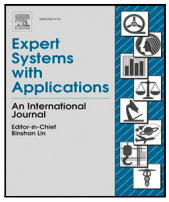
Expert Systems With Applications 237 (2024) 121438

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



A hybrid EMD-GRNN-PSO in intermittent time-series data for dengue fever forecasting

Wiwik Anggraeni [a,*], Eko Mulyanto Yuniarno [b], Reza Fuad Rachmadi [b], Surya Sumpeno [b], Pujiadi Pujiadi [c], Sugiyanto Sugiyanto [d], Joan Santoso [e], Mauridhi Hery Purnomo [b]

[a] Department of Information Systems, Institut Teknologi Sepuluh Nopember, Surabaya, East Java, Indonesia
[b] Department of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, East Java, Indonesia
[c] Dengue Fever Eradication Malang Regency Public Health Office, Malang, East Java, Indonesia
[d] Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Central Java, Indonesia
[e] Department of Information Technology, Institut Sains dan Teknologi Terpadu Surabaya, Surabaya, East Java, Indonesia

ARTICLE INFO

Keywords:
Hybrid method
Empirical mode decomposition
Regression Neural Network
Particle Swarm Optimization
Intermittent
Forecasting

ABSTRACT

Accurate forecasting of dengue cases number is urgently needed as an early warning system to prevent future outbreaks. However, forecasting dengue fever cases with intermittent data characteristics are still rare. In addition, good forecasting accuracy for intermittent data is also challenging to obtain. A hybrid Empirical Mode Decomposition (EMD), Generalized Regression Neural Network (GRNN), and Particle Swarm Optimization (PSO) were proposed to solve the problem. First, data preprocessing is done to ensure that the data is ready for further processing and has a relationship with the dengue fever case number. Second, the decomposition extracts the non-stationarity and nonlinearity patterns of each predictor variable and transforms them into several intrinsic mode functions (IMFs). Third, using various data training and testing ratios and cross-validation, the IMFs of each predictor variable were trained with GRNN to capture the best model of dengue fever cases forecasting. PSO algorithm is used to find the optimal parameters of GRNN so that the parameter searching process is more efficient and accuracy increases. Finally, to see the robustness and effectiveness of the proposed hybrid approach, the forecasting performance of the proposed hybrid model was assessed on 21 datasets with different intermittent conditions, data periods, geographical conditions, diverse numbers, and ranges of data. This approach also compared the comparative benchmark models, using MSE, MAE, and SMAPE as evaluation indicators. The Diebold–Mariano test and the pairwise sample t-test show that the proposed model is more reliable in handling intermittent data.

1. Introduction

Dengue Fever (DF) is a fatal disease that rapidly spreads globally and exhibits increased mortality (WHO, 2017). Over the past 50 years, the dengue case number has increased 30-fold. In addition, it occurs in more than 100 countries (WHO, 2016), with an estimated 50–100 million infections occurring every year as defined in WHO (2020). Forecasting the number of cases can be one of the primary considerations for the early detection of disease development and is expected to decrease the mortality rate (Siriyasatien et al., 2018). As Chen et al. (2019) has presented, forecasting results are also necessary to prevent outbreaks.

In Indonesia, DF usually occurs in the rainy season. No DF cases were noted in some periods in the dry season (Malang, 2020). The situation caused data used in this study have different characteristics from those of the data in previous studies. It contains a lot of zero values, small values, and intermittent characteristics. Forecasting data with zero values is rarely performed in the health sector but frequently in the supply chain. Unfortunately, DF case data in some geographical areas contain zero values and have intermittent characteristics, especially during the dry season. In the case of DF, intermittent time series are characterized by multiple non-case intervals. The data available at the Public Health Office, on the other hand, are too scarce to be considered. If these forecasting data are combined, the result will be less accurate. It renders the process of forecasting data with zero and small values difficult, as defined by Mussumeci and Codeço (2020). These findings have later become research challenges in the field of health data forecasting. In addition, the climate and number of cases

* Corresponding author.
E-mail addresses: wiwik@is.its.ac.id (W. Anggraeni), ekomulyanto@ee.its.ac.id (E.M. Yuniarno), fuad@its.ac.id (R.F. Rachmadi), surya@ee.its.ac.id (S. Sumpeno), dinkes@malangkab.go.id (P. Pujiadi), sugiyanto@dsn.dinus.ac.id (S. Sugiyanto), joan@istts.ac.id (J. Santoso), hery@ee.its.ac.id (M.H. Purnomo).

https://doi.org/10.1016/j.eswa.2023.121438
Received 24 December 2022; Received in revised form 21 August 2023; Accepted 1 September 2023
Available online 7 September 2023

exhibit characteristics that fluctuate dramatically at certain times, and the residuals are likely to show heteroscedasticity. Previous research considers heteroscedasticity as a problem that can reduce the accuracy of forecasting.

Based on our best knowledge, there are still rare studies on DF case forecasting that involved zero values, especially with intermittent characteristics and show heteroscedasticity. The previous studies mainly used data without zero values (Mala & Jat, 2019; Shashvat et al., 2019; Tsai et al., 2018; Zhu et al., 2019). They usually used aggregate data. In Indonesia, the forecast results up to the district level are needed by the Health office to provide more appropriate policies as defined in Malang (2021). Therefore, aggregate data as in previous studies is less suitable for use in this study. In addition, this study forecasts the number of DF cases using several variables where they have different characteristics. The combination of these characteristics renders high accuracy forecasting challenging to achieve. Thus, this study proposes a DF case number forecasting system that combines decomposition methods, machine learning forecasting models, and optimization algorithms to overcome the weaknesses of existing methods and take advantage of combined forecasting.

