needs of different scenarios.

(c) ***Integration of Real-Time Data and Dynamic Updates:*** While this study has preliminarily validated the effectiveness of incremental learning techniques in data-scarce scenarios, how to integrate real-time data more effectively and perform dynamic updates remains a key focus of future research. Future research can explore how to combine stream data analysis techniques with existing incremental learning model to achieve more real-time infectious disease prediction and early warning.

(d) ***Integration and Analysis of Multi-Source Data:*** Current research focuses primarily on single-source datasets. However, future research can explore how to integrate multisource data (such as social media data, epidemiological data, and meteorological data) into prediction models. By fusing multisource data, richer information can be provided to the model, thereby improving the accuracy and stability of predictions. Future research can develop novel data fusion techniques and multimodal learning methods to effectively process and analyse heterogeneous data from different sources.

(e) ***Model interpretability and usability:*** Although the blending ensemble model presented in this study exhibits excellent prediction performance, its complexity may pose challenges to model interpretability and usability. Future research should consider how to improve model interpretability while maintaining performance, especially for nontechnical users. New interpretability methods can be explored, or user-friendly interfaces can be developed to enable public health decision-makers to better understand and apply prediction results.

(f) ***Deep Integration of Biological Features and Machine Learning:*** This study has improved prediction accuracy by introducing the biological feature Rt. Future research can further explore how to integrate more biological features (such as the immune response and host genetic features) in greater depth with machine learning models. By incorporating richer biological data, the biological interpretability and clinical applicability of infectious disease prediction models can be further enhanced.

In conclusion, future research can continue to deepen our understanding in areas such as model optimization, cross-domain applications, real-time data integration, multisource data analysis, model interpretability, and the integration of biological features. Through these efforts, can overcome the limitations of current research and provide more effective and comprehensive solutions for infectious disease prediction and public health emergency response.

REFERENCES

Abdulaal, A., Patel, A., Charani, E., Denny, S., Mughal, N. and Moore, L. (2020). Prognostic Modeling of COVID-19 Using Artificial Intelligence in the United Kingdom: Model Development and Validation. *Journal of Medical Internet Research*. 22(8).

Abraham, A., Le, B., Kosti, I., Straub, P., Velez-Edwards, D. R., Davis, L. K., Newton, J. M., Muglia, L. J., Rokas, A., Bejan, C. A., Sirota, M. and Capra, J. A. (2022). Dense phenotyping from electronic health records enables machine learning-based prediction of preterm birth. *BMC medicine*. 20(1), 333.

Adnan, M., Altalhi, M., Alarood, A. A. and Uddin, M. I. (2022). Modeling the Spread of COVID-19 by Leveraging Machine and Deep Learning Models. *Intelligent Automation and Soft Computing*. 31(3), 1857–1872.

Ahmed, D. M., Hassan, M. M. and Mstafa, R. J. (2022). A review on deep sequential models for forecasting time series data. *Applied Computational Intelligence and Computing*. 2022.

Akbulut, S., Yagin, F. H., Cicek, I. B., Koc, C., Colak, C. and Yilmaz, S. (2023). Prediction of Perforated and Nonperforated Acute Appendicitis Using Machine Learning-Based Explainable Artificial Intelligence. *Diagnostics (Basel, Switzerland)*. 13(6).

Al-Qahtani, A. A. (2020). Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): Emergence, history, basic and clinical aspects. *Saudi journal of biological sciences*. 27(10), 2531–2538.

Alalayah, K. M., Senan, E. M., Atlam, H. F., Ahmed, I. A. and Shatnawi, H. S. A. (2023). Effective Early Detection of Epileptic Seizures through EEG Signals Using Classification Algorithms Based on t-Distributed Stochastic Neighbor Embedding and K-Means. *Diagnostics (Basel, Switzerland)*. 13(11).

Ala'raj, M., Majdalawieh, M. and Nizamuddin, N. (2021). Modeling and forecasting of COVID-19 using a hybrid dynamic model based on SEIRD with ARIMA corrections. *Infectious Disease Modelling*. 6, 98–111.

Albeshri, A. (2021). DLDW: Deep Learning and Dynamic Weighing-based Method for Predicting COVID-19 Cases in Saudi Arabia. *International Journal of Computer Science Network Security*. 21(9), 212–222.

Aldhyani, T. H. H., Alrasheed, M., Al-Adaileh, M. H., Alqarni, A. A., Alzahrani, M. Y. and Alahmadi, A. H. (2021). Deep Learning and Holt-Trend Algorithms for Predicting Covid-19 Pandemic. *Cmc-Computers Materials Continua*. 67(2), 2141–2160.

Alizargar, A., Chang, Y. L., Alkhaleefah, M. and Tan, T. H. (2024). Precision Non-Alcoholic Fatty Liver Disease (NAFLD) Diagnosis: Leveraging Ensemble Machine Learning and Gender Insights for Cost-Effective Detection. *Bioengineering (Basel, Switzerland)*. 11(6).

Alqaissi, E. Y., Alotaibi, F. S. and Ramzan, M. S. (2022). Modern Machine-Learning Predictive Models for Diagnosing Infectious Diseases. *Computational and Mathematical Methods in Medicine*. 2022(1), 1–13.

Alvarez, L., Colom, M., Morel, J. D. and Morel, J. M. (2021). Computing the daily reproduction number of COVID-19 by inverting the renewal equation using a variational technique. *Proceedings of the National Academy of Sciences of the United States of America*. 118(50).

Anastario, M., Rink, E., Firemoon, P., Carnegie, N., Johnson, O., Peterson, M. and Rodriguez, A. M. (2023). Evidence of secular trends during the COVID-19 pandemic in a stepped wedge cluster randomized trial examining sexual and reproductive health outcomes among Indigenous youth. *Trials*. 24(1), 248.

Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., Rabczuk, T. and Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. *Algorithms*. 13(10), 249.

Aronu, C. O., Ekwueme, G. O., Sol-Akubude, V. I. and Okafor, P. N. (2020). Coronavirus (COVID-19) in Nigeria: survival rate. *Scientific African*, e00689.

Arora, P., Kumar, H. and Panigrahi, B. K. (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons Fractals*. 139, 110017.

Atangana, A. and İğret Araz, S. (2020). Mathematical model of COVID-19 spread in Turkey and South Africa: theory, methods, and applications. *Advances in Difference Equations*. 2020(1), 1–89.

Awan, T. M. and Aslam, F. (2020). Prediction of daily COVID-19 cases in European countries using automatic ARIMA model. *Journal of public health research*. 9(3), 1765.

Ayoobi, N., Sharifrazi, D., Alizadehsani, R., Shoeibi, A., Gorriz, J. M., Moosaei, H., Khosravi, A., Nahavandi, S., Chofreh, A. G. and Goni, F. A. (2021). Time series forecasting of new cases and new deaths rate for COVID-19 using deep learning methods. *Results in Physics*. 27, 104495.

Babatunde, H. A., Collins, J., Lukman, R., Saxton, R., Andersen, T. and McDougal, O. M. (2024). SVR Chemometrics to Quantify -Lactoglobulin and -Lactalbumin in Milk Using MIR. *Foods (Basel, Switzerland)*. 13(1).

Babor, M., Paquet-Durand, O., Kohlus, R. and Hitzmann, B. (2023). Modeling and optimization of bakery production scheduling to minimize makespan and oven idle time. *Scientific reports*. 13(1), 235.

Babuna, P., Yang, X., Gyilbag, A., Awudi, D. A., Ngmenbelle, D. and Bian, D. (2020). The Impact of COVID-19 on the Insurance Industry. *International journal of environmental research and public health*. 17(16).

Bai, T., Zhou, S., Pang, Y., Luo, J., Wang, H. and Du, Y. (2023). An image caption model based on attention mechanism and deep reinforcement learning. *Frontiers in Neuroscience*. 17(1), 1270850.

Banyal, S., Dwivedi, R., Gupta, K. D., Sharma, D. K., Al-Turjman, F. and Mostarda, L. (2021). Technology Landscape for Epidemiological Prediction and Diagnosis of COVID-19. *Cmc-Computers Materials Continua*, 1679–1696.

Bar-Or, I., Indenbaum, V., Weil, M., Elul, M., Levi, N., Aguvaev, I., Cohen, Z., Levy, V., Azar, R., Mannasse, B., Shirazi, R., Bucris, E., Mor, O., Sela Brown, A., Sofer, D., Zuckerman, N. S., Mendelson, E. and Erster, O. (2022). National Scale Real-Time Surveillance of SARS-CoV-2 Variants Dynamics by Wastewater Monitoring in Israel. *Viruses*. 14(6).

Barea-Sepúlveda, M., Calle, J. L. P., Ferreiro-González, M. and Palma, M. (2023). Rapid Classification of Petroleum Waxes: A Vis-NIR Spectroscopy and Machine Learning Approach. *Foods (Basel, Switzerland)*. 12(18).

Bolla, G., Berente, D. B., Andrássy, A., Zsuffa, J. A., Hidasi, Z., Csibri, E., Csukly, G., Kamondi, A., Kiss, M. and Horvath, A. A. (2023). Comparison of the diagnostic accuracy of resting-state fMRI driven machine learning algorithms in the detection of mild cognitive impairment. *Scientific reports*. 13(1), 22285.

Borré, A., Seman, L. O., Camponogara, E., Stefenon, S. F., Mariani, V. C. and Coelho, L. D. S. (2023). Machine Fault Detection Using a Hybrid CNN-LSTM Attention-Based Model. *Sensors (Basel, Switzerland)*. 23(9).

Boulant, O., Fekom, M., Pouchol, C., Evgeniou, T., Ovchinnikov, A., Porcher, R. and Vayatis, N. (2020). SEAIR framework accounting for a personalized risk prediction score: application to the Covid-19 epidemic. *Image Processing On Line*. 10, 150–166.

