PeerJ Computer Science

Enhancing infectious disease prediction model selection with multi-objective optimization: an empirical study

Deren Xu[1], Weng Howe Chan[2] and Habibollah Haron[1]

[1] Faculty of Computing, Universiti Teknologi Malaysia, Faculty of Computing, Johor, Johor Bahru, Malaysia
[2] Universiti Teknologi Malaysia, UTM Big Data Centre, Ibnu Sina Institute For Scientific and Industrial Resarch, Universiti Teknologi Malaysia, Johor, Johor Bahru, Malaysia

ABSTRACT

As the pandemic continues to pose challenges to global public health, developing effective predictive models has become an urgent research topic. This study aims to explore the application of multi-objective optimization methods in selecting infectious disease prediction models and evaluate their impact on improving prediction accuracy, generalizability, and computational efficiency. In this study, the NSGA-II algorithm was used to compare models selected by multi-objective optimization with those selected by traditional single-objective optimization. The results indicate that decision tree (DT) and extreme gradient boosting regressor (XGBoost) models selected through multi-objective optimization methods outperform those selected by other methods in terms of accuracy, generalizability, and computational efficiency. Compared to the ridge regression model selected through single-objective optimization methods, the decision tree (DT) and XGBoost models demonstrate significantly lower root mean square error (RMSE) on real datasets. This finding highlights the potential advantages of multi-objective optimization in balancing multiple evaluation metrics. However, this study's limitations suggest future research directions, including algorithm improvements, expanded evaluation metrics, and the use of more diverse datasets. The conclusions of this study emphasize the theoretical and practical significance of multi-objective optimization methods in public health decision support systems, indicating their wide-ranging potential applications in selecting predictive models.

Subjects Computational Biology, Algorithms and Analysis of Algorithms, Data Mining and Machine Learning, Neural Networks

Keywords Multi-objective optimization, Infectious disease prediction, Model selection, Public health, NSGA-II

Submitted 30 April 2024
Accepted 4 July 2024
Published 29 July 2024

Corresponding authors
Deren Xu, 2008xuderen@gmail.com
Weng Howe Chan, cwenghowe@utm.my

Academic editor
Shibiao Wan

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj-cs.2217**

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, Shakibazadeh & Ashoorkhani, 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, Shang & Jing, 2022*; *Li et al., 2023*; *Yang et al., 2022*). Therefore, effective infectious disease prediction models are crucial for the prevention and control of epidemic outbreaks. They 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 (*Akbulut et al., 2023*; *Khoo et al., 2024*; *Tsai, Baldwin & Gopaluni, 2021*; *Xia et al., 2022*; *Ye, Li & Zhang, 2020*). 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 & 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 & 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, Zhao & Zhang, 2023*).

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, we propose an infectious disease prediction model framework 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, we 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, we aim to contribute to the development of infectious disease prediction models and provide stronger scientific support for global public health security.

METHODOLOGY

Data selection and processing

In this study, we utilized the Mexican COVID-19 time series dataset provided by Our World in Data (*Karlinsky & 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, we selected indicators that are highly correlated with the daily new confirmed cases (new_cases) as the main variables (see Fig. 1). 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. (2022)*, *Zhang, Tang & Yu (2023)*.

To further analyze and enhance model performance, we 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)*. We also conducted thorough data cleaning. For the data processing of each model, we followed the same steps. Initially, we used RobustScaler for data scaling, a normalization method known for its robustness to outliers, as noted by *Feng et al. (2014)*, *West (2022)*. Furthermore, we 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. (2023)*. In dividing the training and test sets, we 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.

Model selection

In this study, we 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, Hassan & Mstafa (2022)*, *Lim & Zohren (2021)*.

Feedforward neural networks

Feedforward neural networks (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. The FNN model structure and parameters are shown in Table 1.