The contribution of this study is discussed as follows:

1. This study proposes an adaptive hybrid model between decomposition, GRNN, and PSO, which is the best handling for intermittent dengue forecast cases that cannot be handled by other previous state-of-the-art in regression.
2. Compared to the previous, this model can successfully copes with forecasting data with intermittent characteristics and heterogeneity in data patterns and be proven better through in-depth statistical tests.
3. The combination of GRNN and PSO allows GRNN parameters to be determined automatically without trial and error so that it is not too time-consuming. The GRNN parameters also change dynamically as the model gets additional data. These conditions make GRNN and PSO hybrid models produce better performance.

The remainder of this study is organized as follows. Section 2 discusses previous related works on forecasting DF case numbers. Section 3 presents materials and methods that contain the study area and dataset, DF mathematical modeling, and methodology. Section 4 presents the experiments and results under various conditions, which indicate the reliability of the proposed hybrid model. Section 5 provides the discussion. Finally, Section 6 provides the summary and concluding remarks.

2. Related works

2.1. Dengue fever cases number forecasting approaches

Various models and methods have been introduced to forecast DF incidence. Some combine models and strategies with the expectation that such combination can increase forecasting accuracy, and some only use one technique. Previous studies had used the time series model, statistics, and machine learning approach as presented in Dos Santos et al. (2019). The regression model has been employed to forecast the number of DF cases (Mala & Jat, 2019). It uses data from Google trends and related agencies. Time series models, such as ARIMA, have also been developed to forecast DF incidence and analyze DF cases' temporal patterns (Cortes et al., 2018; Gabriel et al., 2019). In previous work, ARIMA is combined with Exponential Smoothing (Shashvat et al., 2019) and Fuzzy (Anggraeni et al., 2019) to forecast the number of DF cases. The decomposition based on loess models has also been employed to analyze the seasonal impacts between the climatic factors and DF incidence. These factors are subsequently used in forecasting the number of DF incidents (Anggraeni et al., 2019). These time series approaches exhibit good accuracy. However, they involved only one variable, namely, DF cases. The ARIMA approach is usually used when forecasting data using only one variable (Liu et al., 2022; Lu et al., 2021). When the variable is more than one, ARIMAX can be used. ARIMA development called SARIMA has also been combined with decomposition for forecasting (Wen et al., 2022). In addition to the statistical models, predictions of the DF propagation are also carried out using mathematical models, as was done by Abeyrathna et al. (2016).

In recent years, forecasting DF incidence using a machine learning approach has also been performed (Carvajal et al., 2018; Hoang Cao et al., 2018). Moreover, machine learning is used to reveal DF cases' temporal patterns by involving climatic factors (Carvajal et al., 2018). Most of these studies use the Random Forest, Neural Network, and Naive Bayes. A regression model called the Ensemble Penalized Regression algorithm was used to predict several diseases, including dengue (Chen et al., 2018; Guo et al., 2019). However, some of these studies state that the model exhibits poor performance for out-of-sample data (Jain et al., 2019), and bias (Carvajal et al., 2018). In addition, some machine learning approaches are stated to require a lot of data as defined in Ozer et al. (2021).

Currently, research using EMD has been widely done. However, they are more focused on using EMD for feature recognition on different time series frequency domains (Lv & Wang, 2022). Meanwhile, linear and nonlinear pattern recognition is fundamental and meaningful for building models in forecasting. Thus, we need a decomposition method that can extract linear and nonlinear patterns simultaneously so that the accuracy of the model increases. As Liu et al. (2022) has stated, EMD is to be able to improve forecasting accuracy. Most of the relationships between them as variable predictors are also ignored. Whereas. the processing of each part of the decomposition result simultaneously can be influential and valuable for the final forecasting (Lv & Wang, 2022). In addition, the EMD is an adaptive decomposition commonly used in signal processing. EMD has become an effective tool for overcoming prediction difficulties caused by the non-stationarity of time series data in signal processing.

In addition, as Rooki (2016) has stated, GRNN is a method that has high speed in the training process. The performance of GRNN has been found to be superior to that of the radial basis function, multiple linear regression, and multiple nonlinear regression models (Ghritlahre & Prasad, 2018; Li et al., 2018). Furthermore, it has been proven effective in solving nonlinear problems. However, GRNN performs poorly if the inputs are numerous, redundant, and unmodified (Rooki, 2016). This condition led to the modification of the input variables using EMD. Besides, GRNN has parameters whose values can change. The value of this parameter can affect the performance of the forecasting results obtained. Rooki (2016) states that getting the best forecasting performance requires trial and error. It requires much time for the same data. To speed up obtaining optimal parameters, GRNN is combined with PSO. Hybrid approaches can combine the strengths of different techniques (Yang & Li, 2023). GRNN has proven effective in solving non-linear problems but takes time to find optimal parameters, while PSO has the advantage of being fast in achieving convergence in parameter optimization, as stated in Wang et al. (2019). In addition, PSO is an effective evolutionary algorithm and has been widely used in optimization problems (Yang & Li, 2023). Kuranga and Pillay (2022) also claimed that PSO had been successful in forecasting. PSO-based models were more appropriate to solve real-world problems, easy to implement, and had better tuning parameters. Because it is adaptive in tuning its parameters and flexible for all problems, PSO is considered suitable for use in intermittent data. PSO is ideal for dynamic optimization.

