Feature selection to predict COVID-19 new patients in the four southern border provinces of Thailand

Chadaphim Photphanloet, Sherif Eneye Shuaib, Siriprapa Ritraksa, Pakwan Riyapan
Department of Mathematics and Computer Science, Faculty of Science and Technology, Prince of Songkla University, Pattani Campus, Pattani, Thailand

Article Info

Article history:

Received Jul 28, 2024
Revised May 19, 2025
Accepted Oct 14, 2025

Keywords:

Coronavirus disease prediction
Feature selection
Ranking arrangement
Regression problem
Supervised learning model

ABSTRACT

This paper presents a machine learning-based prediction framework that utilizes ensemble feature selection techniques to accurately forecast the number of new coronavirus disease (COVID-19) infections in Thailand's four southern border provinces. The framework used include multiple linear regression (MLR), multilayer perceptron neural networks (MLP-NN), and support vector regression (SVR), to classify short-term trends in new patient cases. The study evaluates the effectiveness of these models across different provinces and demonstrates how integrating feature selection methods: forward selection (FS), backward elimination (BE), and genetic algorithms (GA) enhances prediction accuracy. The findings highlight the adaptability of the models, with each province benefiting from tailored model-feature selection strategies. The results show that the predictive models align closely with real patient data, enabling authorities to anticipate outbreaks and implement timely interventions. Moreover, the proposed methodology can be applied to other epidemics, making it a valuable tool for public health planning and preparedness. The study offers actionable insights for decision-makers, emphasizing the importance of predictive modeling in community-level outbreak management.

Corresponding Author:

Pakwan Riyapan
Department of Mathematics and Computer Science, Faculty of Science and Technology
Prince of Songkla University, Pattani Campus
Pattani 94000, Thailand
Email: pakwan.r@psu.ac.th

1. INTRODUCTION

Coronavirus disease (COVID-19), an infectious disease transmitted through the air, was initially termed the Wuhan virus due to its origins in Wuhan, China, where the first human case was recorded [1]. According to the World Health Organization, as of 31 March 2024, the protracted pandemic has already caused over 774 million confirmed cases [2]. The disease rapidly evolved into a pandemic, negatively affecting people worldwide, leading the World Health Organization (WHO) to declare a global health emergency by the end of January 2020. In Thailand and other parts of the world, the crisis resulted in numerous challenges that led to pressures in the healthcare system and economic strain. A critical challenge was the need to develop rapid and accurate methods for diagnosing and predicting the severity of COVID-19 cases to optimize resource allocation and patient care.

The peak of the COVID-19 pandemic profoundly impacted the Thai economy. The implementation of various restrictions led to temporary disruptions and permanent closures of businesses, causing a notable decline in the gross domestic product (GDP) in 2020 [3]. The tourism sector, in particular, suffered significantly. Strict regulations, including mandatory quarantines for visitors, sharply decreased tourism's contribution to the GDP. The COVID-19 pandemic significantly altered the lifestyles of Thai people. Beyond heightened health concerns, they had to adapt to several changes in their daily routines. Notable shifts included the rise of online shopping, cashless transactions, online education, and remote work [4]. A survey conducted in January 2023 on online shopping behavior indicated that the majority of Thais had increased their online shopping activities. Additionally, remote work became a common practice for many employees during the pandemic [5]. As of March 17, 2024, Thailand had reported approximately 4.76 million confirmed cases of COVID-19. During this same period, the country recorded 34,576 deaths attributable to the virus [6].

As COVID-19 cases continue to rise, the volume of related data grows daily, providing an opportunity for data mining to extract meaningful insights. In light of this, the development and deployment of machine-based predictive models have become increasingly critical. These models provide a rapid and accurate detection of COVID-19, addressing the challenges posed by limited testing resources. Various machine learning techniques have been proposed for COVID-19 diagnosis and severity prediction, with feature selection being a common approach. For instance, Shaban *et al.* [7] proposed a hybrid feature selection method combining a filter-based rapid feature selection technique with a genetic algorithm (GA) to identify the most relevant features from chest CT images for COVID-19 diagnosis. Sun *et al.* [8] introduced an adaptive feature selection guided deep forest (AFS-DF) model for COVID-19 classification, achieving an accuracy of 91.79% using chest CT images. Other methods have explored different techniques for COVID-19 detection. Abraham and Nair [9] combined multiple pre-trained convolutional neural networks (CNNs) to detect COVID-19 from X-ray images, achieving accuracies of 91.16% and 97.44% on different datasets. Pourhomayoun and Shakibi [10] applied a variety of filter and wrapper methods for feature selection, narrowing down 42 features from an initial 112 to predict mortality in COVID-19 patients with 93% accuracy using a neural network algorithm. XGBoost classifiers have also been applied in this field. Wong *et al.* [11] used this approach to predict COVID-19 severity with UK Biobank data, achieving an accuracy of 86.68%. Similarly, Yan *et al.* [12] developed an XGBoost model incorporating demographic factors, resulting in 90% accuracy. Yao *et al.* [13] applied a support vector machine (SVM) classifier, achieving 81.5% accuracy, while Hu *et al.* [14] used logistic regression with an accuracy of 85%, both studies utilizing data from Tongji Hospital [15]. Sun *et al.* [16] created an SVM model using data from the Shanghai Public Health Clinical Centre, achieving 87.75% accuracy. Additional research has tested these models on different populations. An *et al.* [17] evaluated several classifiers using data from the Korean National Health Insurance Service, with linear SVM achieving an AUC of 96.2%. Zagrouba *et al.* [18] applied an SVM model to WHO data, obtaining an accuracy of 96.7%. While these studies achieved high accuracy in predicting COVID-19 severity, many models face limitations when applied to specific populations or geographic regions [19]-[21].