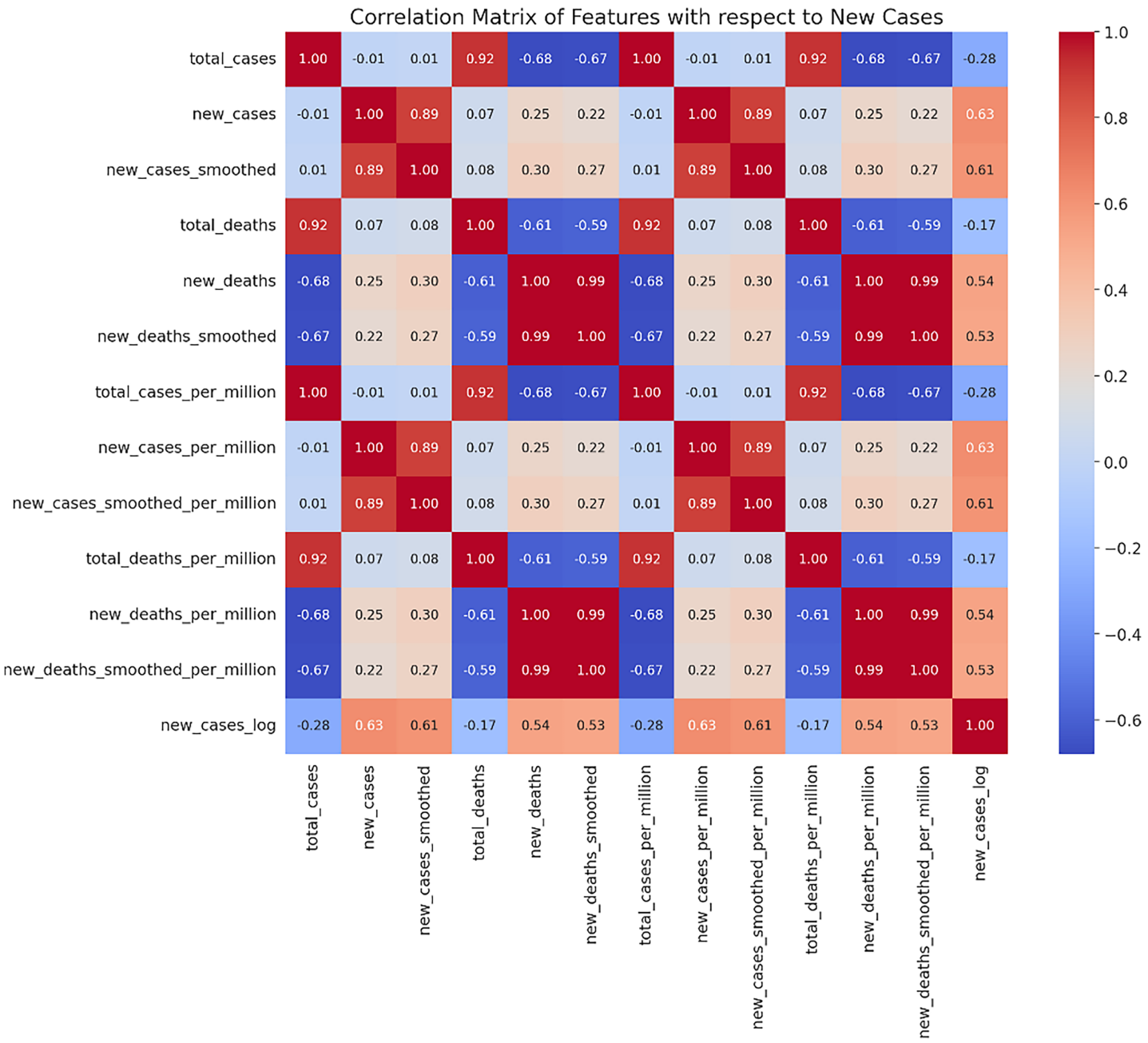

Figure 1 **Correlation matrix heatmap of new cases.** A matrix where both rows and columns are labeled with the same set of features. The cells of the matrix are colored based on the correlation coefficient between two features. Red indicates a positive correlation, while blue represents a negative correlation.
Full-size DOI: 10.7717/peerj-cs.2217/fig-1

Convolutional neural networks

Convolutional neural network (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, CNNs have also been successfully applied

Table 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

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. The CNN model structure and parameters are shown in Table 2.

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. The LSTM model structure and parameters are shown in Table 3.

Temporal convolutional network

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. The TCN model structure and parameters are shown in Table 4.