PSO has been proven to be superior to state-of-the-art techniques in predicting non-linear time series data as stated in Kuranga et al. (2023). Yang and Li (2023) also shows that PSO ranks highest compared to other approaches in Friedman results comparing several evolutionary algorithms on CEC2013 and CEC2017. PSO also provides optimal results from the Whale Optimizer, Grasshopper Optimization Algorithm, Gravitational Search Algorithm, and Gray Wolf Optimizer for tension spring optimization, pressure vessel optimization, and welded beam optimization issues (Yang & Li, 2023).

2.2. Characteristics of the data in dengue fever cases number forecasting

The data and its characteristics are an important component in finding forecasting models because they ensure meaningful and usable results. Forecasting will always involve previous data. Therefore, knowledge of past data patterns or characteristics is urgently needed to help the learning process (Siriyasatien et al., 2018). Thus, variations in the characteristics or patterns of data used become very important things to note.

Previous studies of DF cases used data with characteristics and tendencies (Mala & Jat, 2019; Naish et al., 2014; Tsai et al., 2018; Zhu et al., 2019). Most of these studies use aggregate data and simulation. Some previous studies have used aggregate data on monthly, and weekly and few have used daily data (Naish et al., 2014). Aggregate data on DF cases per year in the Delhi area with the seasonal pattern used by Mala and Jat (2019). A study by Siriyasatien et al. (2016) also uses aggregate data per season, namely winter, rainy, and summer from 3 provinces. Other studies use aggregate data per month in Bangkok (Jain et al., 2019), Paulo and Ribeiro Preto region (Gabriel et al., 2019), Malang Indonesia (Anggraeni et al., 2019), and Brazilian cities (Cortes et al., 2018). Aggregate data in weekly with seasonal data patterns were collected by Carvajal et al. (2018) in Manila, Chen et al. (2018) in Singapore, Guo et al. (2019) in Guangzhou, and Hoang Cao et al. (2018) in a number of places in Vietnam and Singapore. In most of these earlier studies, zero values – also known as intermittent values – were absent. Due to the usage of aggregate data, there is no zero value and the value of the case number is quite high (Chen et al., 2018; Mussumeci & Codeço, 2020).

However, the DF incidence forecasting has been stated to require zero data as defined in Shashvat et al. (2019). Therefore, aggregate data with seasonal patterns containing zero values have been used by Shashvat et al. (2019) and Husnina et al. (2019) to forecast dengue incidence. Study in Husnina et al. (2019) uses monthly aggregate data with zero values, but the main point of the study is to analyze the factors that influence the DF, not yet up to the forecasting process. Nevertheless, the proportion of zero data used is also very small as shown in the incidence rate figure of Husnina et al. (2019). Another researcher, Shashvat et al. (2019) also involved zero data in his aggregate data, but the forecasting used only involves one variable, namely DF case number and has not involved the influence of other variables. The results show that the forecasting model is less able to follow the actual data pattern. Consequently, this proposed study forecasts the number of DF cases using data with varying intermittent patterns that different proportions of zero values. This zero value appears in periods where in reality, there are no DF cases.

2.3. Predictor variables in dengue fever cases number forecasting

Most of DF cases occur during the rainy season and there are no cases during the dry season. This finding indicates that the climatic conditions and season can influence the number of DF cases as shown in Siriyasatien et al. (2018). DF forecasting is undoubtedly associated with precipitating factors. The effect of various factors on the DF incidence has been widely known. Moreover, the impact of meteorological and geographical factors on DF incidence has been described (Mala & Jat, 2019). Variations in various climatic factors effect have also been discussed by Siriyasatien et al. (2016). Zhu et al. (2019) also added the environmental factors, as well as meteorological and hydrological factors as shown in Tsai et al. (2018). Naish et al. (2014) stated that the patient's geographical condition and demographics should be involved in the DF incidence forecast. Likewise human mobility is also stated to influence the DF spread (Abeyrathna et al., 2016). Forest cover is also included as an influencing variable in Husnina et al. (2019). These studies have primarily included certain climatic variations, but not the rainfall, humidity, temperature, wind speed, mosquito density, and population density variables.

However, in the context of climate and DF, the impact of DF incidence is still debated (Carvajal et al., 2018). Unfortunately, in various locations, the influence of climatic factors can also be different (Murray et al., 2013). This is compatible with the development of a DF model due to the nature of this disease and the mosquito life cycle as defined in Dhiman et al. (2010). Such previous studies have mostly involved specific climatic variations. However, they have not involved rainfall, humidity, temperature, wind speed, wind direction, mosquito density, population density, as well as DF case number simultaneously. In addition, they also have not noticed the time lag.

Although EMD, GRNN, and PSO have each been used independently in a number of different sectors, to our knowledge, the suggested hybrid method between them has not yet been used for forecasting in the health sector, particularly in the case of DF, and has intermittent characteristics. There is a significant research gap in DF case forecasting after reviewing previous studies extensively. Hybrid models based on decomposition data processing strategy and modified time-series forecasting models are rarely developed and applied. However, these hybrid models have proven effective in addressing the forecasting difficulties caused by the high nonlinearity, non-stationarity, and uncertainty of time series data of other resources (Zhang et al., 2022). Furthermore, the hybrid of several methods is also stated to increase accuracy because it involves the strengths of each method (Lu et al., 2021). In addition, this study forecasts the number of dengue fever cases involving climate factors, mosquito density, and population density by paying attention to time lag factors. This is consistent with Chen et al. (2019) definition that the accuracy of forecasting can be impacted by a variety of climate factors. Forecasting can be carried out under current actual circumstances and exhibits high accuracy thanks to the use of numerous variables, as demonstrated in Atchadé and Sokadjo (2021).