Five years have passed since COVID-19 was first identified, with 774 million cases and 7 million deaths reported globally. Although the most severe phase of the pandemic is behind us, COVID-19 continues to persist worldwide In Thailand, there has been a recent rise in hospital admissions for COVID-19, particularly among severe pneumonia cases requiring ventilators, and an increase in fatalities. This surge is mainly attributed to the higher transmissibility of the virus and reduced adherence to preventive measures Health officials in Thailand have reported 23,245 cases between January 1 and June 8, with 2,762 cases recorded in the week of June 2-8. Several regions, including the Northeast, Southern, Eastern areas, Bangkok, and surrounding areas, have seen a notable increase in cases [22]. In some regions, particularly the southern provinces, existing machine-learning models for COVID-19 prediction have not been widely tested. There is also a growing need for models that integrate diverse data sources while accounting for the local healthcare context. Thus, this study proposes an ensemble feature selection-based classification system designed to address the specific needs of Thailand's southern provinces. Our approach utilizes multiple machine learning techniques, including multiple linear regression (MLR), multilayer perceptron neural networks (MLP-NN), and support vector regression (SVR), to predict COVID-19 severity with higher accuracy. By focusing on region-specific data, we aim to provide a more tailored prediction model for COVID-19 cases in Thailand.

This article has four detailed sections as follows: section 1 is the introduction. Section 2 is the method consisting of the experimental dataset, feature selection, MLP-NN, SVR, performance evaluation, implementation process, and data preparation. Section 3 is the results and discussion and section 4 is the conclusion.

2. METHODS

2.1. Experimental dataset

The evaluation of the method is carried out by using data derived from the Department of Disease Control, Ministry of Public Health, Thailand, which has COVID-19 situation reports in Thailand [23]. In this paper, we used data from four southern provinces in Thailand as a case study. As highlighted in the introduction, the southern region has experienced a rise in COVID-19 infection cases, primarily due to the prevalence of large family households. These provinces were selected because they are all situated in the southern region, where the increasing number of new patients presents a relevant context for this study. Table 1 shows the area, population, and density of population per area of the four southern provinces, which is the data collected by the Center for Information and Communication Technology, Office of the Permanent Secretary, Ministry of Interior, Thailand. Figure 1 shows the locations of the four southern provinces that were used as a case study consists of Songkhla Province, Pattani Province, Yala Province, and Narathiwat Province.

Table 1. The area, population, and density of population per area of each province

Province	Area (km^2)	Population (people)	Density (people/km^2)
Songkhla	7393.89	1,401,303	189.52
Pattani	1940.35	686,186	352.64
Yala	4,521	542,848	120.07
Narathiwat	4,475.43	774,799	173.12

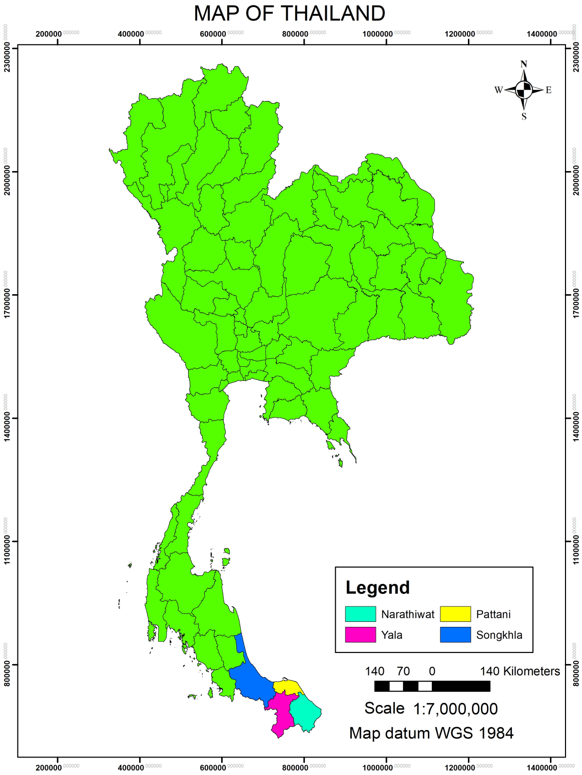

Figure 1. The locations of the four southern provinces that were used as a case study created using ArcGIS software [24]

The data obtained from COVID-19 situation reports consists of COVID-19 new patients, cumulative patients, stay persons, healer persons, and deceased persons, who received 1 dose of vaccination, who received 2 doses of vaccination, and who received 3 doses of vaccination. The method to evaluate the data COVID-19 situation reports every day from the websites of the Department of Disease Control, Ministry of Public Health, and Thailand during the period from the start of April 2021 to the end of December 2021.

Table 2 provides the basic statistical values of daily COVID-19 new patients in four southern provinces. From these data, it can be seen that the number of new COVID-19 patients was high. The average number of new COVID-19 patients for the four southern provinces shows variation with a narrow band between 158 and 247 people; while the standard deviation at each province varies, falling in the range of 155-201 people, the data distribution is relatively broad.

Table 2. The basic statistical values of COVID-19 new patients of four southern provinces

Province	Minimum	Maximum	Average	Standard deviation
Songkhla	0	697	246.89	191.08
Pattani	0	669	180.41	166.43
Yala	0	785	180.01	200.52
Narathiwat	0	618	158.15	155.54

2.2. Feature selection

An essential step in predictive modeling is selecting an appropriate subset of features to serve as input variables. Excluding key features may significantly compromise the model's predictive performance, while incorporating too many irrelevant variables can reduce accuracy and increase the training process's computational complexity. This study explores three feature selection methods: forward selection (FS), backward elimination (BE), and GA.

FS follows a stepwise approach, where features are added to the model incrementally. At each iteration, the algorithm evaluates the inclusion of each unused feature based on its correlation with the dependent variable. The process begins with an empty feature set and sequentially incorporates features that improve model performance. The addition of new features halts when no statistically significant improvement is observed, as demonstrated in Table 3 [25].

Table 3. Illustrative example of the FS and BE processes for identifying relevant features

FS	BE
Input: $\{S_1, S_2, S_3, S_4, S_5\}$	Input: $\{B_1, B_2, B_3, B_4, B_5\}$
Initial set of features: $\{\}$	Initial set of features:
$\Rightarrow \{S_1\}$	$\Rightarrow \{B_1, B_2, B_3, B_4, B_5\}$
$\Rightarrow \{S_1, S_4\}$	$\Rightarrow \{B_1, B_2, B_4, B_5\}$
$\Rightarrow$ Add a feature: $\{S_1, S_4, S_5\}$	$\Rightarrow$ Remove a feature: $\{B_1, B_2, B_5\}$

Feature selection methods were implemented using RapidMiner, which offers a Python integration library available on GitHub. The BE technique, on the other hand, initiates with a full model that includes all candidate features. In each step, the feature with the highest p-value, indicating the weakest contribution, is considered for removal. If this value exceeds a predefined significance threshold, the feature is excluded. This iterative process continues until the model retains only those features that contribute meaningfully, as detailed in Table 3 [26].