Random forest regressor

Random forest regressor (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 is particularly effective because of its ability to process large amounts of data with high accuracy. The RF model structure and parameters are shown in Table 5.

Decision tree regressor

The decision tree regressor (DT) is a type of decision tree used for regression tasks. This model is renowned for its interpretability and effectiveness in capturing non-linear

Table 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

Table 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

Table 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)

Table 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.

Table 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'.

Table 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 randomizedsearchCV, combined with 5-fold cross-validation and 50 iterations
Optimal parameter	n_estimators 200, max_depth None, learning_rate 0.1.

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. The DT model structure and parameters are shown in Table 6.

Extreme gradient boosting regressor (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. The XGBoost model structure and parameters are shown in Table 7.

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. The ridge regression model structure and parameters are shown in Table 8.

Table 8 Ridge model structure and parameters.

Parameter type	Descriptions
Hyperparameter search	Defaults
Optimal parameter	Defaults

Table 9 Model prediction performance metrics.

Name	Accuracy (RMSE)	Generalization (RMSE)	Computational efficiency
FNN	152.025	4,532.84	26.19 s
TCN	231.014	2,477.45	16.65 s
LSTM	170.442	1,0307.7	35.16 s
CNN	1,144.36	5,455.59	10.16 s
RF	18.5805	1,251.51	0.43 s
DT	24.4889	8.1421	0.07 s
xgboost	33.0709	13.0672	0.38 s
Ridge	11.5274	114.852	0.03 s

Multi-objective optimization

In this study, we focus on enhancing the performance of infectious disease prediction models through multi-objective optimization methods. Specifically, we 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 9).

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) framework 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, Li & Liu, 2022*; *Liu, Ruan & Ma, 2023*; *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 framework comprises the subsequent essential stages:

1) **Population initialization:** The population is initialized by randomly selecting algorithm parameters or model configurations.
2) **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.
3) **Genetic operations:** Crossover (cxPassThrough) and mutation (mutate) operations are applied to generate new individuals, exploring the solution space.
4) **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.
5) **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.

RESULTS

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 Fig. 2, 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.

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

1) **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.

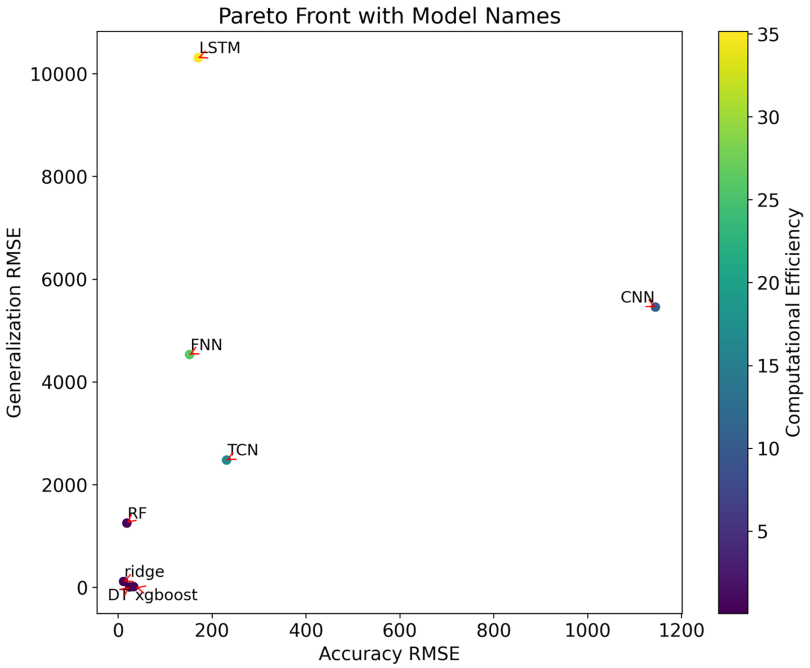

Figure 2 **Multi-objective optimisation of the Pareto frontiers.**
Full-size DOI: 10.7717/peerj-cs.2217/fig-2

2) **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.
3) **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:

1) Models 'DT' and 'XGBoost' both exhibit lower RMSE values for accuracy and generalizability, while also maintaining relatively high computational efficiency (closer to purple), suggesting they may be well-balanced models optimizing accuracy, generalizability, and computational efficiency.
2) The 'Ridge' model shows lower values in terms of accuracy and computational efficiency but slightly lacks in generalizability compared to 'DT' and 'XGBoost'.
3) 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.