3. Materials and methods

3.1. Study area and dataset

The area used in this study is Malang Regency. The Malang Regency region had the highest number of dengue fever cases in Indonesia for several years. It consists of 33 sub-districts that are at different geographical heights. The different geographical conditions affect the DF cases number. The DF cases number in mountainous areas is lesser than those of coastal and forest areas as stated in Chowell et al. (2011). For this reason, these 33 sub-districts were further grouped into sub-districts located in the lowland, medium land, and highland. The area in the lowland group has an altitude of 0–50 meters above sea level (masl), the medium land group, 251–500 masl, and the highland group has > 500 masl (Malang, 2021). In each plateau, a sample was selected from several sub-districts. The sample selection was performed using a correlation test between sub-districts.

This study involved data on temperature, wind speed, humidity, rainfall, population density, mosquito density, and the number of DF cases. Temperature, wind speed, humidity, and rainfall data were recovered from the BMKG station (https://karangploso.jatim.bmkg.go.id). Meanwhile, the mosquito density, the number of DF cases, and the population data were retrieved from the Malang Public Health Office (http://dinkes.malangkab.go.id). The data involved are daily data between January 2016 to January 2020. The descriptive statistics of the data and the distribution of daily data are shown in Table 1. It can be seen that the data involved all have an abnormal distribution. The skewness and kurtosis values shown in Table 1 illustrate this condition. The summary indicates that three of the eight variables involved in this study contain zero values and null values for another two variables, namely temperature, and humidity. In addition, the data involved have different value range variations.

The dataset described in Table 1 is the dataset used for training and testing the best model. Then, to test the robustness of the model, 20 other datasets are used which are applied to 3 different scenarios with different objectives. The other 20 datasets consist of:

Table 1
Descriptive statistics of daily data based on each of the predictor variables entered for the sample area.

Descriptive Statistics								
Pred. Var.	Min	Max	Mean	Std. Dev.	Skewness		Kurtosis	
					Stats	Std.Err.	Stats	Std.Err.
DF Cases	0.000	9.000	0.179	0.608	5.830	0.063	52.346	0.125
Temperature (Celsius)	0.000	27.200	24.029	1.475	−5.888	0.063	91.474	0.125
Humidity (%)	0.000	96.000	78.789	7.970	−1.580	0.063	11.868	0.125
Rainfall (mm)	0.000	99.000	8.654	13.472	2.652	0.063	8.913	0.125
Wind speed (m/s)	0.000	5.000	0.963	0.695	0.603	0.063	1.468	0.125
Population density	2231.307	2287.526	2260.088	16.086	−0.073	0.063	−1.180	0.125
Mosquito density (%)	68.666	98.607	90.037	4.250	−0.884	0.063	1.275	0.125

1. 3 datasets of data in daily periods (each dataset represents low, medium and high plains)
2. 3 datasets of data in a period of 1 week (each dataset represents low, medium and high plains)
3. 3 datasets over a period of 2 weeks (each dataset represents low, medium and high plains)
4. 3 datasets of data in monthly periods (each dataset represents low, medium and high plains)
5. 5 datasets for daily periods for other regions, namely Donomulyo, Dau, Karangploso, Lawang, Jabung
6. 3 other datasets (secondary data) which have very different data characteristics from the others. there are negative data, a large standard deviation, and a smaller proportion of null values.

3.2. Dengue fever mathematical model

The dynamics of DF cases can be explained by dividing individuals into several groups by adopting the Kermack–McKendrick model. The groups include Susceptible (S), Infectious (I), and Recovered (R). This division of individual groups is similar to Katris (2021). The number of individuals in each group in period t is denoted by $S(t)$, $I(t)$, and $R(t)$. The change of the individual number over time $t, t+\Delta$ with $\Delta t \to 0$ is written in the Eqs. (1), (2), and (3):

$$\frac{dS}{dt} = -\lambda I S \tag{1}$$

$$\frac{dI}{dt} = \lambda I S - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I \tag{3}$$

This Kermack–McKendrick equation model assumes that the number of new individuals in the susceptible (S) group is zero and that there is no removal from each group. Thus, modifications are required to enable the model to be more in line with the sample area's facts. The increase of the case number over time is represented geometrically, which is presented in Fig. 1.

Fig. 1 shows that finding the value of the latest case number over time is equivalent to finding the point of the gradient of a tangent line $C(t)$ on the $S(t)$ curve. This tangent needed to correct at each point. It is necessary to make minimal modification and close to zero ($\Delta t \to 0$). The equation is expressed in Eq. (4), where the said value is equal to the tangent slope because of this minimal value. This addition can be modified according to the pattern of growth rate distribution. The same is true for the infectious $I(t)$ and the recovered $R(t)$ functions.

$$\frac{dS}{dt} = \lim_{\Delta t \to 0} \frac{S(t+\Delta t) - S(t)}{\Delta t} = S' \tag{4}$$