Budholiya, K., Shrivastava, S. K. and Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University-Computer and Sciences*. 34(7), 4514–4523.

Błażkiewicz, M. (2022). Evaluation of Geometric Attractor Structure and Recurrence Analysis in Professional Dancers. *Entropy (Basel, Switzerland)*. 24(9).

Caballero, R., Martínez, M. and Peña, E. (2023). Coronary artery properties in atherosclerosis: A deep learning predictive model. *Frontiers in physiology*. 14, 1162436.

Cai, J., Lu, S., Cheng, J., Wang, L., Gao, Y. and Tan, T. (2022a). Collaborative variable neighborhood search for multi-objective distributed scheduling in two-stage hybrid flow shop with sequence-dependent setup times. *Scientific reports*. 12(1), 15724.

Cai, L., Zhao, E., Niu, H., Liu, Y., Zhang, T., Liu, D., Zhang, Z., Li, J., Qiao, P., Lv, H., Ren, P., Zheng, W. and Wang, Z. (2023). A machine learning approach to predict cerebral perfusion status based on internal carotid artery blood flow. *Computers in biology and medicine*. 164, 107264.

Cai, W., Zhang, Q. and Cui, J. (2022b). A Novel Fault Diagnosis Method for Denoising Autoencoder Assisted by Digital Twin. *Computational intelligence and neuroscience*. 2022, 5077134.

Canino, M. P., Cesario, E., Vinci, A. and Zarin, S. (2022). Epidemic forecasting based on mobility patterns: an approach and experimental evaluation on COVID-19 Data. *Social network analysis and mining*. 12(1), 116.

Cao, W., Zhu, J., Wang, X., Tong, X., Tian, Y., Dai, H. and Ma, Z. (2022). Optimizing Spatio-Temporal Allocation of the COVID-19 Vaccine Under Different Epidemiological Landscapes. *Frontiers in public health*. 10, 921855.

Cao, X., Hou, Y., Zhang, X., Xu, C., Jia, P., Sun, X., Sun, L., Gao, Y., Yang, H., Cui, Z., Wang, Y. and Wang, Y. (2020). A comparative, correlate analysis and projection of global and regional life expectancy, healthy life expectancy, and their GAP: 1995-2025. *Journal of global health*. 10(2), 020407.

Castillo Ossa, L. F., Chamoso, P., Arango-López, J., Pinto-Santos, F., Isaza, G. A., Santa-Cruz-González, C., Ceballos-Marquez, A., Hernández, G. and Corchado, J. M. (2021). A hybrid model for COVID-19 monitoring and prediction. *Electronics*. 10(7), 799.

Champredon, D., Papst, I. and Yusuf, W. (2024). ern: An [Formula: see text] package to estimate the effective reproduction number using clinical and wastewater surveillance data. *PloS one*. 19(6), e0305550.

Changyong, F., Hongyue, W., Naiji, L., Tian, C., Hua, H. and Ying, L. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*. 26(2), 105.

Chaudhary, V., Khanna, V., Ahmed Awan, H. T., Singh, K., Khalid, M., Mishra, Y. K., Bhansali, S., Li, C. Z. and Kaushik, A. (2023). Towards hospital-on-chip supported by 2D MXenes-based 5th generation intelligent biosensors. *Biosensors bioelectronics*. 220, 114847.

Chen, J., Guo, C., Lu, M. and Ding, S. (2021). Unifying Diagnosis Identification and Prediction Method Embedding the Disease Ontology Structure From Electronic Medical Records. *Frontiers in public health*. 9, 793801.

Cheng, Q., Collender, P. A., Heaney, A. K., McLoughlin, A., Yang, Y., Zhang, Y., Head, J. R., Dasan, R., Liang, S. and Lv, Q. (2022). Optimizing laboratory-based surveillance networks for monitoring multi-genotype or multi-serotype infections. *PLoS computational biology*. 18(9), e1010575.

Cheuque, C., Querales, M., León, R., Salas, R. and Torres, R. (2022). An Efficient Multi-Level Convolutional Neural Network Approach for White Blood Cells Classification. *Diagnostics (Basel, Switzerland)*. 12(2).

Chew, A. W. Z., Pan, Y., Wang, Y. and Zhang, L. (2021). Hybrid deep learning of social media big data for predicting the evolution of COVID-19 transmission. *Knowledge-Based Systems*. 233, 107417.

Chi, M., An, H., Jin, X. and Nie, Z. (2024). An N-Shaped Lightweight Network with a Feature Pyramid and Hybrid Attention for Brain Tumor Segmentation. *Entropy (Basel, Switzerland)*. 26(2).

Choi, S. H., Park, J. K., An, D., Kim, C. H., Park, G., Lee, I. and Lee, S. (2023). Fault Diagnosis Method for Human Coexistence Robots Based on Convolutional Neural Networks Using Time-Series Data Generation and Image Encoding. *Sensors (Basel, Switzerland)*. 23(24).

Choisy, M., Guégan, J.-F. and Rohani, P. (2007). Mathematical modeling of infectious diseases dynamics. *Encyclopedia of infectious diseases: modern methodologies*. 379.

Ciobanu-Caraus, O., Aicher, A., Kernbach, J. M., Regli, L., Serra, C. and Staartjes, V. E. (2024). A critical moment in machine learning in medicine: on reproducible and interpretable learning. *Acta neurochirurgica*. 166(1), 14.

Cooper, I., Mondal, A., Antonopoulos, C. G. and Mishra, A. (2022). Dynamical analysis of the infection status in diverse communities due to COVID-19 using a modified SIR model. *Nonlinear dynamics*. 109(1), 19–32.

Corsi, A., de Souza, F. F., Pagani, R. N. and Kovaleski, J. L. (2021). Big data analytics as a tool for fighting pandemics: a systematic review of literature. *Journal of ambient intelligence and humanized computing*. 12(10), 9163–9180.

Cui, J., Li, K., Hao, J., Dong, F., Wang, S., Rodas-González, A., Zhang, Z., Li, H. and Wu, K. (2022). Identification of Near Geographical Origin of Wolfberries

by a Combination of Hyperspectral Imaging and Multi-Task Residual Fully Convolutional Network. *Foods (Basel, Switzerland)*. 11(13).

Cui, R., Hua, W., Qu, K., Yang, H., Tong, Y., Li, Q., Wang, H., Ma, Y., Liu, S., Lin, T., Zhang, J., Sun, J. and Liu, C. (2021). An Interpretable Early Dynamic Sequential Predictor for Sepsis-Induced Coagulopathy Progression in the Real-World Using Machine Learning. *Frontiers in medicine*. 8, 775047.

de Lima, C. L., da Silva, C. C., da Silva, A. C. G., Luiz Silva, E., Marques, G. S., de Araújo, L. J. B., Albuquerque Júnior, L. A., de Souza, S. B. J., de Santana, M. A., Gomes, J. C., de Freitas Barbosa, V. A., Musah, A., Kostkova, P., Dos Santos, W. P. and da Silva Filho, A. G. (2020). COVID-SGIS: A Smart Tool for Dynamic Monitoring and Temporal Forecasting of Covid-19. *Frontiers in public health*. 8, 580815.

De Salazar, P. M., Lu, F., Hay, J. A., Gómez-Barroso, D., Fernández-Navarro, P., Martínez, E. V., Astray-Mochales, J., Amillategui, R., García-Fulgueiras, A., Chirlaque, M. D., Sánchez-Migallón, A., Larrauri, A., Sierra, M. J., Lipsitch, M., Simón, F., Santillana, M. and Hernán, M. A. (2022). Near real-time surveillance of the SARS-CoV-2 epidemic with incomplete data. *PLoS computational biology*. 18(3), e1009964.

Deng, L., Guo, S., Yin, J., Zeng, Y. and Chen, K. (2022). Multi-objective optimization of water resources allocation in Han River basin (China) integrating efficiency, equity and sustainability. *Scientific Reports*. 12(1), 798.

Deng, Z., Zhang, J., Li, J. and Zhang, X. (2021). Application of Deep Learning in Plant-Microbiota Association Analysis. *Frontiers in genetics*. 12, 697090.

Deressa, C. T. and Duressa, G. F. (2021). Analysis of Atangana–Baleanu fractional-order SEAIR epidemic model with optimal control. *Advances in Difference Equations*. 2021(1), 1–25.

Dervishi, A. (2024a). A multimodal stacked ensemble model for cardiac output prediction utilizing cardiorespiratory interactions during general anesthesia. *Scientific Reports*. 14(1), 7478.

Dervishi, A. J. S. R. (2024b). A multimodal stacked ensemble model for cardiac output prediction utilizing cardiorespiratory interactions during general anesthesia. 14(1), 7478.

Diaz Perez, F. J., Chinarro, D., Otin, R. P., Martín, R. D., Diaz, M. and Mouhaffel, A. G. (2020). Comparison of Growth Patterns of COVID-19 Cases through the ARIMA and Gompertz Models. Case Studies: Austria, Switzerland, and Israel. *Rambam Maimonides medical journal*. 11(3).

Didier, A. J., Nigro, A., Noori, Z., Omballi, M. A., Pappada, S. M. and Hamouda, D. M. (2024). Application of machine learning for lung cancer survival prognostication-A systematic review and meta-analysis. *Frontiers in artificial intelligence*. 7, 1365777.

Ding, D. and Zhang, R. (2022). China's COVID-19 Control Strategy and Its Impact on the Global Pandemic. *Frontiers in public health*. 10, 857003.

Diraco, G., Rescio, G., Siciliano, P. and Leone, A. (2023). Review on human action recognition in smart living: Sensing technology, multimodality, real-time processing, interoperability, and resource-constrained processing. *Sensors*. 23(11), 5281.

Dissanayake, S., Krishna, R., Pathirana, P. N., Horne, M. K., Szmulewicz, D. J. and Corben, L. A. (2023). A Bayesian Network Approach for Friedreich Ataxia Severity Classification using Probability Modelling. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2023, 1–4.