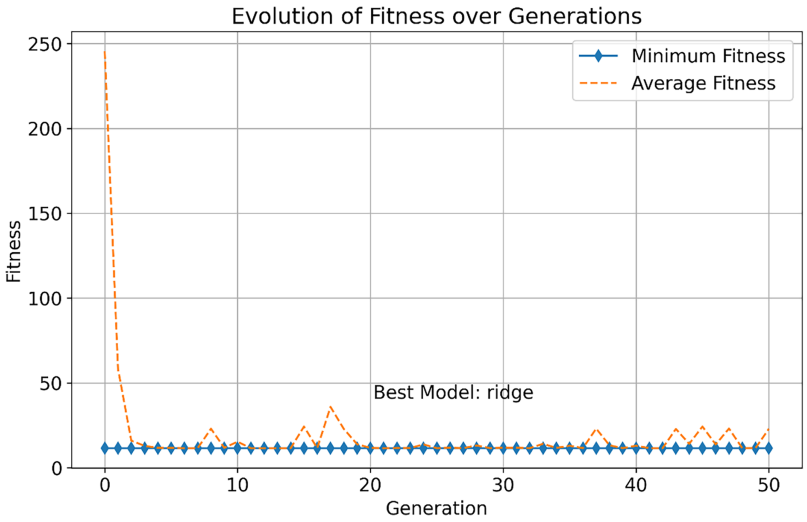

Figure 3 **Evolution of fitness over generations.** Full-size DOI: 10.7717/peerj-cs.2217/fig-3

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. 3, we 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.

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.

Case study

In order to validate the effectiveness of the method, we 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, we 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 10), 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.

Table 10 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

In contrast, the ridge regression model selected by the single-objective optimization method has a relatively high RMSE value on this dataset.

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.

DISCUSSION

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 they 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, Shrivastava & Sharma, 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 they 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, Ramaboa & Mongwe, 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, Wang & Li, 2022*). 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, Nikitas & Sotiriadis, 2021*).

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 (*Cheng 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. *Tan et al. (2017)* 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.

Comparison of different algorithms

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

1) **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, Ruan & Ma, 2023*). 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, Jain & Dhar, 2019*).

2) **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., 2022*). 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*).
3) **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 & Ye, 2023*). Shortcoming: MOEA/D may require more complex parameterization and higher computational resources for some problems (*Sun, 2023*).

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*). They 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., 2024*). These models perform mediocre in dealing with data drift because they 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., 2024*). 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.

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 framework 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.

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 framework 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 framework 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, we 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.

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 & Miani, 2024*). Models selected by multi-objective optimisation methods, such as the 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 virus (*Cai 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., 2023*). This flexibility in modelling is essential for monitoring outbreaks in real time and predicting their trends.

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:

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, we cannot confidently predict the performance of these models under other different infectious disease conditions.

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.

CONCLUSIONS

The primary goal of this research is to create and verify a multi-objective optimisation framework to improve predictive model selection for infectious disease data. By combining numerous criteria such as prediction accuracy, generalization ability, and computational efficiency, we 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 framework. 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 framework would result in more accurate and useful prediction models.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Deren Xu conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Weng Howe Chan conceived and designed the experiments, performed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

- Habibollah Haron analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data are available from Our World In Data (Mexico, Indonesia, Iran): https://ourworldindata.org/explorers/coronavirus-data-explorer?facet=none&country=~MEX&Metric=Confirmed+cases&Interval=New+per+day&Relative+to+Population=true&Color+by+test+positivity=true.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.2217#supplemental-information.

REFERENCES

Ahmed DM, Hassan MM, Mstafa RJ. 2022. A review on deep sequential models for forecasting time series data. *Applied Computational Intelligence and Soft Computing* **2022**:1–19 DOI 10.1155/2022/6596397.