By solving Eq. (4), we finally acquired Eq. (5)

$$S' = \lambda I(t) S(t) \tag{5}$$

Let $N(t) = S(t) + I(t) + R(t)$ represent the total population and C the compartment of the total population size $N(t)$. Individuals who leave group C during period $(0, t)$ are denoted by X. The residence time X has a probability density function $F(t)$, where $F(t) = Prob[X < t]$. Suppose $G(t)$ is a survival function denoted by the $G(t) = 1 - F(t) = Prob[X < t]$. For $\tau > 0, G(t-\tau)$ is an individual infected at time $\tau > 0$ and still infectious at time $(t > \tau)$. Thus, numerous individuals are infected at time r and remain infectious at time t, as stated by $\lambda I(\tau) S(\tau) G(t-\tau)$. Thus, the SIR model in Eqs. (1), (2), and (3) turns into Eqs. (6), (7), and (8).

$$S(t) = \int_0^t \lambda I(\tau) S(\tau) G(t-\tau) d\tau \tag{6}$$

$$(t) = I_0(t) + \int_0^t \lambda I(\tau) S(\tau) G(t-\tau) d\tau \tag{7}$$

$$R(t) = I_0(t) + \int_0^t \lambda I(\tau) S(\tau)(1 - G(t-\tau)) d\tau \tag{8}$$

When the infectious period has an exponential distribution, refer to the equation $\begin{cases} F(t) = 1 - e^{-\lambda t}, t \geq 0 \\ F(t) = 0, t < 0 \end{cases}$ with the mean of the infectious period is $\frac{1}{\gamma}$. Then, we have $G(t) = e^{-\gamma t}$. Let $N(t) = N_0 e^{-\gamma t}$, then $I_0(t) = I_0(t) e^{-\gamma t}$ and $I_0\prime = -\gamma I_0(t)$ so that Eq. (7) becomes:

$$\begin{aligned} I\prime(t) &= I_0\prime(t) + \lambda I(\tau) S(\tau) - \gamma \int_0^t \lambda I(\tau) S(\tau) e^{-\gamma\tau} d\tau \\ I\prime(t) &= I_0\prime(t) + \lambda I(t) S(t) - \gamma\ (I(t) - I_0(t)) \\ I\prime(t) &= \lambda I(t) S(t) - \gamma\ (I(t)) \end{aligned} \tag{9}$$

Using the same method, Eqs. (6) and (8) becomes:

$$\begin{aligned} R\prime(t) &= \gamma(I(t)) \\ S\prime(t) &= \lambda I(t) S(t) \end{aligned} \tag{10}$$

If we include demographic factors, such as births, mortality, and population growth, the logistic growth function can be used to model the changes in each patient group. The logistic growth function indicates the exponential growth and population decline with the assumption that $d1 = d2 = d3 = d$ can be written as Eq. (11).

$$N\prime(t) = (b-d)\ N(t) - \frac{N^2(t)}{K} \tag{11}$$

where b and d are the birth and death rates and $(b-d)K$ is the population density.

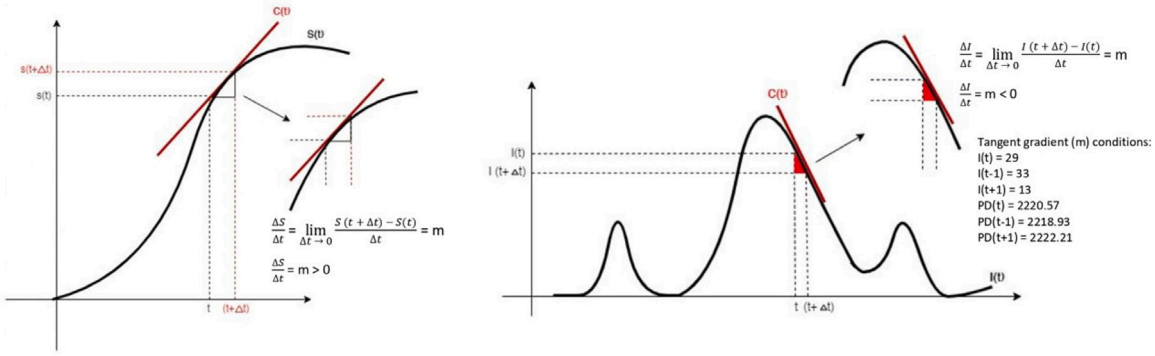


Fig. 1. The increase in the number of individuals in the susceptible (S) and infectious (I) groups over time from a geometrical point of view.