Dixon, S., Keshavamurthy, R., Farber, D. H., Stevens, A., Pazdernik, K. T. and Charles, L. E. (2022). A Comparison of Infectious Disease Forecasting Methods across Locations, Diseases, and Time. *Pathogens (Basel, Switzerland)*. 11(2).

Du, H., Xu, Q., Jiang, L., Bu, Y., Li, W. and Yan, J. (2024). Stepwise Identification Method of Thermal Load for Box Structure Based on Deep Learning. *Materials (Basel, Switzerland)*. 17(2).

Du, H., Yu, M., Xue, H., Lu, X., Chang, Y. and Li, Z. (2022). Association between sarcopenia and cognitive function in older Chinese adults: Evidence from the China health and retirement longitudinal study. *Frontiers in public health*. 10, 1078304.

Du, M., Huang, X., Li, S., Xu, L., Yan, B., Zhang, Y., Wang, H. and Liu, X. (2020). A Nomogram Model to Predict Malignant Cerebral Edema in Ischemic Stroke

Patients Treated with Endovascular Thrombectomy: An Observational Study. *Neuropsychiatric disease and treatment*. 16, 2913–2920.

Efimov, D. and Ushirobira, R. (2021). On an interval prediction of COVID-19 development based on a SEIR epidemic model. *Annual reviews in control*. 51, 477–487.

El Taha, L., Beyrouthy, C., Tamim, H. and Ghazeeri, G. (2022). Knowledge and attitudes among Lebanese pregnant women and women seeking fertility treatment during the COVID-19 outbreak: a cross-sectional survey. *BMJ open*. 12(3), e057873.

Entezari, A., Liu, N. C., Zhang, Z., Fang, J., Wu, C., Wan, B., Swain, M. and Li, Q. (2023). Nondeterministic multiobjective optimization of 3D printed ceramic tissue scaffolds. *Journal of the mechanical behavior of biomedical materials*. 138, 105580.

Estiri, H., Strasser, Z. H., Klann, J. G., Naseri, P., Wagholikar, K. B. and Murphy, S. N. (2021a). Predicting COVID-19 mortality with electronic medical records. *NPJ digital medicine*. 4(1), 1–10.

Estiri, H., Strasser, Z. H., Klann, J. G., Naseri, P., Wagholikar, K. B. and Murphy, S. N. (2021b). Predicting COVID-19 mortality with electronic medical records. *NPJ digital medicine*.

Feng, H. and Zhang, X. (2023). A novel encoder-decoder model based on Autoformer for air quality index prediction. *PloS one*. 18(4), e0284293.

Feng, Y., Hu, J., Duan, R. and Chen, Z. (2021). Credibility Assessment Method of Sensor Data Based on Multi-Source Heterogeneous Information Fusion. *Sensors (Basel, Switzerland)*. 21(7).

Ferrández, M., Ivorra, B., Ortigosa, P., Ramos, A. and Redondo, J. (2019). Application of the Be-CoDiS model to the 2018-19 Ebola Virus Disease outbreak in the Democratic Republic of Congo. *ResearchGate preprint*. 23, 1–17.

Ferrández, M. R., Ivorra, B., Redondo, J. L., Ramos, M. and Ortigosa, P. M. (2023). A multi-objective approach to identify parameters of compartmental epidemiological models—Application to Ebola Virus Disease epidemics.

Communications in Nonlinear Science and Numerical Simulation. 120(6), 107165.

Fialho, B. C., Gauss, L., Soares, P. F., Medeiros, M. Z. and Lacerda, D. P. (2023). Vaccine Innovation Meta-Model for Pandemic Contexts. *Journal of pharmaceutical innovation*, 1–49.

Ganaie, M. A. and Hu, M. (2021). Ensemble deep learning: A review. *arXiv preprint arXiv:2104.02395*.

Gao, Q., Shang, W. P. and Jing, M. X. (2022). Effect of Nucleic Acid Screening Measures on COVID-19 Transmission in Cities of Different Scales and Assessment of Related Testing Resource Demands-Evidence from China. *International journal of environmental research and public health*. 19(20).

Ghafouri-Fard, S., Mohammad-Rahimi, H., Motie, P., Minabi, M. A. S., Taheri, M. and Nateghinia, S. (2021). Application of machine learning in the prediction of COVID-19 daily new cases: A scoping review. *Heliyon*. 7(10), e08143.

Gong, H., Wang, M., Zhang, H., Elahe, M. F. and Jin, M. (2022). An Explainable AI Approach for the Rapid Diagnosis of COVID-19 Using Ensemble Learning Algorithms. *Frontiers in public health*. 10, 874455.

Goyal, S. and Singh, R. (2021). Detection and classification of lung diseases for pneumonia and Covid-19 using machine and deep learning techniques. *Journal of Ambient Intelligence and Humanized Computing*, 1–21.

Guadiana-Alvarez, J. L., Hussain, F., Morales-Menendez, R., Rojas-Flores, E., García-Zendejas, A., Escobar, C. A., Ramírez-Mendoza, R. A. and Wang, J. (2022). Prognosis patients with COVID-19 using deep learning. *BMC Medical Informatics and Decision Making*. 22(1), 1–18.

Guan, J., Zhao, Y., Wei, Y., Shen, S., You, D., Zhang, R., Lange, T. and Chen, F. (2022). Transmission dynamics model and the coronavirus disease 2019 epidemic: applications and challenges. *Medical Review*. 2(1), 89–109.

Gul, S., Khan, M. S. and Ur-Rehman, A. (2024). Triple-0: Zero-shot denoising and dereverberation on an end-to-end frozen anechoic speech separation network. *PloS one*. 19(7), e0301692.

Guo, Y., Liu, X., Wang, X., Zhu, T. and Zhan, W. (2022). Automatic Decision-Making Style Recognition Method Using Kinect Technology. *Frontiers in psychology*. 13, 751914.

Gupta, A., Jain, V. and Singh, A. (2022). Stacking ensemble-based intelligent machine learning model for predicting post-COVID-19 complications. *New Generation Computing*. 40(4), 987–1007.

Ha, K., Cho, S. and MacLachlan, D. (2005). Response models based on bagging neural networks. *Journal of Interactive Marketing*. 19(1), 17–30.

Hamer, W. H. (1906). *Epidemic disease in England: the evidence of variability and of persistency of type*. Bedford Press.

Hasan, M., Abedin, M. Z., Hajek, P., Coussement, K., Sultan, M. N. and Lucey, B. (2024a). A blending ensemble learning model for crude oil price forecasting. *Annals of Operations Research*. 120(6), 1–31.

Hasan, M., Abedin, M. Z., Hajek, P., Coussement, K., Sultan, M. N. and Lucey, B. (2024b). A blending ensemble learning model for crude oil price forecasting. *Annals of Operations Research*. 44(5), 1–31.

Hasan, M., Marjan, M. A., Uddin, M. P., Afjal, M., Kardy, S., Ma, S. and Nam, Y. (2023). Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. *Frontiers in plant science*. 14, 1234555.

Hasrod, T., Nuapia, Y. B. and Tutu, H. (2024). Comparison of individual and ensemble machine learning models for prediction of sulphate levels in untreated and treated Acid Mine Drainage. *Environmental monitoring and assessment*. 196(4), 332.

Hebbar, R., Papadopoulos, P., Reyes, R., Danvers, A. F., Polsinelli, A. J., Moseley, S. A., Sbarra, D. A., Mehl, M. R. and Narayanan, S. (2021). Deep multiple instance learning for foreground speech localization in ambient audio from wearable devices. *EURASIP journal on audio, speech, and music processing*. 2021(1), 7.

Hernández-Giottonini, K. Y., Arellano-Reynoso, B., Rodríguez-Córdova, R. J., de la Vega-Olivas, J., Díaz-Aparicio, E. and Lucero-Acuña, A. (2023). Enhancing

Therapeutic Efficacy against Brucella canis Infection in a Murine Model Using Rifampicin-Loaded PLGA Nanoparticles. *ACS omega*. 8(51), 49362–49371.

Hlongwane, R., Ramaboa, K. and Mongwe, W. (2024). Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. *PloS one*. 19(5), e0303566.

Hong, S., Wu, H., Xu, X. and Xiong, W. (2022). Early Warning of Enterprise Financial Risk Based on Decision Tree Algorithm. *Computational intelligence and neuroscience*. 2022, 9182099.

Hu, Y., Yang, Q., Zhang, J., Peng, Y., Guang, Q. and Li, K. (2023). Methods to predict osteonecrosis of femoral head after femoral neck fracture: a systematic review of the literature. *Journal of orthopaedic surgery and research*. 18(1), 377.

Hu, Z., Li, P. and Liu, Y. (2022). Enhancing the Performance of Evolutionary Algorithm by Differential Evolution for Optimizing Distillation Sequence. *Molecules (Basel, Switzerland)*. 27(12).

Huang, Y., Zhang, Z., Jiao, A., Ma, Y. and Cheng, R. (2024). A Comparative Visual Analytics Framework for Evaluating Evolutionary Processes in Multi-Objective Optimization. *IEEE transactions on visualization and computer graphics*. 30(1), 661–671.

Husnayain, A., Shim, E., Fuad, A. and Su, E. C.-Y. (2021). Predicting new daily COVID-19 cases and deaths using search engine query data in South Korea from 2020 to 2021: infodemiology study. *Journal of Medical Internet Research*. 23(12), e34178.

Inyang, U. I., Petrunin, I. and Jennions, I. (2023). Diagnosis of Multiple Faults in Rotating Machinery Using Ensemble Learning. *Sensors (Basel, Switzerland)*. 23(2).

Ismail, W. N., Alsalamah, H. A. and Mohamed, E. (2023). GA-Stacking: A New Stacking-Based Ensemble Learning Method to Forecast the COVID-19 Outbreak. *Computers, Materials and Continua*. 74(2).