Akbulut S, Yagin FH, Cicek IB, Koc C, Colak C, Yilmaz S. 2023. Prediction of perforated and nonperforated acute appendicitis using machine learning-based explainable artificial intelligence. *Diagnostics* **13(6)**:1173 DOI 10.3390/diagnostics13061173.

Alalayah KM, Senan EM, Atlam HF, Ahmed IA, Shatnawi HSA. 2023. Effective early detection of epileptic seizures through EEG signals using classification algorithms based on t-distributed stochastic neighbor embedding and K-means. *Diagnostics* **13(11)**:1957 DOI 10.3390/diagnostics13111957.

Babor M, Paquet-Durand O, Kohlus R, Hitzmann B. 2023. Modeling and optimization of bakery production scheduling to minimize makespan and oven idle time. *Scientific Reports* **13**:235 DOI 10.1038/s41598-022-26866-9.

Barea-Sepúlveda M, Calle JLP, Ferreiro-González M, Palma M. 2023. Rapid classification of petroleum waxes: a Vis-NIR spectroscopy and machine learning approach. *Foods* **12(18)**:3362 DOI 10.3390/foods12183362.

Bolla G, Berente DB, Andrássy A, Zsuffa JA, Hidasi Z, Csibri E, Csukly G, Kamondi A, Kiss M, Horvath AA. 2023. Comparison of the diagnostic accuracy of resting-state fMRI driven machine learning algorithms in the detection of mild cognitive impairment. *Scientific Reports* **13**:22285 DOI 10.1038/s41598-023-49461-y.

Budholiya K, Shrivastava SK, Sharma V. 2022. An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University-Computer and Information Sciences* **34(7)**:4514–4523 DOI 10.1016/j.jksuci.2020.10.013.

Cai J, Lu S, Cheng J, Wang L, Gao Y, Tan T. 2022. Collaborative variable neighborhood search for multi-objective distributed scheduling in two-stage hybrid flow shop with sequence-dependent setup times. *Scientific Reports* **12**:15724 DOI 10.1038/s41598-022-19215-3.

Cai L, Zhao E, Niu H, Liu Y, Zhang T, Liu D, Zhang Z, Li J, Qiao P, Lv H, Ren P, Zheng W, 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 DOI 10.1016/j.compbiomed.2023.107264.

Cao W, Zhu J, Wang X, Tong X, Tian Y, Dai H, Ma Z. 2022. Optimizing spatio-temporal allocation of the COVID-19 vaccine under different epidemiological landscapes. *Frontiers in Public Health* **10**:921855 DOI 10.3389/fpubh.2022.921855.

Cheng Q, Collender PA, Heaney AK, McLoughlin A, Yang Y, Zhang Y, Head JR, Dasan R, Liang S, Lv Q, Liu Y, Yang C, Chang HH, Waller LA, Zelner J, Lewnard JA, Remais JV. 2022. Optimizing laboratory-based surveillance networks for monitoring multi-genotype or multi-serotype infections. *PLOS Computational Biology* **18(9)**:e1010575 DOI 10.1371/journal.pcbi.1010575.

Cui J, Li K, Hao J, Dong F, Wang S, Rodas-González A, Zhang Z, Li H, 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)**:1936 DOI 10.3390/foods11131936.

Dixon S, Keshavamurthy R, Farber DH, Stevens A, Pazdernik KT, Charles LE. 2022. A comparison of infectious disease forecasting methods across locations, diseases, and time. *Pathogens (Basel, Switzerland)* **11(2)**:185 DOI 10.3390/pathogens11020185.

Du H, Xu Q, Jiang L, Bu Y, Li W, Yan J. 2024. Stepwise identification method of thermal load for box structure based on deep learning. *Materials (Basel, Switzerland)* **17(2)**:357 DOI 10.3390/ma17020357.

Entezari A, Liu NC, Zhang Z, Fang J, Wu C, Wan B, Swain M, Li Q. 2023. Nondeterministic multiobjective optimization of 3D printed ceramic tissue scaffolds. *Journal of the Mechanical Behavior of Biomedical Materials* **138**:105580 DOI 10.1016/j.jmbbm.2022.105580.