The GA is a method for finding the optimal solution to a problem by mimicking natural evolution, based on Charles Darwin's theory of natural selection [27]. It is one of the techniques in artificial intelligence, where new generations are created through processes similar to genetic evolution, inheriting various characteristics from previous generations. The flowchart describing the GA is presented in Figure 2.

Furthermore, the steps involved in this genetic method are described as follows:

Step 1 Chromosome encoding: chromosome encoding is an important step in the GA. Each position in a string represents a gene in a chromosome that contains a feature of the solution. An n-bit binary integer can be used to encode the mix of randomly chosen properties that make up each chromosome. Each bit, or gene, stands for a distinct feature, with 0 denoting a "not selected feature" and 1 denoting a "chosen feature." Each chromosome is encoded by an 8-bit binary number that represents each of the eight features in our data as consists of new COVID-19 patients (NP_t), cumulative patients (CP_t), stayed person (SP_t), healer person (HP_t), deceased people (DP_t), people who received 1 dose of vaccination ($RV_{1,t}$), people who received 2 doses of vaccinations ($RV_{2,t}$), and people who received 3 doses of vaccinations ($RV_{3,t}$), respectively. Here, there are 256 chromosomes, and each set is unique.

Step 2 Population initialization: in this step, the population is randomly generated by selecting 10 chromosomes at random from the 256 available. Several features on this specific chromosome will be used as inputs for the prediction model, which serves as a fitness function for the GA.

Step 3 Fitness evaluation: for this step, the fitness score that determines each chromosome's suitability can be calculated from the fitness function. In this paper, prediction models for the COVID-19 new patients are used as fitness functions, i.e., MLR, MLP-NN, SVR, to calculate the fitness score of each chromosome.

Step 4 Selection: the most suited chromosomes are chosen, and the genes from these chromosomes are passed on to the following generation. Chromosome pairs are chosen to serve as the reproduction's "parents." The high fitness scores of chromosomes have a better probability of getting chosen.

Step 5 Crossover: in this step, a crossover point is randomly selected for each mating pair, guided by a crossover probability within the interval [0,1] [28]. This operation involves the exchange of genetic material between parent chromosomes to generate two new offspring. These offspring are then incorporated into the new population, ensuring that the size of the next generation remains equal to that of the original population.

Step 6 Mutation: the general concept of mutation is to randomly select the genes and change the values of the genes under the probability in the range of [0,1] [28] to avoid duplicate chromosome problems.

Step 7 Replacement: the current population is replaced with the new population of the same size, and then the algorithm is repeated for steps 3–6 until it reaches a termination condition, i.e., it converges to the solution or its iterations reach the maximum number.

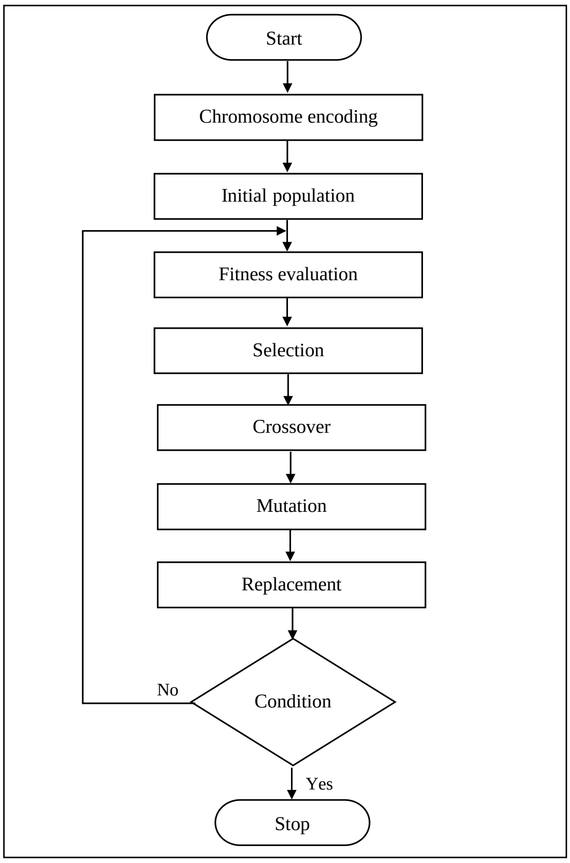

Figure 2. The procedure of the GA

In this paper, independent variables are COVID-19 new patients, cumulative patients, stay persons, healer persons, deceased persons, those who received 1 dose of vaccination, those who received 2 doses of vaccination, and those who received 3 doses of vaccination; whereas, the dependent variable is for COVID-19 new patients at each province 5 days ahead. The MLR model is expressed in (1):

$$Z = a_0 + a_1Y_1 + a_2Y_2 + ... + +a_pY_p + e \quad (1)$$

where Z is the predicted COVID-19 new patients, Y_i is the ith choosing independent variables, a_i is the weight of selected independent variables for $i = 1, 2, ..., p$, a_0 is a constant, e is an error value, and p is the number of choosing independent variables. The optimum values of a_0 and a_p can be found when the least square error occurs between the actual COVID-19 new patient and the predicted COVID-19 new patient. We utilized RapidMiner software, which offers a Python library on GitHub for MLR integration.

2.3. Multilayer perceptron neural network

One of several neural networks in the feed-forward artificial neural network category is a MLP-NN [29]. MLP-NN can solve the problems of data sets with linear and nonlinear characteristics, in which COVID-19 new patients can be used as a training called, called the backpropagation, and the activation function is the sigmoid function. Figure 3 shows the network layer, which consists of an input layer, hidden layers, and an output layer, each of which is fully connected to all nodes.