Ivorra, B., Ferrández, M. R., Vela-Pérez, M. and Ramos, A. M. (2020). Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking

into account the undetected infections. The case of China. *Communications in nonlinear science and numerical simulation*. 88, 105303.

Jahanshahi, H., Munoz-Pacheco, J. M., Bekiros, S. and Alotaibi, N. D. (2021). A fractional-order SIRD model with time-dependent memory indexes for encompassing the multi-fractional characteristics of the COVID-19. *Chaos, Solitons Fractals*. 143, 110632.

Jarman, E. L., Alain, M., Conroy, N. and Omam, L. A. (2022). A case report of monkeypox as a result of conflict in the context of a measles campaign. *Public health in practice (Oxford, England)*. 4, 100312.

Ji, B., Pi, W., Liu, W., Liu, Y., Cui, Y., Zhang, X. and Peng, S. (2023). HyperVR: a hybrid deep ensemble learning approach for simultaneously predicting virulence factors and antibiotic resistance genes. *NAR genomics and bioinformatics*. 5(1), lqad012.

Jia, W. W., Zhu, F. Y. and Li, F. R. (2018). [Change pattern of heartwood of Larix olgensis plantation.]. *Ying yong sheng tai xue bao = The journal of applied ecology*. 29(7), 2277–2285.

Jiang, Y., Li, Q., Trevisan, G., Linhares, D. C. L. and MacKenzie, C. (2021). Investigating the relationship of porcine reproductive and respiratory syndrome virus RNA detection between adult/sow farm and wean-to-market age categories. *PloS one*. 16(7), e0253429.

Jiang, Z., Yang, S., Dong, S., Pang, Q., Smith, P., Abdalla, M., Zhang, J., Wang, G. and Xu, Y. (2023). Simulating soil salinity dynamics, cotton yield and evapotranspiration under drip irrigation by ensemble machine learning. *Frontiers in plant science*. 14, 1143462.

Jin, W., Dong, S., Yu, C. and Luo, Q. (2022). A data-driven hybrid ensemble AI model for COVID-19 infection forecast using multiple neural networks and reinforced learning. *Computers in Biology and Medicine*. 146, 105560.

Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Journal, Data Mining: The ASA Data Science*. 15(4), 531–538.

Kalule, R., Abderrahmane, H. A., Alameri, W. and Sassi, M. (2023). Stacked ensemble machine learning for porosity and absolute permeability prediction of carbonate rock plugs. *Scientific reports*. 13(1), 9855.

Kang, Z., Fan, R., Zhan, C., Wu, Y., Lin, Y., Li, K., Qing, R. and Xu, L. (2024). The Rapid Non-Destructive Differentiation of Different Varieties of Rice by Fluorescence Hyperspectral Technology Combined with Machine Learning. *Molecules (Basel, Switzerland)*. 29(3).

Karema, C., Wen, S., Sidibe, A., Smith, J. L., Gosling, R., Hakizimana, E., Tanner, M., Noor, A. M. and Tatarsky, A. (2020). History of malaria control in Rwanda: implications for future elimination in Rwanda and other malaria-endemic countries. *Malaria journal*. 19(1), 356.

Karlinsky, A. and Kobak, D. (2021). The World Mortality Dataset: Tracking excess mortality across countries during the COVID-19 pandemic. *medRxiv : the preprint server for health sciences*.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*. 115(772), 700–721.

Keya, A. J., Shajeeb, H. H., Rahman, M. S. and Mridha, M. F. (2023). FakeStack: Hierarchical Tri-BERT-CNN-LSTM stacked model for effective fake news detection. *PloS one*. 18(12), 294701.

Khan, I. U., Aslam, N., Aljabri, M., Aljameel, S. S., Kamaleldin, M. M. A., Alshamrani, F. M. and Chrouf, S. M. B. (2021). Computational Intelligence-Based Model for Mortality Rate Prediction in COVID-19 Patients. *International Journal of Environmental Research and Public Health*. 18(12), 20.

Khan, M. F. F., Dung, M. D. and Sakamura, K. (2023). Predicting COVID-19 Infected Cases: Exploring Stacked Generalization with Japanese Data. *International Conference On Systems Engineering*. 44(1), 59–68.

Khatun, D., Hossain, M. Y., Rahman, O. and Hossain, M. F. (2022). Estimation of life history parameters for river catfish Eutropiichthys vacha: insights from multi-models for sustainable management. *Heliyon*. 8(10), e10781.

Khoo, L. S., Lim, M. K., Chong, C. Y. and McNaney, R. (2024). Machine Learning for Multimodal Mental Health Detection: A Systematic Review of Passive Sensing Approaches. *Sensors (Basel, Switzerland)*. 24(2).

Kim, J. and Ahn, I. (2021). Infectious disease outbreak prediction using media articles with machine learning models. *Scientific reports*. 11(1), 4413.

Kozyrev, E. A., Ermakov, E. A., Boiko, A. S., Mednova, I. A., Kornetova, E. G., Bokhan, N. A. and Ivanova, S. A. (2023). Building Predictive Models for Schizophrenia Diagnosis with Peripheral Inflammatory Biomarkers. *Biomedicines*. 11(7).

Krause, C., Bergmann, E. and Schmidt, S. V. (2024). Epigenetic modulation of myeloid cell functions in HIV and SARS-CoV-2 infection. *Molecular biology reports*. 51(1), 342.

Kucharski, A. J., Russell, T. W., Diamond, C., Liu, Y., Edmunds, J., Funk, S., Eggo, R. M., Sun, F., Jit, M. and Munday, J. D. (2020). Early dynamics of transmission and control of COVID-19: a mathematical modelling study. *The lancet infectious diseases*. 20(5), 553–558. ISSN 1473-3099.

Kumar, R. L., Khan, F., Din, S., Band, S. S., Mosavi, A. and Ibeke, E. (2021). Recurrent neural network and reinforcement learning model for COVID-19 prediction. *Frontiers in public health*. 9.

Kumari, K., Jain, S. and Dhar, A. (2019). Computationally efficient approach for identification of fuzzy dynamic groundwater sampling network. *Environmental monitoring and assessment*. 191(5), 310.

Lange, O. (2023). Health economic evaluation of preventive digital public health interventions using decision-analytic modelling: a systematized review. *BMC health services research*. 23(1), 268.

Larsen, S. L. and Kraay, A. N. M. (2024). Transparent transmission models for informing public health policy: the role of trust and generalizability. *Proceedings. Biological sciences*. 291(2015), 20232273.

Lashley, F. R. and Durham, J. D. (2007). *Emerging infectious diseases: trends and issues*. Springer Publishing Company.

Le Fouest, S. and Mulleners, K. (2024). Optimal blade pitch control for enhanced vertical-axis wind turbine performance. *Nature communications*. 15(1), 2770.

Lee, J., Kim, J. N., Dallan, L. A. P., Zimin, V. N., Hoori, A., Hassani, N. S., Makhlouf, M. H. E., Guagliumi, G., Bezerra, H. G. and Wilson, D. L. (2024). Deep learning segmentation of fibrous cap in intravascular optical coherence tomography images. *Scientific reports*. 14(1), 4393.

Lenatti, M., Narteni, S., Paglialonga, A., Rampa, V. and Mongelli, M. (2023). Dual-View Single-Shot Multibox Detector at Urban Intersections: Settings and Performance Evaluation. *Sensors (Basel, Switzerland)*. 23(6).

Li, C. (2022). IGBT Fault Prediction Combining Terminal Characteristics and Artificial Intelligence Neural Network. *Computational and mathematical methods in medicine*. 2022, 7459354.

Li, G., Wang, Y., Zhang, C., Xu, C. and Zhan, L. (2024a). Study on the Impact of Building Energy Predictions Considering Weather Errors of Neighboring Weather Stations. *Sensors (Basel, Switzerland)*. 24(4).

Li, H., Yang, Y., Chen, J., Li, Q., Chen, Y., Zhang, Y., Cai, S., Zhan, M., Wu, C., Lin, X. and Xiang, J. (2024b). Epidemiological Characteristics of Overseas-Imported Infectious Diseases Identified through Airport Health-Screening Measures: A Case Study on Fuzhou, China. *Tropical medicine and infectious disease*. 9(6).

Li, J., An, L., Cheng, Y. and Wang, H. (2024c). Research on sound quality of roller chain transmission system based on multi-source transfer learning. *Scientific reports*. 14(1), 11226.

Li, S., Zhu, L., Zhang, L., Zhang, G., Ren, H. and Lu, L. (2023a). Urbanization-Related Environmental Factors and Hemorrhagic Fever with Renal Syndrome: A Review Based on Studies Taken in China. *International journal of environmental research and public health*. 20(4).

Li, W., Gong, J., Zhou, J., Zhang, L., Wang, D., Li, J., Shi, C. and Fan, H. (2021). An evaluation of COVID-19 transmission control in Wenzhou using a modified SEIR model. *Epidemiology Infection*. 149.

Li, X., Patel, V., Duan, L., Mikuliak, J., Basran, J. and Osgood, N. D. (2024d). Real-Time Epidemiology and Acute Care Need Monitoring and Forecasting for COVID-19 via Bayesian Sequential Monte Carlo-Leveraged Transmission Models. *International journal of environmental research and public health*. 21(2).

Li, X., Yu, Q., Yang, Y., Tang, C. and Wang, J. (2023b). An evolutionary ensemble model based on GA for epidemic transmission prediction. *Journal of Intelligent & Fuzzy Systems*. 44(5), 7469–7481.

Li, X., Yu, Q., Yang, Y., Tang, C. and Wang, J. (2023c). An evolutionary ensemble model based on GA for epidemic transmission prediction. *Journal of Intelligent & Fuzzy Systems*. 44(5), 7469–7481.

Li, X., Yu, Q., Yang, Y., Tang, C. and Wang, J. (2023d). An evolutionary ensemble model based on GA for epidemic transmission prediction. *Journal of Intelligent and Systems, Fuzzy*. 44(5), 7469–7481.