Feng C, Wang H, Lu N, Chen T, He H, Lu Y, Tu XM. 2014. Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry* **26(2)**:105–109 DOI 10.3969/j.issn.1002-0829.2014.02.009.

Feng H, Zhang X. 2023. A novel encoder-decoder model based on autoformer for air quality index prediction. *PLOS ONE* **18(4)**:e0284293 DOI 10.1371/journal.pone.0284293.

Fialho BC, Gauss L, Soares PF, Medeiros MZ, Lacerda DP. 2023. Vaccine innovation meta-model for pandemic contexts. *Journal of Pharmaceutical Innovation* **18(3)**:1145–1193 DOI 10.1007/s12247-023-09708-7.

Gao Q, Shang WP, Jing MX. 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)**:13343 DOI 10.3390/ijerph192013343.

Hernández-Giottonini KY, Arellano-Reynoso B, Rodríguez-Córdova RJ, de la Vega-Olivas J, Díaz-Aparicio E, Lucero-Acuñacorresponding author A. 2023. Enhancing therapeutic efficacy against Brucella canis infection in a murine model using rifampicin-loaded PLGA nanoparticles. *ACS Omega* **8(51)**:49362–49371 DOI 10.1021/acsomega.3c07892.

Hlongwane R, Ramaboa K, Mongwe W. 2024. Enhancing credit scoring accuracy with a comprehensive evaluation of alternative data. *PLOS ONE* **19(5)**:e0303566 DOI 10.1371/journal.pone.0303566.

Hu Z, Li P, Liu Y. 2022. Enhancing the performance of evolutionary algorithm by differential evolution for optimizing distillation sequence. *Molecules* **27(12)**:3802 DOI 10.3390/molecules27123802.

Hu Y, Yang Q, Zhang J, Peng Y, Guang Q, 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 DOI 10.1186/s13018-023-03858-7.

Huang Y, Zhang Z, Jiao A, Ma Y, 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 DOI 10.1109/TVCG.2023.3326921.

Husnayain A, Shim E, Fuad A, Chia-Yu Su E. 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 DOI 10.2196/34178.

Joseph VR. 2022. Optimal ratio for data splitting. *Statistical Analysis and Data Mining* **15(4)**:531–538 DOI 10.1002/sam.11583.

Karlinsky A, Kobak D. 2021. The world mortality dataset: tracking excess mortality across countries during the COVID-19 pandemic. *medRxiv* DOI 10.1101/2021.01.27.21250604.

Khatun D, Hossain MY, Rahman O, Hossain MF. 2022. Estimation of life history parameters for river catfish Eutropiichthys vacha: insights from multi-models for sustainable management. *Heliyon* **8(10)**:e10781 DOI 10.1016/j.heliyon.2022.e10781.

Khoo LS, Lim MK, Chong CY, McNaney R. 2024. Machine learning for multimodal mental health detection: a systematic review of passive sensing approaches. *Sensors* **24(2)**:348 DOI 10.3390/s24020348.

Kozyrev EA, Ermakov EA, Boiko AS, Mednova IA, Kornetova EG, Bokhan NA, Ivanova SA. 2023. Building predictive models for schizophrenia diagnosis with peripheral inflammatory biomarkers. *Biomedicines* **11(7)**:1990 DOI 10.3390/biomedicines11071990.

Kumari K, Jain S, Dhar A. 2019. Computationally efficient approach for identification of fuzzy dynamic groundwater sampling network. *Environmental Monitoring and Assessment* **191(5)**:310 DOI 10.1007/s10661-019-7467-3.

Lange O. 2023. Health economic evaluation of preventive digital public health interventions using decision-analytic modelling: a systematized review. *BMC Health Services Research* **23**:268 DOI 10.1186/s12913-023-09280-3.

Le Fouest S, Mulleners K. 2024. Optimal blade pitch control for enhanced vertical-axis wind turbine performance. *Nature Communications* **15(1)**:2770 DOI 10.1038/s41467-024-46988-0.

Li Y, Meng X, Zhang Z, Song G. 2020. A machining state-based approach to tool remaining useful life adaptive prediction. *Sensors* **20(23)**:6975 DOI 10.3390/s20236975.