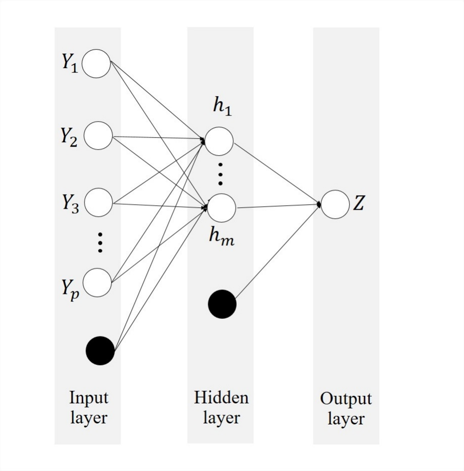

Figure 3. MLP-NN

For the input layer, each node represents COVID-19 new patients, cumulative patients, stay persons, healer persons, deceased persons, who received 1 dose of vaccination, those who received 2 doses of vaccination, and those who received 3 doses of vaccination, whereas the output layer gives the predicted COVID-19 new patients at each province 5 days ahead as a result. In this paper, the hidden layer has only 1 layer and the number of hidden nodes is equal to m where the value of each hidden node is calculated from the sum of the multiplications of input data from (2) and the weights using the sigmoid function as the activation function. The value of the output node can be calculated from the sum of the multiplications of hidden node values and the weights by using the sigmoid function as the activation function. This is shown in (3).

$$h_k = \sum_{i=1}^{p} \omega_{ik} Y_i + \omega_0 \tag{2}$$

$$Z = \sum_{j=1}^{m} \beta_j h_k \tag{3}$$

where h_k is the value of the kth hidden node for $1, 2, ..., m$, ω_{jk} is the weight from Y_i to h_k in the closed interval of $[-1, 1]$, ω_0 is the bias value from the hidden layer to the output node, p is the number of input nodes, Z is the output node, and β_j is the weight from h_k to Z in the closed interval of $[-1, 1]$.

2.4. Support vector regression

SVR is a type of SVM algorithm adapted for regression tasks. It functions as an approximation method designed to predict continuous values using data points referred to as support vectors. The goal of SVR is to determine a function $f(x)$, as shown in (4), that deviates from the actual observed values (Y_i) by no more than a predefined margin ε, while maintaining the flattest possible regression function [30].

In this study, a convex optimization approach is applied to SVR by incorporating a soft margin loss function. This involves the introduction of slack variables $(\vartheta_i, \vartheta_i^*)$ to handle constraints that cannot be satisfied within the εinsensitive zone. The corresponding optimization formulation is presented in (5):

$$f(x) = \langle w, x \rangle + b \tag{4}$$

Here, $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$, with $\langle \cdot, \cdot \rangle$ representing the dot product in $\mathbb{R}^n$, and $\|w\|^2 = \langle w, w \rangle$. The parameter b is a bias term.

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\vartheta_i + \vartheta_i^*) \tag{5}$$

subject to:

$$\begin{aligned} Y_i - \langle w, x_i \rangle - b &\leq \varepsilon + \vartheta_i, \\ \langle w, x_i \rangle + b - Y_i &\leq \varepsilon + \vartheta_i^*, \\ \vartheta_i,\, \vartheta_i^* &\geq 0. \end{aligned}$$

The constant C determines the balance between the model's flatness and the degree to which deviations larger than ε are penalized.

2.5. Performance evaluation

In this research, four standard statistical models are used as assessment criteria [31], namely the root mean square error (RMSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and the correlation coefficient (R^2).These metrics ensure a comprehensive evaluation of model performance by capturing different aspects of prediction accuracy. RMSE measures the model's prediction error, with larger errors penalized more heavily. It is computed as (6):

$$RMSE = \sqrt{\frac{1}{2}\sum_{j=1}^{P}(Y_j + \hat{Y}_j)^2} \tag{6}$$

where $\hat{Y}_j$ and Y_j are the predicted value and the observed value for $j = 1, 2, ..., P$ and P is the number of data to predict COVID-19 new patients. RMSE can measure goodness-of-fit and describe the predicted average error. MAE calculates the average absolute difference between observed and predicted values, while MAPE expresses the error as a percentage, making it useful for comparing errors across datasets with different scales. They can be computed as (7) and (8):

$$MAPE = \frac{100}{P}\sum_{j=1}^{P}|\frac{Y_j + \hat{Y}_j}{Y_j}| \tag{7}$$

$$MAE = \frac{1}{P}\sum_{j=1}^{P}|Y_j - \hat{Y}_j| \tag{8}$$

The statistical model that we use to find the relationship between the predicted values and the observed values is the correlation coefficient, R^2, which can be calculated by (9):

$$R^2 = \frac{\sum_{j=1}^{P}(\hat{Y}_j - \mu_{\hat{Y}})(Y_j - \mu_Y)}{\sqrt{\sum_{j=1}^{P}(\hat{Y}_j - \mu_{\hat{Y}})^2}\sqrt{\sum_{j=1}^{P}(Y_j - \mu_Y)^2}} \tag{9}$$

where $\mu_{\hat{Y}}$ and μ_Y are the average predicted value and the average observed value.

2.6. Implementation process and data preparation

To predict the number of new COVID-19 patients in each province 5 days ahead (NP_{t+5}), we propose a model that integrates a feature selection method with supervised learning for regression problems. This model ranks the COVID-19 new patient predictions across four provinces. Given the absence of a clear premise that input features influence the prediction of new COVID-19 cases, we utilize a feature selection method to identify only the associated features. Subsequently, these features serve as input to the supervised learning models for regression.

Each feature $(NP_t, CP_t, SP_t, HP_t, DP_t, RV_{1,t}, RV_{2,t}, RV_{3,t})$ possesses a distinct unit of measurement. As a result, the raw data must be standardized to ensure comparability. This is accomplished by using a standard normalization method, which rescales all values to fall within the interval $[0, 1]$ [32]. The resulting normalized values are stored in $Data_j$, where $j = 1, 2, 3, 4$. Following normalization, the data are further transformed into a mean of zero and a standard deviation of one. The normalization process is defined as:

$$y^{'} = \frac{y - \bar{y}}{s} \tag{10}$$

where $y^{'}$ is the normalized data, y is the original data, $\bar{y}$ is the mean of the original data, and s is the standard deviation of $y^{'}$.

The flowchart illustrating the proposed method for predicting COVID-19 cases 5 days in advance, using provinces as case studies, is presented in Figure 4.