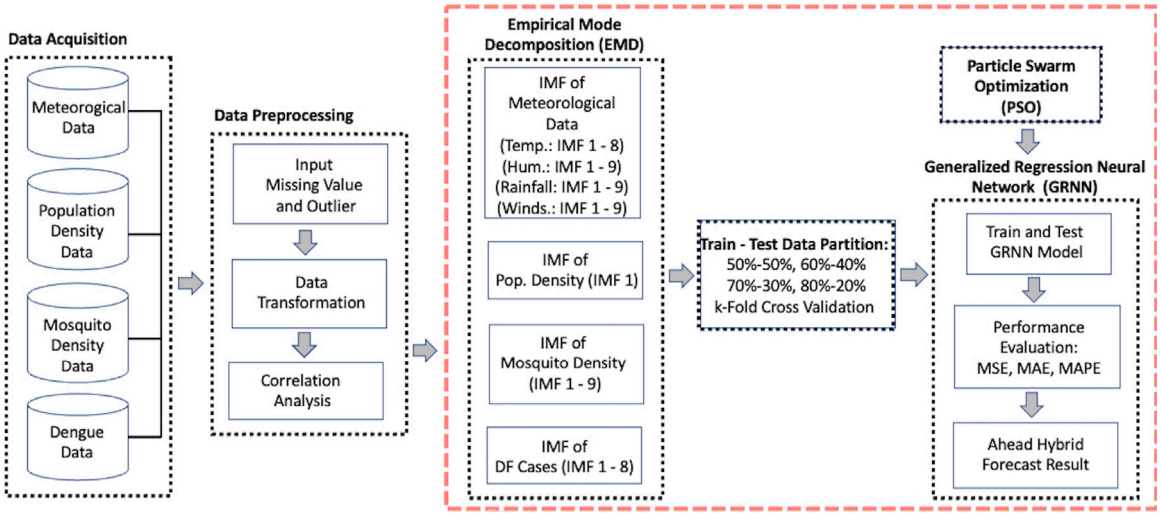


Fig. 2. The workflow of the hybrid method.

Therefore, if the logistic growth function is substituted for the model, Eq. (12) is formulated specifically for this case, where b is valued at 4.8677E−05; d, at 0.0139; and K, at 80.138.

$$\begin{aligned} S\prime(t) &= (b)\ N(t) - \frac{N^2(t)}{K} - \lambda I(t)S(t) - dS(t) \\ I\prime(t) &= \lambda I(t)S(t) - (\gamma\ + d)(I(t)) \\ R\prime(t) &= \gamma\ (I(t) - dR(t)) \\ N(t) &= S(t) + I(t) + R(t) \end{aligned} \tag{12}$$

3.3. Methodology

In this study, the predictor variables are temperature, wind speed, rainfall, humidity, mosquito density, and population density. The targeted variable is the number of DF cases. The workflow of the hybrid method is shown in Fig. 2.

3.3.1. Preprocessing data

Handling Missing Value and Outlier. The missing values and outlier data were handled using interpolation-extrapolation, linear regression, and averages of previous periods. The selection of each approach is based on the data pattern and data characteristics.

Data Transformation. Conversion of the total population and area of region data into population density is done using several formulas by involving the Exponential Population Rate equation.

Correlation Analysis. This analysis is used to determine the relationship between the targeted variable and the predictor variables. This analysis is based on an analysis of risk factors that correlate with the number of different DF cases as presented by Murray et al. (2013). Likewise, Chen et al. (2019) also stated that other combination variables could produce different resolutions. The variables involved are different from previous studies. These include temperature, wind speed, humidity, rainfall, mosquito density, population density, and the number of DF cases associated with time lags. In this study, the Spearman rank coefficient is used to identify the optimal effect lag between DF cases and predictor variables as shown in Jayaraj et al. (2019).

3.3.2. Empirical Mode Decomposition (EMD)

EMD is an adaptive method commonly used to analyze non-linear and non-stationary signals (Boudraa, 2007; Liu et al., 2022). The EMD algorithm decomposes univariate signals into a finite number of orthogonal oscillation modes called Intrinsic Mode Function (IMF) (Liu et al., 2021). Conditions that must be met include (1) the extreme number and the number of zero crossings must be equal to or different from one of them and (2) the average value of the local maxima (upper envelope) and the local minimum (lower envelope) is zero in each period.

For signal $X(t)$, EMD estimates N IMFs set $c_i(t)$ with $i = 1$ to N dan residue signal $r(t)$, so that $x(t) = \sum_{i=1}^{N} c_i(t) + r(t)$. EMD algorithms use iteration processes to obtain IMFs. A repetitive process called the filtering process is used. The steps of EMD are as follows:

1. Estimate all local minima and local maxima $x_0(t)$
2. Interpolate all local minima to get a lower envelope signal $x_{low}(t)$ and then interpolate all local maxima to estimate the upper envelope signal $x_{up}(t)$.

3. Calculate the average value between the upper and lower envelopes $m(t) = \frac{x_{low}(t)+x_{up}(t)}{2}$.
4. Extracts the average number of signals to get oscillation mode $c(t) = x(t) - m(t)$
5. If $c(t)$ meets the criteria then stop, so that we define $d(t) = c(t)$ as IMF. If it does not meet, set $x^{'(t)} = c(t)$. Repeat the process from step 1.
6. After the first IMF was obtained, the same step was applied iteratively to residuals to obtain the rest of the IMF where $h_0(t) = c(t)$ and a $r(t) = x(t) - h_0(t)$. The process will stop after the standard criteria stop being met.

All of the above processes are performed on all variable predictors. The process is summarized in the pseudocode shown in Algorithm 1.

Algorithm 1 Empirical Mode Decomposition (EMD) Algorithm

```
t ← 1
x(t) ← x_0(t)
repeat
    repeat
        x(t) = c(t)
        lminima ← getLocalMinima(x(t))
        lmaxima ← getLocalMaxima(x(t))
        lminima ← interpolation(lminima)
        x_low(t) ← getLowerEnvelope(lminima)
        x_up(t) ← getUpperEnvelope(lmaxima)
        m(t) ← (x_low(t) + x_up(t)) / 2
        c(t) ← x(t) − m(t)
    until c(t) = IMF
    t ← t + 1
    h_0(t) = c(t)
    r(t) = x(t) − h_0(t)
until hasMonotoneValue(r(t))
```

3.3.3. Data partition

Data is partitioned into training data and testing data. Training data is used to find a model, while testing data is used to find the best model. The proportion of training and testing data includes 50%:50%, 60%:40%, 70%:30%, and 80%:20%. In addition, we use 5-fold cross-validation to have a varied and robust model. The proportion that generates the smallest error is to be used in the next stage.