Li Y, Wang Y, Shen Z, Miao F, Wang J, Sun Y, Zhu S, Zheng Y, Guan S. 2022. A biodegradable magnesium alloy vascular stent structure: design, optimisation and evaluation. *Acta Biomaterialia* **142(5)**:402–412 DOI 10.1016/j.actbio.2022.01.045.

Li G, Wang Y, Zhang C, Xu C, Zhan L. 2024. Study on the impact of building energy predictions considering weather errors of neighboring weather stations. *Sensors* **24(4)**:1157 DOI 10.3390/s24041157.

Li S, Zhu L, Zhang L, Zhang G, Ren H, Lu L. 2023. 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)**:3328 DOI 10.3390/ijerph20043328.

Liao Y, Han L, Wang H, Zhang H. 2022. Prediction models for railway track geometry degradation using machine learning methods: a review. *Sensors* **22(19)**:7275 DOI 10.3390/s22197275.

Lim B, Zohren S. 2021. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A* **379(2194)**:20200209 DOI 10.1098/rsta.2020.0209.

Liu C, Ruan K, Ma X. 2023. DMEformer: a newly designed dynamic model ensemble transformer for crude oil futures prediction. *Heliyon* **9(6)**:e16715 DOI 10.1016/j.heliyon.2023.e16715.

Liu X, Tian J, Duan P, Yu Q, Wang G, Wang Y. 2024. GrMoNAS: a granularity-based multi-objective NAS framework for efficient medical diagnosis. *Computers in Biology and Medicine* **171(4)**:108118 DOI 10.1016/j.compbiomed.2024.108118.

Liu T, Ye A. 2023. Domain knowledge-assisted multi-objective evolutionary algorithm for channel selection in brain-computer interface systems. *Frontiers in Neuroscience* **17**:1251968 DOI 10.3389/fnins.2023.1251968.

Lv CX, An SY, Qiao BJ, 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**:839 DOI 10.1186/s12879-021-06503-y.

Mathieu E, Ritchie H, Rodés-Guirao L, Appel C, Gavrilov, D, Giattino C, Hasell J, Macdonald B, Dattani S, Beltekian D, Ortiz-Ospina E, Roser M. 2020. Coronavirus pandemic (COVID-19). Our World in Data. *Available at https://ourworldindata.org/coronavirus*.

Mirzania M, Shakibazadeh E, 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**:1118 DOI 10.1186/s12913-022-08510-4.

Mohammed S, Sha'aban YA, Umoh IJ, Salawudeen AT, Ibn Shamsah SM. 2023. A hybrid smell agent symbiosis organism search algorithm for optimal control of microgrid operations. *PLOS ONE* **18(6)**:e0286695 DOI 10.1371/journal.pone.0286695.

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 DOI 10.1186/s13690-021-00750-w.

Padilla-García EA, Cervantes-Culebro H, Rodriguez-Angeles A, Cruz-Villar CA. 2023. Selection/control concurrent optimization of BLDC motors for industrial robots. *PLOS ONE* **18(8)**:e0289717 DOI 10.1371/journal.pone.0289717.

Papafotis K, Nikitas D, Sotiriadis PP. 2021. Magnetic field sensors' calibration: algorithms' overview and comparison. *Sensors* **21(16)**:5288 DOI 10.3390/s21165288.

Piscitelli P, 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)**:314 DOI 10.3390/ijerph21030314.

Sassano M, Mariani M, Quaranta G, Pastorino R, Boccia S. 2022. Polygenic risk prediction models for colorectal cancer: a systematic review. *BMC Cancer* **22**:65 DOI 10.1186/s12885-021-09143-2.

Sharma S, Alsmadi I, Alkhawaldeh RS, Al-Ahmad BJC. 2022. Data-driven analysis and predictive modeling on COVID-19. *Concurrency and Computation: Practice & Experience* **34(28)**:e7390 DOI 10.1002/cpe.7390.

Silvestri S, Islam S, Papastergiou S, Tzagkarakis C, Ciampi M. 2023. A machine learning approach for the NLP-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem. *Sensors* **23(2)**:651 DOI 10.3390/s23020651.