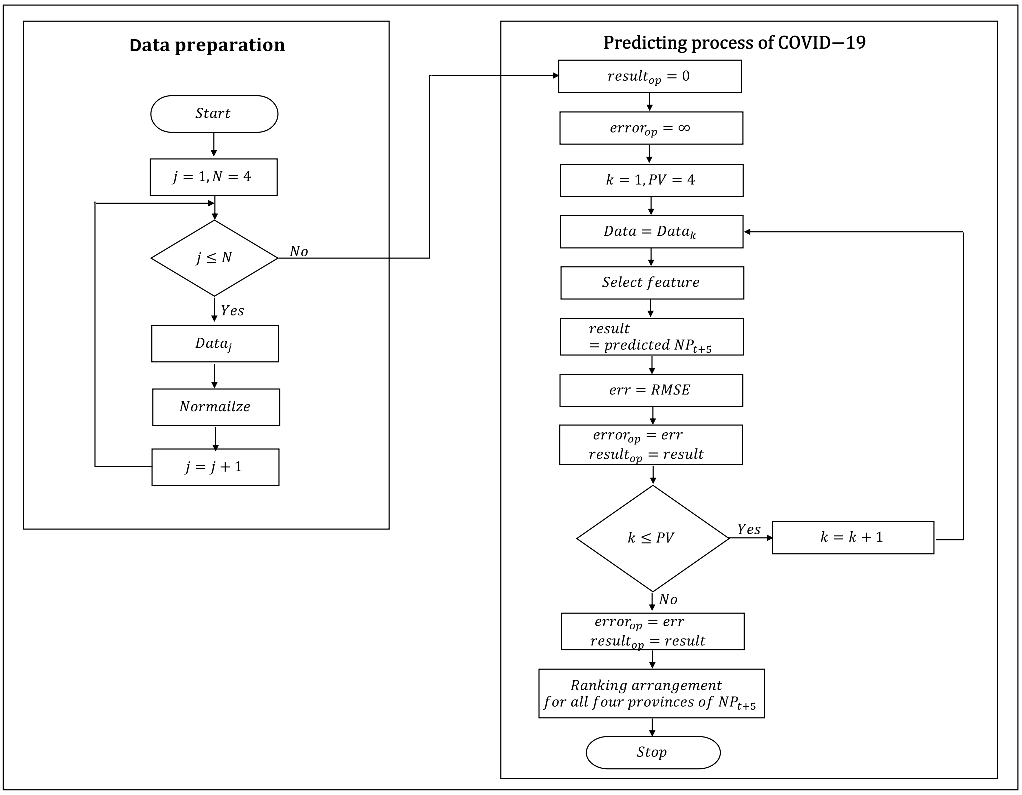

Figure 4. The flowchart of the proposed method for predicting COVID-19 new patients for 5 days ahead, using provinces as case studies

The prediction procedure begins by utilizing the dataset from each province, denoted as $Data = Data_k$ for $k = 1, 2, \ldots, PV$. Relevant features for forecasting new COVID-19 cases are extracted from $Data$ using various feature selection techniques. During each iteration, the data $Data_k$ is input into supervised regression models to estimate the number of new patients expected over the next five days. The predicted values and their corresponding root mean square errors (RMSE) are stored in $result$ = predicted NP_{t+5} and $err = RMSE$, respectively. The prediction yielding the lowest RMSE is retained as the initial optimal outcome, labeled $result_{op}$, with its associated error stored as $error_{op}$. In cases where the total number of provinces exceeds four, the final prediction output is based on the value stored in $result_{op}$.

For each province, we utilize the optimal prediction results $result_{op}$ to rank the COVID-19 new patients at $t + 5$ in order from 1 to 4.

The prediction process combines feature selection with supervised learning models for regression, necessitating specific parameter settings as shown in Tables 4 and 5.

Table 4. Parameter settings for the feature selection methods

Feature selection	Parameter setting
FS	-
BE	-
GA	Population size $= 10$
	Maximum number of iterations $= 100$
	Probability of crossover $= 0.5$
	Probability of mutation $= \frac{1}{\sigma}$
	where σ is the total number of features in a chromosome

Table 5. Configuration of parameters for supervised regression models

Model	Parameter setting
MLR	-
MLP-NN	Acceptable error $= 0.0001$
	Learning rate $= 0.01$
	Maximum number of iterations $= 200$
	Number of hidden layers $= 1$
	Number of hidden nodes $= m = \lceil \frac{p+1}{2} \rceil + 1$
	where p is the total number of input nodes
SVR	$C = 0$
	$\varepsilon = 0.0001$
	Kernel type = dot product (linear)
	Maximum number of iterations $= 100,000$

3. RESULTS AND DISCUSSION

This study utilizes data collected from four provinces. To develop and test the models, 80% of the data, arranged chronologically, was allocated for training, while the remaining 20% served as the test set. Four commonly used evaluation metrics were applied to assess the accuracy of the predictions: RMSE, MAE, MAPE, and the coefficient of determination (R^2). The outcomes demonstrate that careful model selection and performance evaluation are essential for accurately forecasting new COVID-19 cases, as they reveal both the strengths and limitations of each modeling approach.

The experiments were conducted in three distinct phases using the same dataset. In the first phase, data from each province were used to forecast the number of new cases over a five-day horizon using three supervised regression models: MLR, MLP-NN, and SVR. Among these, SVR generally yielded superior performance across most provinces, indicating that nonlinear models may offer advantages for this data type.

The second phase introduced feature selection by applying three techniques such as FS, BE, and GA, to identify the most relevant features. These selected features were then used with the same three regression models to predict new cases over the next five days. This step aimed to improve performance by eliminating irrelevant or redundant information.

In the third and final phase, the optimal prediction results obtained from the second phase were used to rank the provinces based on the projected number of new cases over the following five days.

During the first phase, no feature selection methods were applied. The models directly utilized the raw features from each province's dataset to perform five-day-ahead predictions using the three supervised learning algorithms.

Table 6 demonstrates that SVR consistently yields a lower RMSE than MLR and MLP-NN, particularly in Songkhla and Pattani provinces, indicating that SVR effectively captures the data's underlying non-linear relationships. In contrast, MLP-NN performs best in Yala province, where data patterns appear more complex.