3.3.4. Particle swarm optimization - general regression neural network model

The General Regression Neural Network (GRNN) was developed based on the theory of intelligent statistical learning. GRNN has high computing efficiency, good regularization capabilities, and strong durability capabilities as stated in Ghritlahre and Prasad (2018). This approach avoids the complex weight-training process. The nonlinear approximation and mapping capabilities of the GRNN model are determined only by one factor referred to as the spread factor σ_j. General Regression Neural Network Model (GRNN) is based on nonlinear regression (Masikos et al., 2015). The nonlinear formula of GRNN in this study is expressed in Eq. (13), where y denotes DF cases variable; R, the predictor variable vector in the form of residual input $(r_1, r_2, \ldots, r_7)$; $E[y|R]$, the expected output of y given the input vector r; and $f(R, y)$, a density function joint probability of r and y.

$$E[y/R] = \frac{\int_{-\infty}^{\infty} y f(R, y) dy}{\int_{-\infty}^{\infty} f(R, y) dy} \tag{13}$$

The GRNN structure consists of four layers, namely the input layer, pattern layer, summation layer, and output layer. In this study, the number of predictors was 6 variables. Each predictor is decomposed with EMD so that the total obtained by the IMF is 53 IMF and 6 residues. Furthermore, these 59 data series serve as GRNN inputs. The input layer transmits the information to the pattern layer. A pattern layer is used to conduct forecasting. Then it passes through the summation layer, which consists of two neurons called S-Summation and D-Summation neurons. These two neurons in the summation layer give the following Eq. (14) dan Eq. (15).

$$S = \sum_{t=1}^{n} W_i exp[-D(r, r_t)] \tag{14}$$

$$D = \sum_{t=1}^{n} exp[-D(r, r_t)] \tag{15}$$

Output layer gives the predicted value f to input vector r as the Eq. (16), where n indicates the number of training patterns and W_i is the weight connecting the i^th neuron in the pattern layer to the summation layer and the Gaussian D function is defined as Eq. (17). P indicates the number of elements of an input vector. The terms x_j and x_{tj} represent the j^th element of r and r_t; respectively. The σ_j is the spread factor. The σ_j is the spread factor. We choose the smoothing parameter σ_j using the PSO algorithm to improve the accuracy of forecasting speedily as stated in Ghritlahre and Prasad (2018).

$$F(R) = \frac{S}{D} = \frac{\sum_{t=1}^{n} W_i exp[-D(r, r_t)]}{\sum_{t=1}^{n} exp[-D(r, r_t)]} \tag{16}$$

$$D(r, r_i) = \sum_{j=i}^{p} \left(\frac{r_j - r_{ij}}{\sigma_j}\right)^2 \tag{17}$$

In this study, the non-linear regression function F(R) was equipped with PSO-GRNN. The feasibility and effectiveness of the PSO-GRNN mathematical model depend largely on the spread factor. Clearly, the challenge of fitting a high fidelity regression function is turned into the following learning model's optimal solution as shown in Eq. (18), where $y(r_t)$ indicates the t^th actual values and f_t indicates the t^th forecasting values.

$$min \ \frac{1}{n} \sum_{t=1}^{n} \left[y(r_t) - f_t\right]^2 \tag{18}$$

To complete the optimization model and improve the accuracy of PSO-GRNN, the study proposes a dynamic PSO algorithm. The PSO algorithm is a well-known search algorithm based on collaborative searches of a collection of particles, which has advantages in search accuracy and search efficiency. To further improve search efficiency and accuracy, we adopted a PSO algorithm with inertial weights and dynamic learning factors. The purpose of this design is to complete the dynamic search of a collection of particles and weight the relationship between global search capabilities and local search capabilities to get a better optimal set of solutions.

The concept of dynamic PSO algorithms is that the position of the particle consists of various spread factors, and the learning error of the PSO-GRNN model is denoted in the fitness value. Each particle is a potential solution for the initial spread factor PSO-GRNN model. In the search process, all particles look for optimal solutions in solution space with current optimal particles and update the individual position of particles, extreme individual values, and extreme values of populations.

$$x_j(i) = x_1(1), \ x_2(1), \ \ldots, \ x_{jN}(i) \ v(i) = v_1(1), \ v_2(1), \ \ldots, \ v_{jN}(i)$$

x indicates the position of the particle, v is the velocity of the particle, i is the iteration, j denotes the particle index, and N is the number of particles.

Meanwhile, the particle status update formula is shown in Eqs. (19) and (20).

$$v_j(i) = v_j(i-1) + c_1 k_1 (P_{best,j} - x_j(i-1)) + c_2 k_2 (G_{best} - x_j(i-1)) \tag{19}$$

$$x_j(i) = v_j(i) + x_j(i-1) \tag{20}$$

j = 1, 2, …, N Represents the number of particles., $P_{best,j} = P_{best,1}$, $P_{best,2}, \ldots, P_{best,j}$ representing the best personal of j^th particle, G_{best}