Li, Y., Meng, X., Zhang, Z. and Song, G. (2020). A Machining State-Based Approach to Tool Remaining Useful Life Adaptive Prediction. *Sensors (Basel, Switzerland)*. 20(23).

Li, Y., Wang, Y., Shen, Z., Miao, F., Wang, J., Sun, Y., Zhu, S., Zheng, Y. and Guan, S. (2022). A biodegradable magnesium alloy vascular stent structure: Design, optimisation and evaluation. *Acta biomaterialia*. 142, 402–412.

Liao, Y., Han, L., Wang, H. and Zhang, H. (2022). Prediction Models for Railway Track Geometry Degradation Using Machine Learning Methods: A Review. *Sensors (Basel, Switzerland)*. 22(19).

Liao, Z., Lan, P., Fan, X., Kelly, B., Innes, A. and Liao, Z. (2021). SIRVD-DL: A COVID-19 deep learning prediction model based on time-dependent SIRVD. *Computers in Biology and Medicine*. 138, 104868.

Lim, B. and Zohren, S. (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*. 379(2194), 20200209.

Lim, K. M., Lee, C. P., Lee, Z. Y. and Alqahtani, A. (2023). EnViTSA: Ensemble of Vision Transformer with SpecAugment for Acoustic Event Classification. *Sensors (Basel, Switzerland)*. 23(22).

Liu, C., Ruan, K. and Ma, X. (2023a). DMEformer: A newly designed dynamic model ensemble transformer for crude oil futures prediction. *Heliyon*. 9(6), e16715.

Liu, C. and Su, H. (2024). Prediction of martensite start temperature of steel combined with expert experience and machine learning. *Science and technology of advanced materials*. 25(1), 2354655.

Liu, M., Ning, J., Du, Y., Cao, J., Zhang, D., Wang, J. and Chen, M. (2020a). Modelling the evolution trajectory of COVID-19 in Wuhan, China: experience and suggestions. *Public health*. 183, 76–80.

Liu, S. and Ozay, M. (2023). Task guided representation learning using compositional models for zero-shot domain adaptation. *Neural networks : the official journal of the International Neural Network Society*. 165, 370–380.

Liu, T., Huang, J., He, Z., Zhang, Y., Yan, N., Zhang, C. J. P. and Ming, W.-K. (2023b). A real-world data validation of the value of early-stage SIR modelling to public health. *Scientific Reports*. 13(1), 9164.

Liu, T. and Ye, A. (2023). Domain knowledge-assisted multi-objective evolutionary algorithm for channel selection in brain-computer interface systems. *Frontiers in neuroscience*. 17, 1251968.

Liu, X., Tian, J., Duan, P., Yu, Q., Wang, G. and Wang, Y. (2024). GrMoNAS: A granularity-based multi-objective NAS framework for efficient medical diagnosis. *Computers in biology and medicine*. 171, 108118.

Liu, X.-X., Fong, S. J., Dey, N., Crespo, R. G. and Herrera-Viedma, E. (2021). A new SEAIRD pandemic prediction model with clinical and epidemiological data analysis on COVID-19 outbreak. *Applied Intelligence*. 51(7), 4162–4198.

Liu, Z., Huang, S., Lu, W., Su, Z., Yin, X., Liang, H. and Zhang, H. (2020b). Modeling the trend of coronavirus disease 2019 and restoration of operational capability of metropolitan medical service in China: a machine learning and mathematical model-based analysis. *Global health research and policy*. 5(1), 1–11.

Lou, J., Wang, B., Li, J., Ni, P., Jin, Y., Chen, S., Xi, Y., Zhang, R. and Duan, G. (2022). The CRISPR-Cas system as a tool for diagnosing and treating infectious diseases. *Molecular biology reports*. 49(12), 11301–11311.

Lu, S. Y., Zhang, Z., Zhang, Y. D. and Wang, S. H. (2021). CGENet: A Deep Graph Model for COVID-19 Detection Based on Chest CT. *Biology*. 11(1).

Lu, Z., Yu, Y., Chen, Y., Ren, G., Xu, C. and Wang, S. (2022). Stability analysis of a nonlocal SIHRDP epidemic model with memory effects. *Nonlinear dynamics*. 109(1), 121–141.

Lv, C. X., An, S. Y., Qiao, B. J. and Wu, W. (2021). Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. *BMC infectious diseases*. 21(1), 839.

Ma, P., Wang, S., Zhou, J., Li, T., Fan, X., Fan, J. and Wang, S. (2020). Meteorological rhythms of respiratory and circulatory diseases revealed by Harmonic Analysis. *Heliyon*. 6(5).

Ma, Z., Wang, B., Luo, W., Jiang, J., Liu, D., Wei, H. and Luo, H. (2024). Air pollutant prediction model based on transfer learning two-stage attention mechanism. *Scientific reports*. 14(1), 7385.

Madewell, Z. J., Yang, Y., Longini, I. M., Halloran, M. E., Vespignani, A. and Dean, N. E. (2023). Rapid review and meta-analysis of serial intervals for SARS-CoV-2 Delta and Omicron variants. *BMC infectious diseases*. 23(1), 429.

Mahajan, A., Sharma, N., Aparicio-Obregon, S., Alyami, H., Alharbi, A., Anand, D., Sharma, M. and Goyal, N. (2022a). A Novel Stacking-Based Deterministic Ensemble Model for Infectious Disease Prediction. *Mathematics*. 10(10), 1714.

Mahajan, A., Sharma, N., Aparicio-Obregon, S., Alyami, H., Alharbi, A., Anand, D., Sharma, M. and Goyal, N. (2022b). A novel stacking-based deterministic ensemble model for infectious disease prediction. *Mathematics*. 10(10), 1714.

Mahajan, A., Sharma, N., Aparicio-Obregon, S., Alyami, H., Alharbi, A., Anand, D., Sharma, M. and Goyal, N. (2022c). A novel stacking-based deterministic ensemble model for infectious disease prediction. *Mathematics*. 10(10), 1714.

Mahanty, C., Kumar, R., Mishra, B. K., Hemanth, D. J., Gupta, D. and Khanna, A. (2020). Prediction of COVID-19 active cases using exponential and non-linear growth models. *Expert Systems*.

Mandal, M., Jana, S., Nandi, S. K., Khatua, A., Adak, S. and Kar, T. (2020). A model based study on the dynamics of COVID-19: Prediction and control. *Chaos, Solitons Fractals*. 136, 109889.

Mathieu, E., Ritchie, H., Rodés-Guirao, L., Appel, C., Giattino, C., Hasell, J., Macdonald, B., Dattani, S., Beltekian, D. and Ortiz-Ospina, E. (2020). Coronavirus pandemic (COVID-19). *Our world in data*.

McConnell, J. R., Wilson, A. I., Stohl, A., Arienzo, M. M., Chellman, N. J., Eckhardt, S., Thompson, E. M., Pollard, A. M. and Steffensen, J. P. (2018). Lead pollution recorded in Greenland ice indicates European emissions tracked plagues, wars, and imperial expansion during antiquity. *Proceedings of the National Academy of Sciences of the United States of America*. 115(22), 5726–5731.

Meem, A. T., Khan, M. M., Masud, M. and Aljahdali, S. (2022). Prediction of covid-19 based on chest x-ray images using deep learning with CNN. *Computer Systems Science and Engineering*, 1223–1240.

Meraihi, Y., Gabis, A. B., Mirjalili, S., Ramdane-Cherif, A. and Alsaadi, F. E. (2022). Machine Learning-Based Research for COVID-19 Detection, Diagnosis, and Prediction: A Survey. *SN Computer Science*. 3(4), 1–35.

Milligan, W. R., Fuller, Z. L., Agarwal, I., Eisen, M. B., Przeworski, M. and Sella, G. (2021). Impact of essential workers in the context of social distancing for epidemic control. *PLoS One*. 16(8), e0255680.

Minor, N. R., Ramuta, M. D., Stauss, M. R., Harwood, O. E., Brakefield, S. F., Alberts, A., Vuyk, W. C., Bobholz, M. J., Rosinski, J. R., Wolf, S. *et al.* (2023). Metagenomic sequencing detects human respiratory and enteric viruses in air samples collected from congregate settings. *Scientific reports*. 13(1), 21398.

Mirzania, M., Shakibazadeh, E. and Ashoorkhani, M. (2022). Challenges for implementation of inter-sectoral efforts to improve outbreak response using consolidated framework for implementation research; Iran's COVID-19 experience. *BMC health services research*. 22(1), 1118.

Mohammed, S., Sha'aban, Y. A., Umoh, I. J., Salawudeen, A. T. and Ibn Shamsah, S. M. (2023). A hybrid smell agent symbiosis organism search algorithm for optimal control of microgrid operations. *PloS one*. 18(6), e0286695.

Muñoz, L., Alonso-García, M., Villarreal, V., Hernández, G., Nielsen, M., Pinto-Santos, F., Saavedra, A., Areiza, M., Montenegro, J. and Sittón-Candanedo, I. (2022). A Hybrid System For Pandemic Evolution Prediction. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*. 11(1), 111–128.

Nair, A., Ahirwar, A., Singh, S., Lodhi, R., Lodhi, A., Rai, A., Jadhav, D. A., Harish, Varjani, S., Singh, G., Marchand, J., Schoefs, B. and Vinayak, V. (2023). Astaxanthin as a King of Ketocarotenoids: Structure, Synthesis, Accumulation, Bioavailability and Antioxidant Properties. *Marine drugs*. 21(3).

Narkhede, P., Walambe, R., Poddar, S. and Kotecha, K. (2021). Incremental learning of LSTM framework for sensor fusion in attitude estimation. *PeerJ. Computer science*. 7, e662.

Nath, A. and Sahu, G. K. (2019). Exploiting ensemble learning to improve prediction of phospholipidosis inducing potential. *Journal of theoretical biology*. 479, 37–47.