Table 6. The COVID-19 new patient predicting results of each province using three models

Model		Province Songkhla	Pattani	Yala	Narathiwat
MLR	RMSE	432.5	203.54	177.101	179.124
	MAE	394.631	196.974	170.552	174.316
	MAPE	342.31	286.52	279.05	631.08
	R^2	0.044	0.772	0.684	0.681
MLP-NN	RMSE	186.511	247.583	**106.858**	170.837
	MAE	141.807	239.562	56.132	158.609
	MAPE	95.30	334.20	93.01	650.24
	R^2	0.118	0.598	0.426	0.014
SVR	RMSE	**178.71**	**116.839**	113.108	**152.22**
	MAE	149.904	103.093	99.433	145.019
	MAPE	146.61	181.19	231.73	577.26
	R^2	0.329	0.571	0.182	0.687

The results reveal that each model has a different performance across the four provinces, likely reflecting the varied nature of data across these regions. MLP-NN performs best for Yala province, suggesting that this model may better capture the complex patterns in the data specific to Yala province. In contrast, SVR outperforms the other models in Songkhla, Pattani, and Narathiwat provinces, indicating it might be more suited to regions with data characteristics favoring non-linear relationships. This implies that local conditions, such as population movement, healthcare interventions, or reporting practices, may influence model performance. In the second phase, three feature selection methods were combined with three supervised regression models, resulting in nine distinct method combinations for forecasting the number of new patients. Table 7 presents the best-performing method combination identified for each province. The addition of feature selection improves prediction accuracy, emphasizing the importance of selecting relevant features to reduce noise and computational overhead. Our findings align with prior studies by Shaban *et al.* [7] and Pourhomayoun and Shakibi [10] which demonstrated that feature selection techniques improve model performance without compromising prediction accuracy. Similarly, Pudjihartono *et al.* [33], as well as Theng and Bhoyar [34], emphasized that feature selection enhances generalization by eliminating irrelevant features, reducing noise, and mitigating the risk of overfitting. The five-day-ahead prediction results for each province, obtained after applying feature selection techniques before training the three supervised regression models, exhibited lower RMSE values compared to those generated in the first phase without feature selection. The results demonstrate that applying feature selection techniques improves model performance by focusing on relevant features, reducing noise, and thus lowering the RMSE values across all provinces. For example, using FS with MLP-NN in Songkhla province achieves an RMSE of 51.5, the lowest for that region, confirming the value of reducing dimensionality in feature selection. Similarly, BW with MLP-NN performs best in Pattani province, reflecting the benefit of BE in this region. In Yala and Narathiwat provinces, SVR combined with FS or BW achieves the lowest RMSE, reinforcing the strength of SVR in non-linear modeling when relevant features are carefully selected. These results highlight the importance of tailoring the model-feature selection strategy to each region's unique data characteristics.

Table 7. The five-day forecasts of new patient cases in each province generated using a combination of feature selection methods and supervised regression models

Province	Feature selection	Supervised learning model	RMSE	MAE	MAPE	R^2
Songkhla	FS	MLP-NN	51.506	35.689	15.07	0.931
Pattani	BW	MLP-NN	50.048	36.772	41.79	0.822
Yala	FS	SVR	46.391	40.389	84.40	0.729
Narathiwat	BW/FS	SVR	46.210	35.201	108.43	0.698

For the third phase, we propose to use the optimal prediction results from the second phase to take new patient ranking of four ranks from all four provinces. The ranking analysis is valuable for identifying regions with higher risks of increasing new patients, enabling healthcare authorities to prioritize interventions. Our findings show that 62.96% of predictions achieved full ranking accuracy, highlighting the robustness of the proposed methodology. However, occasional inaccuracies suggest the need for further refinement. This highlights the inherent challenge in forecasting, given the unpredictable nature of the COVID-19 pandemic. Factors such as sudden outbreaks, policy changes, or new variants may impact forecasting accuracy. The ranking results for predicting new patients of all four provinces are shown as Table 8.

Table 8. The ranking results for predicting new patients of all four provinces

The number of times as the correct ranking	The number of times (times)	As a percentage
Zero correct rank	1	1.85
One correct rank	1	1.85
Two correct ranks	17	31.48
Three correct ranks	1	1.85
Four correct ranks	34	62.96
Total	54	100

This study utilized daily data from provinces in Thailand, which included the number of new cases, cumulative cases, active cases, recovered individuals, deaths, and the number of people who received one, two, or three doses of the COVID-19 vaccine. These data were sourced from the Department of Disease Control, Ministry of Public Health, Thailand. For each province, 275 data points were initially used, with 80% allocated for training the models and 20% reserved for testing. The optimal prediction model was determined based on the results with the lowest RMSE value.

In the first experimental phase, where all features at the monitoring station were included without prior selection of features, SVR produced the most accurate forecasts for the three provinces, i.e., Songkhla, Pattani, and Narathiwat provinces. In contrast, MLP-NN achieved the best performance for Yala province.

In the second phase, after applying feature selection techniques to retain only the most relevant features before feeding the data into the supervised learning models, it was observed that prediction performance improved across all three models.

This improvement suggests that feature selection is critical in enhancing prediction accuracy, especially in multi-dimensional datasets. FS and BW consistently improve RMSE across regions, confirming that eliminating irrelevant or redundant features leads to better generalization of the models. This finding aligns with prior studies emphasizing the value of feature selection in predictive modeling. However, some limitations remain in the study. First, the dataset size, with only 275 observations per province, may restrict the generalizability of the models to other regions or future scenarios. Second, while RMSE is the primary criterion for selecting optimal predictions, other metrics such as MAE or MAPE might provide additional insights into model performance, especially for regions with high variability. Future work could explore incorporating ensemble learning techniques to combine predictions from multiple models, potentially improving forecasting robustness. Additionally, expanding the dataset by including data from more provinces or longer timeframes could enhance the model's performance and applicability to other contexts.

From RMSE in Table 7, the results from using FW and BE can also be used depending on the data of each province. This flexibility in model selection highlights the adaptability of the approach, allowing healthcare practitioners to select models based on specific regional needs. The findings provide valuable insights for policymakers, who can leverage these predictions to allocate resources more efficiently and develop targeted interventions to mitigate the impact of COVID-19.

4. CONCLUSION

From the data we studied, we successfully predicted the number of new COVID-19 patients and ranked the number of patients by province, focusing on the province where we live and nearby provinces. Our study shows that the predictive models we implemented, while designed for short-term forecasts, closely align with real data, demonstrating their effectiveness. This suggests that the mathematical models we used are practical tools for anticipating trends in new patient numbers. By providing early warning signals, these models equip healthcare authorities with the ability to prepare in advance, allocate resources efficiently, and implement timely interventions to mitigate the impact of future outbreaks.