Sun J. 2023. A multi-objective optimization based doherty power amplifier and its matching network optimization method. *PLOS ONE* **18(12)**:e0293371 DOI 10.1371/journal.pone.0293371.

Tan L, Wang H, Yang C, Niu B. 2017. A multi-objective optimization method based on discrete bacterial algorithm for environmental/economic power dispatch. *Natural Computing* **16**:549–565 DOI 10.1007/s11047-017-9620-7.

Tian S, Sun S, Mao W, Qian S, Zhang L, Zhang G, Xu B, Chen M. 2021. Development and validation of prognostic nomogram for young patients with kidney cancer. *International Journal of General Medicine* **14**:5091–5103 DOI 10.2147/IJGM.S331627.

Tsai Y, Baldwin SA, Gopaluni B. 2021. Identifying indicator species in ecological habitats using deep optimal feature learning. *PLOS ONE* **16(9)**:e0256782 DOI 10.1371/journal.pone.0256782.

Vukašinović A, Klisic A, Ostanek B, Kafedžić S, Zdravković M, Ilić I, Sopić M, Hinić S, Stefanović M, Bogavac-Stanojević N, Marc J, Nešković AN, 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)**:14308 DOI 10.3390/ijms241814308.

Wang YC, Houng YC, Chen HX, Tseng SM. 2023. Network anomaly intrusion detection based on deep learning approach. *Sensors (Basel, Switzerland)* **23(4)**:2171 DOI 10.3390/s23042171.

Wang Y, Yan Z, Wang D, Yang M, Li Z, Gong X, Wu Di, Gong X, Wang Y, Zhai L, Zhang W, Wang Y. 2022. Prediction and analysis of COVID-19 daily new cases and cumulative cases: times series forecasting and machine learning models. *BMC Infectious Diseases* **22**:495 DOI 10.1186/s12879-022-07472-6.

Wang Z, Zhao C, Zhang W. 2023. Multi-objective design and optimization of squeezed branch pile based on orthogonal test. *Scientific Reports* **13**:22508 DOI 10.1038/s41598-023-49936-y.

West RM. 2022. Best practice in statistics: the use of log transformation. *Annals of Clinical Biochemistry* **59(3)**:162–165 DOI 10.1177/00045632211050531.

Xia Z, Qin L, Ning Z, Zhang X. 2022. Deep learning time series prediction models in surveillance data of hepatitis incidence in China. *PLOS ONE* **17(4)**:e0265660 DOI 10.1371/journal.pone.0265660.

Yang M, Shi L, Chen H, Wang X, Jiao J, Liu M, Yang J. 2022. 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 DOI 10.1186/s12939-022-01723-3.

Yang J, Wang Y, Li X. 2022. Prediction of stock price direction using the LASSO-LSTM model combines technical indicators and financial sentiment analysis. *PeerJ. Computer Science* **8**:e1148 DOI 10.7717/peerj-cs.1148.

Ye S, Li J, 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 DOI 10.1186/s40104-020-00515-5.

Zhang Y, Tang S, Yu G. 2023. An interpretable hybrid predictive model of COVID-19 cases using autoregressive model and LSTM. *Scientific Reports* **13**:6708 DOI 10.21203/rs.3.rs-2261448/v1.

Zhang L, Wong C, Li Y, Huang T, Wang J, Lin C. 2024. Artificial intelligence assisted diagnosis of early tc markers and its application. *Discover Oncology* **15(1)**:172 DOI 10.1007/s12672-024-01017-w.

Zhang M, Zhao C, Cheng Q, Xu J, Xu N, Yu L, Feng W. 2023. A score-based method of immune status evaluation for healthy individuals with complete blood cell counts. *BMC Bioinformatics* **24**:467 DOI 10.1186/s12859-023-05603-7.

Zhao X, Yang K, He X, Wei Z, Zhang J, 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**:5672 DOI 10.1038/s41598-024-56028-y.

Zhao D, Zhang H, Cao Q, Wang Z, Zhang R. 2022. The research of SARIMA model for prediction of hepatitis B in mainland China. *Medicine* **101(23)**:e29317 DOI 10.1097/MD.0000000000029317.