Nguyen, R., Sokhansanj, B. A., Polikar, R. and Rosen, G. L. (2023). Complet+: a computationally scalable method to improve completeness of large-scale protein sequence clustering. *PeerJ*. 11, e14779.

Nisar, K. S., Ahmad, S., Ullah, A., Shah, K., Alrabaiah, H. and Arfan, M. (2021). Mathematical analysis of SIRD model of COVID-19 with Caputo fractional derivative based on real data. *Results in Physics*. 21, 103772.

Noordman, C. R., Yakar, D., Bosma, J., Simonis, F. F. J. and Huisman, H. (2023). Complexities of deep learning-based undersampled MR image reconstruction. *European radiology experimental*. 7(1), 58.

Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M. and Qasem, S. N. (2024). Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases. *Diagnostics (Basel, Switzerland)*. 14(2).

Oka, M. (2021). Interpreting a standardized and normalized measure of neighborhood socioeconomic status for a better understanding of health differences. *Archives of Public Health*. 79(1), 226.

Olum, R., Ahaisibwe, B., Atuhairwe, I., Balizzakiwa, T., Kizito, P., Apiyo, M., Kalanzi, J., Nabawanuka, A., Bahatungire, R. and Kerry, V. (2024). Readiness To Manage Ebola Virus Disease Among Emergency Healthcare Workers in Uganda: A Nationwide Multicenter Survey. *Research square*.

Ortiz-Barrios, M., Petrillo, A., Arias-Fonseca, S., McClean, S., de Felice, F., Nugent, C. and Uribe-López, S. A. (2024). An AI-based multiphase framework for improving the mechanical ventilation availability in emergency departments during respiratory disease seasons: a case study. *International journal of emergency medicine*. 17(1), 45.

Pacheco, C. and de Lacerda, C. (2021). Function estimation and regularization in the SIRD model applied to the COVID-19 pandemics. *Inverse Problems in Science and Engineering*. 29(11), 1613–1628.

Padhi, A., Pradhan, S., Sahoo, P. P., Suresh, K., Behera, B. K. and Panigrahi, P. K. (2020). Studying the effect of lockdown using epidemiological modelling of COVID-19 and a quantum computational approach using the Ising spin interaction. *Scientific reports*. 10(1), 1–14.

Padilla-García, E. A., Cervantes-Culebro, H., Rodriguez-Angeles, A. and Cruz-Villar, C. A. (2023). Selection/control concurrent optimization of BLDC motors for industrial robots. *PloS one*. 18(8), e0289717.

Pant, D., Pokharel, S., Mandal, S., Kc, D. B. and Pati, R. (2023). DFT-aided machine learning-based discovery of magnetism in Fe-based bimetallic chalcogenides. *Scientific reports*. 13(1), 3277.

Papafotis, K., Nikitas, D. and Sotiriadis, P. P. (2021). Magnetic Field Sensors' Calibration: Algorithms' Overview and Comparison. *Sensors (Basel, Switzerland)*. 21(16).

Peng, X., Xu, H., Liu, J., Wang, J. and He, C. (2023). Voice disorder classification using convolutional neural network based on deep transfer learning. *Scientific reports*. 13(1), 7264.

Perramon-Malavez, A., Bravo, M., de Rioja, V. L., Català, M., Alonso, S., Álvarez Lacalle, E., López, D., Soriano-Arandes, A. and Prats, C. (2023). A semi-empirical risk panel to monitor epidemics: multi-faceted tool to assist

healthcare and public health professionals. *Frontiers in public health*. 11, 1307425.

Pinter, G., Felde, I., Mosavi, A., Ghamisi, P. and Gloaguen, R. (2020). COVID-19 pandemic prediction for Hungary; a hybrid machine learning approach. *Mathematics*. 8(6), 890.

Piscitelli, P. and Miani, A. (2024). Climate Change and Infectious Diseases: Navigating the Intersection through Innovation and Interdisciplinary Approaches. *International journal of environmental research and public health*. 21(3).

Pishgar, M., Harford, S., Theis, J., Galanter, W., Rodriguez-Fernandez, J. M., Chaisson, L. H., Zhang, Y., Trotter, A., Kochendorfer, K. M., Boppana, A. and Darabi, H. (2022). A process mining- deep learning approach to predict survival in a cohort of hospitalized COVID-19 patients. *Bmc Medical Informatics and Decision Making*. 22(1).

Popescu, S. and Myers, N. (2021). Interdisciplinary Information for Infectious Disease Response: Exercising for Improved Medical/Public Health Communication and Collaboration. *Disaster medicine and public health preparedness*. 15(5), 546–550.

Postnikov, E. B. (2020). Estimation of COVID-19 dynamics "on a back-of-envelope": Does the simplest SIR model provide quantitative parameters and predictions? *Chaos, Solitons Fractals*. 135, 109841.

Pradines, B. and Rogier, C. (2018). Contribution of the French army health service in support of expertise and research in infectiology in Africa. *New microbes and new infections*. 26, S78–S82.

Qin, C., Wang, X., Xu, G. and Ma, X. (2022). Advances in Cuffless Continuous Blood Pressure Monitoring Technology Based on PPG Signals. *BioMed research international*. 2022, 8094351.

Qiu, H., Luo, L., Su, Z., Zhou, L., Wang, L. and Chen, Y. (2020). Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure. *BMC medical informatics and decision making*. 20(1), 83.

Ramanuja, E., Santhiya, C. and Padmavathi, S. (2022). Day-Level Forecasting of COVID-19 Transmission in India Using Variants of Supervised LSTM Models: Modeling and Recommendations. *Journal of Information Technology Research (JITR)*. 15(1), 1–14.

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B. and Yang, G.-Z. (2016). Deep learning for health informatics. *IEEE journal of biomedical and health informatics*. 21(1), 4–21.

Ren, J. (2021). Pop Music Trend and Image Analysis Based on Big Data Technology. *Computational intelligence and neuroscience*. 2021, 4700630.

Roda, W. C., Varughese, M. B., Han, D. and Li, M. Y. (2020). Why is it difficult to accurately predict the COVID-19 epidemic? *Infectious disease modelling*. 5, 271–281.

Rodrigues, J., Studer, E., Streuber, S., Meyer, N. and Sandi, C. (2020). Locomotion in virtual environments predicts cardiovascular responsiveness to subsequent stressful challenges. *Nature communications*. 11(1), 5904.

Rodrigues, P. M., Madeiro, J. P. and Marques, J. A. L. (2023). Enhancing Health and Public Health through Machine Learning: Decision Support for Smarter Choices. *Bioengineering (Basel, Switzerland)*. 10(7).

Ruan, X., Du, P., Zhao, K., Huang, J., Xia, H., Dai, D., Huang, S., Cui, X., Liu, L. and Zhang, J. (2020). Mechanism of Dayuanyin in the treatment of coronavirus disease 2019 based on network pharmacology and molecular docking. *Chinese medicine*. 15, 62.

Ruan, Y., Huang, T., Zhou, W., Zhu, J., Liang, Q., Zhong, L., Tang, X., Liu, L., Chen, S. and Xie, Y. (2023). The lead time and geographical variations of Baidu Search Index in the early warning of COVID-19. *Scientific reports*. 13(1), 14705.

Saba Raoof, S. and Durai, M. S. (2022). A comprehensive review on smart health care: applications, paradigms, and challenges with case studies. *Contrast Media & Molecular Imaging*. 2022(1), 4822235.

Sah, S., Surendiran, B., Dhanalakshmi, R., Mohanty, S. N., Alenezi, F. and Polat, K. (2022). Forecasting COVID-19 Pandemic Using Prophet, ARIMA, and Hybrid

Stacked LSTM-GRU Models in India. *Computational and Mathematical Methods in Medicine*. 2022.

Sahai, A. K., Rath, N., Sood, V. and Singh, M. P. (2020). ARIMA modelling forecasting of COVID-19 in top five affected countries. *Diabetes and Metabolic Syndrome Clinical Research and Reviews*. 14(5).

Sangphukieo, A., Laomettachit, T. and Ruengjitchatchawalya, M. (2020). Photosynthetic protein classification using genome neighborhood-based machine learning feature. *Scientific reports*. 10(1), 7108.

Santra, A. and Dutta, A. (2022). A Comprehensive Review of Machine Learning Techniques for Predicting the Outbreak of Covid-19 Cases.

Saqib, M. (2021). Forecasting COVID-19 outbreak progression using hybrid polynomial-Bayesian ridge regression model. *Applied Intelligence*. 51(5), 2703–2713.

Sassano, M., Mariani, M., Quaranta, G., Pastorino, R. and Boccia, S. (2022). Polygenic risk prediction models for colorectal cancer: a systematic review. *BMC cancer*. 22(1), 65.

Sayed, S. A.-F., Elkorany, A. M. and Mohammad, S. S. (2021). Applying different machine learning techniques for prediction of COVID-19 severity. *Ieee Access*. 9, 135697–135707.

Seal, S., Yang, H., Trapotsi, M. A., Singh, S., Carreras-Puigvert, J., Spjuth, O. and Bender, A. (2023). Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data. *Journal of cheminformatics*. 15(1), 56.

Sharma, S., Alsmadi, I., Alkhawaldeh, R. S. and Al-Ahmad, B. (2022). Data-driven analysis and predictive modeling on COVID-19. *Concurrency and Practice, Computation: and Experience*. 34(28), e7390.

Shashvat, K., Basu, R., Bhondekar, P. and Kaur, A. (2019). An ensemble model for forecasting infectious diseases in India. *Tropical Biomed.* 36(4), 822–832.

Shen, J., Wang, S., Sun, H., Huang, J., Bai, L., Wang, X., Dong, Y. and Tang, Z. (2024a). A novel non-negative Bayesian stacking modeling method for Cancer

survival prediction using high-dimensional omics data. *BMC Medical Research Methodology*. 24(1), 105.