Our approach is not limited to these specific provinces or even to COVID-19; the methodology can be extended to other provinces and adapted to forecast new outbreaks of different infectious diseases. This flexibility ensures that our work has broad applicability in public health planning. Further applications could include supporting regional decision-makers in tailoring healthcare strategies or integrating the models with real-time surveillance systems to enhance outbreak monitoring.

While the study demonstrates promising results, future research could extend these models by incorporating additional variables, such as mobility patterns or environmental factors, to improve prediction accuracy. Exploring ensemble models could also enhance forecast robustness. Ultimately, our findings contribute to the growing body of research on epidemic forecasting, demonstrating how mathematical models can serve as vital tools in community-level outbreak management and preparedness.

ACKNOWLEDGMENTS

The authors are very grateful to the experts for their appropriate and constructive suggestions to improve this article.

FUNDING INFORMATION

The authors state no funding was involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Chadaphim Photphanloet	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	
Sherif Eneye Shuaib				✓		✓			✓	✓				
Siriprapa Ritraksa	✓			✓	✓	✓	✓	✓	✓	✓				
Pakwan Riyapan	✓			✓		✓	✓		✓	✓		✓	✓	

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject Administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding Acquisition
Fo	: **Fo**rmal Analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

REFERENCES

[1] M. Khan *et al.*, "COVID-19: A global challenge with old history, epidemiology, and progress so far," *Molecules*, vol. 26, no. 39, pp. 1–25, Dec. 2020, doi: 10.3390/molecules26010039.

[2] World Health Organization, "Update on COVID-19 epidemiological update – 12 April 2024," Apr. 12, 2024. [Online]. Available: https://www.who.int/publications/m/item/covid-19-epidemiological-update-edition-166. (Accessed: Nov. 30, 2024).

[3] S. Sudsawasd, T. Charoensedtasin, N. Laksanapanyakul, and P. Pholphirul, "Modelling the overall impacts of COVID-19 on the Thai economy," *Cogent Economics & Finance*, vol. 11, no. 2, pp. 1–22, 2023, doi: 10.1080/23322039.2023.2242171.

[4] N. A. F. A. Aniqoh, A. Z. Nihayah, and F. Amalia, "The role of digital banking industry towards consumer behavior during the COVID-19," *Journal of Digital Marketing and Halal Industry*, vol. 4, no. 2, pp. 75–88, 2022, doi: 10.21580/jdmhi.2022.4.2.13378.

[5] A. S. Weiler and B. Gilitwala, "Why Bangkokians use online food delivery services after COVID-19 restrictions have been lifted," *Rajagiri Management Journal*, vol. 18, no. 2, pp. 151–166, 2024, doi: 10.1108/RAMJ-08-2023-0244.

[6] Statista, "Number of novel coronavirus (COVID-19) confirmed, recovered, and death cases in Thailand," Mar. 17, 2024. [Online]. Available: https://www.statista.com/statistics/1099913/thailand-number-of-novel-coronavirus-cases/. (Accessed: Nov. 30, 2024).

[7] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, "A new COVID-19 patients detection strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier," *Knowledge-Based Systems*, vol. 205, pp. 1-18, Oct. 2020, doi: 10.1016/j.knosys.2020.106270.

[8] L. Sun *et al.*, "Adaptive feature selection guided deep forest for COVID-19 classification with chest CT," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2798–2805, Aug. 2020, doi: 10.1109/JBHI.2020.3019505.

[9] B. Abraham and M. S. Nair, "Computer-aided detection of COVID-19 from X-ray images using multi-CNN and Bayesnet classifier," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1436–1445, Oct. 2020, doi: 10.1016/j.bbe.2020.08.005.

[10] M. Pourhomayoun and M. Shakibi, "Using artificial intelligence for medical condition prediction and decision-making for COVID-19 patients," in *Advances in Computer Vision and Computational Biology: Proceedings from IPCV'20, HIMS'20, BIOCOMP'20, and BIOENG'20*, Cham, Switzerland: Springer, Aug. 2021, pp. 617–624, doi: 10.1007/978-3-030-71051-4_49.

[11] K. C. Wong, Y. Xiang, and H. C. So, "Uncovering clinical risk factors and prediction of severe COVID-19: A machine learning approach based on UK biobank data," *MedRxiv*, vol. 7, no. 9, Sep. 2021, doi: 10.2196/29544.

[12] L. Yan *et al.*, "Prediction of criticality in patients with severe COVID-19 infection using three clinical features: A machine learning-based prognostic model with clinical data in Wuhan," *MedRxiv*, vol. 2020, no. 2, pp. 1–15, Mar. 2020, doi: 10.1101/2020.02.27.20028027.

[13] H. Yao, Y. Lu, Z. He, and L. Sun, "Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests," *Frontiers in Cell and Developmental Biology*, vol. 8, no. 683, pp. 1–12, Jul. 2020, doi: 10.3389/fcell.2020.00683.

[14] C. Hu, X. Li, J. Wang, and Z. Zhang, "Early prediction of mortality risk among patients with severe COVID-19, using machine learning," *International Journal of Epidemiology*, vol. 49, no. 6, pp. 1918–1929, Sep. 2020, doi: 10.1093/ije/dyaa171.

[15] T. Chen, J. Yang, L. Han, and Q. Zhao, "Clinical characteristics of 113 deceased patients with coronavirus disease 2019: Retrospective study," *BMJ*, vol. 368, pp. 1–12, Mar. 2020, doi: 10.1136/bmj.m1091.

[16] L. Sun *et al.*, "Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19," *Journal of Clinical Virology*, vol. 128, pp. 1–10, Jul. 2020, doi: 10.1016/j.jcv.2020.104431.

[17] C. An, Y. Li, H. Zheng, and K. Kim, "Machine learning prediction for mortality of patients diagnosed with COVID-19: A nationwide Korean cohort study," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, Oct. 2020, doi: 10.1038/s41598-020-75767-2.

[18] R. Zagrouba *et al.*, "Modelling and simulation of COVID-19 outbreak prediction using supervised machine learning," *Computer, Materials & Continua*, vol. 66, no. 3, pp. 2397–2407, 2021, doi: 10.32604/cmc.2021.014042.

[19] Y. Jenny, A. A. Soltan, and D. A. Clifton, "Machine learning generalizability across healthcare settings: Insights from multi-site COVID-19 screening," *NPJ Digital Mmedicine*, vol. 5, no. 1, pp. 69–75, 2022, doi: 10.1038/s41746-022-00614-9.

[20] S. Harvineet, V. Mhasawade, and R. Chunara, "Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database," *PLoS Digital Health*, vol. 1, no. 4, pp. 1-17, 2022, doi: 10.1371/journal.pdig.0000023.

[21] J. Yang *et al.* "Generalizability assessment of AI models across hospitals in a low-middle and high income country," *Nature Communications*, vol. 15, no. 1, pp. 8270–8280, 2024, doi: 10.1038/s41467-024-52618-6.

[22] T. K. Times, "COVID-19 cases see a sharp jump in Thailand as doctors warn of spread in 'peak season'," Jun. 11, 2024. [Online]. Available: https://www.khmertimeskh.com/501504144/covid-19-cases-see-sharp-jump-in-thailand-as-doctors-warn-of-spread-in-peak-season/. (Accessed: Nov. 30, 2024).

[23] Department of Disease Control, "Thailand COVID-19 Dashboard." [Online]. Available: https://ddc.moph.go.th/covid19-dashboard/. (Accessed: Oct. 19, 2024).

[24] Esri, "ArcGIS Pro," Esri, Redlands, CA, USA. [Online]. Available: https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview.

[25] A. Jadhav and S. K. Shandilya, "Towards effective feature selection in estimating software effort using machine learning," *Journal of Software: Evolution and Process*, vol. 36, no. 5, 2024, doi: 10.1002/smr.2588.

[26] F. Bagherzadeh, M. J. Mehrani, M. Basirifard, and J. Roostaei, "Comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance," *Journal of Water Process Engineering*, vol. 41, 2021, doi: 10.1016/j.jwpe.2021.102033.

[27] C. Darwin, *On the Origin of Species by Means of Natural Selection*. Good Press, 2023.

[28] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: past, present, and future," *Multimedia Tools and Applications*, vol. 80, pp. 8091–8126, 2021, doi: 10.1007/s11042-020-10139-6.

[29] A. Mohammadzadeh, M. H. Sabzalian, O. Castillo, R. Sakthivel, F. F. M. El-Sousy, and S. Mobayen, "Multilayer Perceptron (MLP) Neural Networks," in *Neural Networks, Learning Algorithms in MATLAB*, Cham, Switzerland: Springer, pp. 5-21, 2022, doi: 10.1007/978-3-031-14571-1_2.

[30] C. Campbell and Y. Ying, *Learning with Support Vector Machines*, 1st ed., Cham, Switzerland: Springer, 2022, doi: 10.1007/978-3-031-01552-6.

[31] X. Li and X. Zhang, "A comparative study of statistical and machine learning models on carbon dioxide emissions prediction of China," *Environmental Science and Pollution Research*, vol. 30, pp. 117485–117502, 2023, doi: 10.1007/s11356-023-30428-5.

[32] P. Cuevas-Delgado, D. Dudzik, V. Miguel, S. Lamas, and C. Barbas, "Data-dependent normalization strategies for untargeted metabolomics—a case study," *Analytical and Bioanalytical Chemistry*, vol. 412, no. 27, pp. 6391–6405, 2020, doi: 10.1007/s00216-020-02594-9.

[33] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, pp. 1-17, 2022, doi: 10.3389/fbinf.2022.927312.

[34] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: a survey of more than two decades of research," *Knowledge and Information Systems*, vol. 66, no. 3, pp. 1575–1637, 2024, doi: 10.1007/s10115-023-02010-5.

BIOGRAPHIES OF AUTHORS

Chadaphim Photphanloet is an assistant professor at the Department of Mathematics and Computer Science, Prince of Songkla University, Thailand, where she has been a faculty member since 2020. Chadaphim Photphanloet graduated with second-class honors. B.Sc. degree in Mathematics from Prince of Songkla University, Thailand, in 2014; M.Sc. in Applied Mathematics and Computational Science from Chulalongkorn University, Thailand, in 2016; and Ph.D. in Applied Mathematics and Computational Science from Chulalongkorn University, Thailand, in 2020. Her research interests are primarily in machine learning, data mining, mathematical modeling, and statistical analysis, particularly in applying mathematical skills to real-world situations. She can be contacted at email: chadaphim.p@psu.ac.th.

Sherif Eneye Shuaib completed his B.Tech in Mathematics with Computer Science from the Federal University of Technology, Minna, Nigeria, in 2014 and his M.Sc. in Applied Mathematics from Prince of Songkla University, Thailand, in 2020. He has also worked as an editor at the Publication Unit, Prince of Songkla University, Pattani Campus, Thailand. His research interests lie in mathematical biology, disease modeling, and data analytics. He can be contacted at email: sherifeneyeshuaib@gmail.com.

Siriprapa Ritraksa is an Assistant Professor at the Department of Mathematics and Computer Science, Prince of Songkla University, Thailand, where she has been a faculty member since 2012. Siriprapa Ritraksa graduated with a first-class honors B.Sc. degree in Applied Mathematics from Prince of Songkla University, Thailand in 2009 and M.Sc. in Applied Mathematics and Computational Science from Chulalongkorn University, Thailand in 2011 and completed her Ph.D. in Applied Mathematics and Computational Science from Chulalongkorn University, Thailand in 2021. Her research interests are primarily in mathematical modeling, simulation, and visualization of blood vessels and image processing. She can be contacted at email: siriprapa.r@psu.ac.th.

Pakwan Riyapan is an Associate Professor at the Department of Mathematics and Computer Science, Prince of Songkla University, Thailand, where she has been a faculty member since 2005. She graduated with a second-class honors B.Sc. in Mathematics from Prince of Songkla University, Thailand, in 2003, and an M.Sc. in Mathematics from Kasetsart University, Thailand in 2005. She then received an M.Sc. in Applied Mathematics from Heriot-Watt University, UK, in 2008 and completed her Ph.D. in Applied Mathematics from the University of Leeds, UK, in 2013. Her research interests are pattern formation, mathematical modeling, mode interactions, and differential equations. She can be contacted at email: pakwan.r@psu.ac.th.