Shen, J., Wang, S., Sun, H., Huang, J., Bai, L., Wang, X., Dong, Y. and Tang, Z. (2024b). A novel non-negative Bayesian stacking modeling method for Cancer survival prediction using high-dimensional omics data. *BMC Medical Research Methodology*. 24(1), 105.

Shen, X. and Li, X. (2024). Deep-learning methods for unveiling large-scale single-cell transcriptomes. *Cancer biology medicine*. 20(12), 972–80.

Shin, H. G., Choi, Y. H. and Yoon, C. P. (2021). Movement Path Data Generation from Wi-Fi Fingerprints for Recurrent Neural Networks. *Sensors (Basel, Switzerland)*. 21(8).

Shulgin, B., Stone, L. and Agur, Z. (1998). Pulse vaccination strategy in the SIR epidemic model. *Bulletin of mathematical biology*. 60(6), 1123–1148.

Shyaa, M. A., Zainol, Z., Abdullah, R., Anbar, M., Alzubaidi, L. and Santamaría, J. (2023). Enhanced Intrusion Detection with Data Stream Classification and Concept Drift Guided by the Incremental Learning Genetic Programming Combiner. *Sensors (Basel, Switzerland)*. 23(7).

Silvestri, S., Islam, S., Papastergiou, S., Tzagkarakis, C. and Ciampi, M. (2023). A Machine Learning Approach for the NLP-Based Analysis of Cyber Threats and Vulnerabilities of the Healthcare Ecosystem. *Sensors (Basel, Switzerland)*. 23(2).

Singh, S., Sundram, B. M., Rajendran, K., Law, K. B. and Gill, B. S. (2020). Coronavirus Pandemic Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models. *The Journal of Infection in Developing Countries*. 14(9), 971–976.

Sivaraman, N. K., Gaur, M., Baijal, S., Muthiah, S. B. and Sheth, A. (2022). Exo-SIR: An Epidemiological Model to Analyze the Impact of Exogenous Spread of Infection. *International Journal of Data Science and Analytics*, 1–16.

Soliman, M., Lyubchich, V. and Gel, Y. R. (2019). Complementing the power of deep learning with statistical model fusion: Probabilistic forecasting of influenza in Dallas County, Texas, USA. *Epidemics*. 28, 100345.

Song, Z., Jia, X., Bao, J., Yang, Y., Zhu, H. and Shi, X. (2021). Spatio-Temporal Analysis of Influenza-Like Illness and Prediction of Incidence in High-Risk Regions in the United States from 2011 to 2020. *International journal of environmental research and public health*. 18(13).

Soni, M., Khan, I. R., Basir, S., Chadha, R., Alguno, A. C. and Bhowmik, T. (2022). Light-Weighted Deep Learning Model to Detect Fault in IoT-Based Industrial Equipment. *Computational intelligence and neuroscience*. 2022, 2455259.

Su, L., Hong, N., Zhou, X., He, J., Ma, Y., Jiang, H., Han, L., Chang, F., Shan, G. and Zhu, W. (2020). Evaluation of the secondary transmission pattern and epidemic prediction of COVID-19 in the four metropolitan areas of China. *Frontiers in medicine*. 7, 171.

Sun, J. (2023). A multi-objective optimization based doherty power amplifier and its matching network optimization method. *PloS one*. 18(12), e0293371.

Sun, Z., Yuan, Y., Xiong, X., Meng, S., Shi, Y. and Chen, A. (2024). Predicting academic achievement from the collaborative influences of executive function, physical fitness, and demographic factors among primary school students in China: ensemble learning methods. *BMC public health*. 24(1), 274.

Swaraj, A., Verma, K., Kaur, A., Singh, G. and Sales, L. (2021). Implementation of Stacking Based ARIMA Model for Prediction of Covid-19 Cases in India. *Journal of Biomedical Informatics*. 121(9), 103887.

Tan, C. V., Singh, S., Lai, C. H., Zamri, A., Dass, S. C., Aris, T. B., Ibrahim, H. M. and Gill, B. S. (2022). Forecasting COVID-19 Case Trends Using SARIMA Models during the Third Wave of COVID-19 in Malaysia. *International journal of environmental research and public health*. 19(3).

Tan, L., Wang, H., Yang, C. and Niu, B. J. N. c. (2017). A multi-objective optimization method based on discrete bacterial algorithm for environmental/economic power dispatch. 16, 549–565.

Tao, Z., Bing-Qiang, H., Huiling, L., Hongbin, S., Pengfei, Y. and Hongsheng, D. (2021). 18F-FDG-PET/CT Whole-Body Imaging Lung Tumor Diagnostic Model: An Ensemble E-ResNet-NRC with Divided Sample Space. *BioMed research international*. 2021, 8865237.

Tian, S., Sun, S., Mao, W., Qian, S., Zhang, L., Zhang, G., Xu, B. and Chen, M. (2021). Development and Validation of Prognostic Nomogram for Young Patients with Kidney Cancer. *International journal of general medicine*. 14, 5091–5103.

Tripto, N. I., Kabir, M., Bayzid, M. S. and Rahman, A. (2020). Evaluation of classification and forecasting methods on time series gene expression data. *PloS one*. 15(11), e0241686.

Tsai, Y., Baldwin, S. A. and Gopaluni, B. (2021). Identifying indicator species in ecological habitats using Deep Optimal Feature Learning. *PloS one*. 16(9), e0256782.

van de Pas, R., Mans, L. and Koutsoumpa, M. (2023). An exploratory review of investments by development actors in health workforce programmes and job creation. *Human Resources for Health*. 21(1), 54.

Vobugari, N., Raja, V., Sethi, U., Gandhi, K., Raja, K. and Surani, S. R. (2022). Advancements in Oncology with Artificial Intelligence-A Review Article. *Cancers*. 14(5).

Vukašinović, A., Klisic, A., Ostanek, B., Kafedžić, S., Zdravković, M., Ilić, I., Sopić, M., Hinić, S., Stefanović, M., Bogavac-Stanojević, N., Marc, J., Nešković, A. N. and Kotur-Stevuljević, J. (2023). Redox Status and Telomere-Telomerase System Biomarkers in Patients with Acute Myocardial Infarction Using a Principal Component Analysis: Is There a Link? *International journal of molecular sciences*. 24(18).

Wang, H., Tao, G., Ma, J., Jia, S., Chi, L., Yang, H., Zhao, Z. and Tao, J. (2022a). 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.

Wang, J., Zhang, H., Chen, N., Zeng, T., Ai, X. and Wu, K. (2023a). PorcineAI-Enhancer: Prediction of Pig Enhancer Sequences Using Convolutional Neural Networks. *Animals : an open access journal from MDPI*. 13(18).

Wang, L., Liu, Y., Chen, H., Qiu, S., Liu, Y., Yang, M., Du, X., Li, Z., Hao, R., Tian, H., Song, . and H. (2023b). Search-engine-based surveillance using artificial intelligence for early detection of coronavirus disease outbreak. *Journal of Big Data*. 10(1), 169.

Wang, Y., Cao, Z., Zeng, D., Wang, X. and Wang, Q. (2020a). Using deep learning to predict the hand-foot-and-mouth disease of enterovirus A71 subtype in Beijing from 2011 to 2018. *Scientific reports*. 10(1), 12201.

Wang, Y., Xu, C., Li, Y., Wu, W., Gui, L., Ren, J. and Yao, S. (2020b). An Advanced Data-Driven Hybrid Model of SARIMA-NNNAR for Tuberculosis Incidence Time Series Forecasting in Qinghai Province, China. *Infection and drug resistance*. 13, 867–880.

Wang, Y., Yan, Z., Wang, D., Yang, M., Li, Z., Gong, X., Wu, D., Zhai, L., Zhang, W. and Wang, Y. (2022b). Prediction and analysis of COVID-19 daily new cases and cumulative cases: times series forecasting and machine learning models. *BMC Infectious Diseases*. 22(1), 1–12.

Wang, Y., You, L., Chyr, J., Lan, L., Zhao, W., Zhou, Y., Xu, H., Noble, P. and Zhou, X. (2020c). Causal Discovery in Radiographic Markers of Knee Osteoarthritis and Prediction for Knee Osteoarthritis Severity With Attention-Long Short-Term Memory. *Frontiers in public health*. 8, 604654.

Wang, Y. C., Houng, Y. C., Chen, H. X. and Tseng, S. M. (2023c). Network Anomaly Intrusion Detection Based on Deep Learning Approach. *Sensors (Basel, Switzerland)*. 23(4).

Wang, Z., Zhao, C. and Zhang, W. (2023d). Multi-objective design and optimization of squeezed branch pile based on orthogonal test. *Scientific reports*. 13(1), 22508.

Wangping, J., Ke, H., Yang, S., Wenzhe, C., Shengshu, W., Shanshan, Y., Jianwei, W., Fuyin, K., Penggang, T. and Jing, L. (2020). Extended SIR prediction of the epidemics trend of COVID-19 in Italy and compared with Hunan, China. *Frontiers in medicine*. 7, 169.

West, R. M. J. A. o. C. B. (2022). Best practice in statistics: The use of log transformation. 59(3), 162–165.

White, L., Basurra, S., Alsewari, A. A., Saeed, F. and Addanki, S. M. (2024). Temporal meta-optimiser based sensitivity analysis (TMSA) for agent-based models and applications in children's services. *Scientific reports*. 14(1), 9105.

Wu, C., Zheng, P., Xu, X., Chen, S., Wang, N. and Hu, S. (2020). Discovery of the Environmental Factors Affecting Urban Dwellers' Mental Health: A Data-

Driven Approach. *International journal of environmental research and public health*. 17(21).

Xia, Z., Qin, L., Ning, Z. and Zhang, X. (2022). Deep learning time series prediction models in surveillance data of hepatitis incidence in China. *PloS one*. 17(4), e0265660.

Xie, M., Lin, S., Dong, K. and Zhang, S. (2023). Short-Term Prediction of Multi-Energy Loads Based on Copula Correlation Analysis and Model Fusions. *Entropy (Basel, Switzerland)*. 25(9).

Xu, C., Li, H., Yang, J., Peng, Y., Cai, H., Zhou, J., Gu, W. and Chen, L. (2023). Interpretable prediction of 3-year all-cause mortality in patients with chronic heart failure based on machine learning. *BMC medical informatics and decision making*. 23(1), 267.

Xu, J., Shao, Z., Jia, S., Sha, J., Li, J., Gao, F., Shi, X., Wang, J., Jin, C., Jiang, M., Tian, H., Cao, J., Pu, H., Xu, L. and Lu, L. (2024). A comprehensive stem cell laboratory module with blended learning for medical students at Tongji University. *Biochemistry and molecular biology education : a bimonthly publication of the International Union of Biochemistry and Molecular Biology*. 52(3), 291–298.

Yang, J., Wang, Y. and Li, X. (2022a). Prediction of stock price direction using the LASSO-LSTM model combines technical indicators and financial sentiment analysis. *PeerJ. Computer science*. 8, e1148.

Yang, M., Shi, L., Chen, H., Wang, X., Jiao, J., Liu, M., Yang, J. and Sun, G. (2022b). Critical policies disparity of the first and second waves of COVID-19 in the United Kingdom. *International journal for equity in health*. 21(1), 115.

Yang, Q., Wang, J., Ma, H. and Wang, X. (2020a). Research on COVID-19 based on ARIMA model-Taking Hubei, China as an example to see the epidemic in Italy. *Journal of Infection and Public Health*. 13(10), 1415–1418.

Yang, S., Cui, L., Wang, L., Wang, T. and You, J. (2024). Enhancing multimodal depression diagnosis through representation learning and knowledge transfer. *Heliyon*. 10(4), 25959.

Yang, X., Zhao, P., Dong, Y., Shen, X., Shen, H., Li, J., Jiang, G., Wang, W., Dai, H., Dong, J., Gao, S. and Si, X. (2020b). An improved recombinase polymerase amplification assay for visual detection of Vibrio parahaemolyticus with lateral flow strips. *Journal of food science*. 85(6), 1834–1844.

Yang, Z., Zeng, Z., Wang, K., Wong, S.-S., Liang, W., Zanin, M., Liu, P., Cao, X., Gao, Z. and Mai, Z. (2020c). Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. *Journal of thoracic disease*. 12(3), 165.

Yarsky, P. (2021). Using a genetic algorithm to fit parameters of a COVID-19 SEIR model for US states. *Mathematics and Computers in Simulation*. 185, 687–695.

Ye, S., Li, J. and Zhang, Z. (2020). Multi-omics-data-assisted genomic feature markers preselection improves the accuracy of genomic prediction. *Journal of animal science and biotechnology*. 11(1), 109.

Yu, H., Cong, Y., Sun, G., Hou, D., Liu, Y. and Dong, J. (2024). Open-Ended Online Learning for Autonomous Visual Perception. *IEEE transactions on neural networks and learning systems*. 35(8), 10178–10198.

Yu, Y., Tan, J., Yang, Y., Zhang, B., Yao, X., Sang, S. and Deng, S. (2023). The Differential Diagnostic Value of Radiomics Signatures Between Single-Nodule Pulmonary Metastases and Second Primary Lung Cancer in Patients with Colorectal Cancer. *Technology in cancer research treatment*. 22, 15330338231175735.

Zaghloul, M., Salem, M. and Ali-Eldin, A. (2021). A new framework based on features modeling and ensemble learning to predict query performance. *PloS one*. 16(10), e0258439.

Zandavi, S. M., Rashidi, T. H. and Vafaee, F. (2021). Dynamic Hybrid Model to Forecast the Spread of COVID-19 Using LSTM and Behavioral Models Under Uncertainty. *IEEE Transactions on Cybernetics*.

Zeng, Z., Yao, S., Zheng, J. and Gong, X. (2021). Development and validation of a novel blending machine learning model for hospital mortality prediction in ICU patients with Sepsis. *BioData mining*. 14(1), 1–15.

Zhan, C., Tse, C. K., Lai, Z., Hao, T. and Su, J. (2020). Prediction of COVID-19 spreading profiles in South Korea, Italy and Iran by data-driven coding. *PLoS One*. 15(7), e0234763.

Zhang, H. and Fu, R. (2020). A Hybrid Approach for Turning Intention Prediction Based on Time Series Forecasting and Deep Learning. *Sensors (Basel, Switzerland)*. 20(17).

Zhang, L., Wong, C., Li, Y., Huang, T., Wang, J. and Lin, C. (2024a). Artificial intelligence assisted diagnosis of early tc markers and its application. *Discover oncology*. 15(1), 172.

Zhang, M., Yang, W., Chen, D., Fu, C. and Wei, F. (2024b). AM-MSFF: A Pest Recognition Network Based on Attention Mechanism and Multi-Scale Feature Fusion. *Entropy (Basel, Switzerland)*. 26(5).

Zhang, M., Zhao, C., Cheng, Q., Xu, J., Xu, N., Yu, L. and Feng, W. (2023a). A score-based method of immune status evaluation for healthy individuals with complete blood cell counts. *BMC bioinformatics*. 24(1), 467.

Zhang, S., Yang, X., Wang, Y., Zhao, Z., Liu, J., Liu, Y., Sun, C. and Zhou, C. (2020). Automatic Fish Population Counting by Machine Vision and a Hybrid Deep Neural Network Model. *Animals : an open access journal from MDPI*. 10(2).

Zhang, T., Zhou, X., Zhang, P., Duan, Y., Cheng, X., Wang, X. and Ding, G. (2022). Hardness Prediction of Laser Powder Bed Fusion Product Based on Melt Pool Radiation Intensity. *Materials (Basel, Switzerland)*. 15(13).

Zhang, Y., Tang, S. and Yu, G. J. S. R. (2023b). An interpretable hybrid predictive model of COVID-19 cases using autoregressive model and LSTM. 13(1), 6708.

Zhao, D., Zhang, H., Cao, Q., Wang, Z. and Zhang, R. (2022). The research of SARIMA model for prediction of hepatitis B in mainland China. *Medicine*. 101(23), e29317.

Zhao, X., Yang, K., He, X., Wei, Z., Zhang, J. and Yu, X. (2024). Mix proportion and microscopic characterization of coal-based solid waste backfill material based on response surface methodology and multi-objective decision-making. *Scientific reports*. 14(1), 5672.

Zhao, X., Zhou, Q., Wang, A., Zhu, F., Meng, Z. and Zuo, C. (2021). The impact of awareness diffusion on the spread of COVID-19 based on a two-layer SEIR/V–UA epidemic model. *Journal of Medical Virology*. 93(7), 4342–4350.

Zheng, N., Du, S., Wang, J., Zhang, H., Cui, W., Kang, Z., Yang, T., Lou, B., Chi, Y. and Long, H. (2020a). Predicting COVID-19 in China using hybrid AI model. *IEEE transactions on cybernetics*. 50(7), 2891–2904.

Zheng, Y., Li, Z., Xin, J. and Zhou, G. (2020b). A spatial-temporal graph based hybrid infectious disease model with application to COVID-19. *arXiv preprint arXiv:2010.09077*.

Zhong, T., Qin, G., Guo, Q. and Wang, E. (2020). Perioperative management for patients with coronavirus disease 2019. *Zhong nan da xue xue bao. Yi xue ban = Journal of Central South University. Medical sciences*. 45(5), 609–612.

Zhou, T., Liu, Q., Yang, Z., Liao, J., Yang, K., Bai, W., Lu, X. and Zhang, W. (2020). Preliminary prediction of the basic reproduction number of the Wuhan novel coronavirus 2019-nCoV. *Journal of Evidence-Based Medicine*. 13(1), 3–7.

Zhu, W., Cheng, K., Guo, Y. and Chen, Y. (2022). Comprehensive Evaluation of the Tendency of Vertical Collusion in Construction Bidding Based on Deep Neural Network. *Computational intelligence and neuroscience*. 2022, 2897672.

Zhuang, M., Li, Y., Tan, X., Xing, L. and Lu, X. (2021). Analysis of public opinion evolution of COVID-19 based on LDA-ARMA hybrid model. *Complex Intelligent Systems*. 7(6), 3165–3178.

Zivkovic, M., Bacanin, N., Venkatachalam, K., Nayyar, A., Djordjevic, A., Strumberger, I. and Al-Turjman, F. (2021). COVID-19 cases prediction by using hybrid machine learning and beetle antennae search approach. *Sustainable Cities and Society*. 66, 102669.

Zreiq, R., Kamel, S., Boubaker, S., Algahtani, F. D., Alzain, M. A., Alshammari, F., Aldhmadi, B. K., Alshammari, F. S. and J. Araúzo-Bravo, M. (2022). Predictability of COVID-19 Infections Based on Deep Learning and Historical Data. *Applied Sciences*. 12(16), 8029.

LIST OF PUBLICATIONS

Journal Articles

(a) Xu, D., Chan, W. H., & Haron, H. (2024). Enhancing infectious disease prediction model selection with multi-objective optimization: an empirical study. PeerJ Computer Science, 10, e2217.

(b) Xu, D., Chan, W. H., Haron, H., Nies, H. W., & Moorthy, K. (2024). From COVID-19 to monkeypox: a novel predictive model for emerging infectious diseases. BioData Mining, 17(1), 42.

Non-Indexed conference proceedings

(a) Xu, D., Yusuf, S. M., & Haron, H. (2022). Big data analytic tools in prediction of COVID-19 cases. In N. A. Ismail & M. Y. Abdullah (Eds.), Big data and machine learning with applications (pp. 175-188). Penerbit UTM Press. ISBN: 978-983-52-1857-6