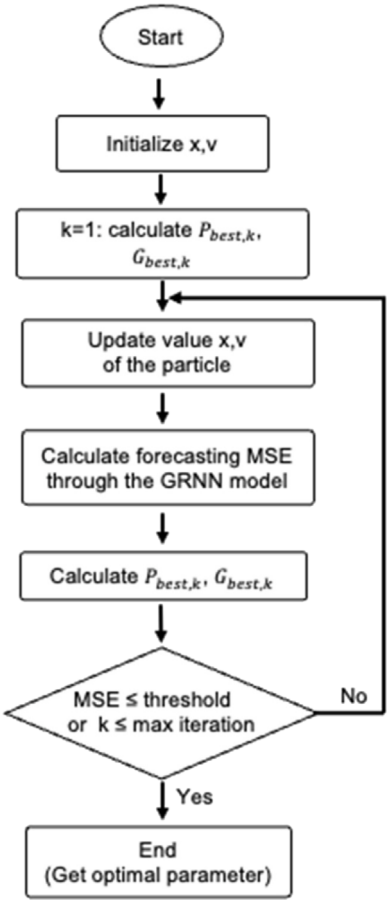


Fig. 3. The flowchart of the PSO-GRNN Process.

representing the global best of the entire population, c_1, c_2 indicate learning factor, and k_1, k_2 are random number between 0 and 1.

Eq. (19) is used to calculate the velocity of a new particle based on the previous velocity, the distance between the current position with the best particle (personal best), and the distance between the current position with the best global position. The particle then moves to a new position based on the Eq. (20). This PSO algorithm is run with a certain number of iterations until it reaches the stop criteria so that a solution will be obtained at the global best. This equation will be simulated in space with certain dimensions with a number of iterations so that in each iteration the position of the particle will increasingly lead to the intended target (minimization of fitness function). This is done until the maximum iteration is achieved or the achievement of other stop criteria. All stages of the process between PSO and GRNN can be explained in Fig. 3.

For further details, technically, the process through which PSO-GRNN is passed is indicated in pseudocode from Algorithm 2.

Algorithm 2 PSO-GRNN Algorithm

Require: dataset_train , dataset_test, max_iteration, error_threshold
repeat
  $model \leftarrow initSurogateModel()$
  $\sigma \leftarrow initSmoothFactor()$
  $x_p, v_p \leftarrow initPSOParamter()$
  $k \leftarrow 0$
  while $MSE \leq error\_threshold$ and $k \leq max\_iteration$ **do**
    $V_s, X_s, P_s, G_s \leftarrow updatePSOParameter(V_s, X_s, P_s, G_s)$
    $fitness \leftarrow calculateFitness(V_s, X_s, P_s, G_s)$
    $optimal\_particle\_position \leftarrow getOptimalParticle(fitness, V_s, X_s, P_s, G_s)$
    $MSE \leftarrow calculateMSE(optimal\_particle\_position, dataset\_train)$
    $k \leftarrow k + 1$
  end while
  $\sigma \leftarrow getOptimalSmoothFactor(P_g)$
  $performance \leftarrow calculateMSE(model, \sigma, dataset\_test)$
until hasMeetRequirement()

4. Experiments and results

To find out the best performance of the hybrid model and the robustness of this model, we devised several comparative studies and acquired scientifically sound results.

4.1. Evaluation criteria

The experiment's forecasting accuracy evaluation method is crucial. It is critical to check the validity of the forecasting model utilizing numerous test indicators (Nur Adli Zakaria et al., 2021; Wang et al., 2020). In this study, the metrics used to evaluate the forecasting model performance are Mean Absolute Error (MAE), Mean Square Error (MSE), and Symmetric Mean Absolute Percentage Error (SMAPE) (Chakraborty et al., 2019; Katris, 2021; Kuranga & Pillay, 2022; Liu et al., 2022). However, to select the best model, we used MSE (Anggraeni et al., 2019; Wang et al., 2020).

4.2. Validation method

The validation process is needed to know which methods have superior forecasting capabilities (Du et al., 2019; Nur Adli Zakaria et al., 2021; Wang et al., 2020). One of the tests that can be used in the validation process is Diebold–Mariano (DM) test (Liu et al., 2021; Wang et al., 2020; Zhang et al., 2022). In this study, the DM test was used to test the superior forecasting ability of our proposed new hybrid approach compared to its benchmarks in some time horizon ahead forecasting. DM statistics are defined by Eq. (21) (Liu et al., 2021; Wang et al., 2020; Zhang et al., 2022).

$$DM = \frac{1}{N}\sum_{i=1}^{N}(e_{i+1}^{1} - e_{i+1}^{2})/\sqrt{S^2 N} \tag{21}$$

Here, Eq. (21) shows the variance estimation of $(e_{i+1}^{1} - e_{i+1}^{2})$. Null hypothesis indicates the statement $(e_{i+1}^{1} = e_{i+1}^{2})$ and the alternative hypothesis suggests otherwise. The null hypothesis is rejected means that the predictive effects of the two forecasting models are significantly different. The DM statistic is a random variable subject to the standard normal distribution. The critical value at a level is defined as $Z\frac{\alpha}{2}$. The performance of the proposed approach is significantly different from others if the absolute value of the DM result is less than the critical value (Du et al., 2019; Liu et al., 2021; Wang et al., 2020; Zhang et al., 2022).

This study also uses a 2-way comparison test with paired sample t-test to convince the results of the DM test and show whether the proposed method is better than the others. The null hypothesis indicates that $(e_{i+1}^{1} - e_{i+1}^{2}) \geq 0$. Here, e_{i+1}^{1} represents the predictive performance of the proposed hybrid model. Rejecting the null hypothesis means that the proposed hybrid approach is not better than the others. The area where the null hypothesis is rejected is t-value $< -t_{df,\alpha}$.

4.3. Experimental results

The number of DF cases in each period is juxtaposed with the variables involved in one sample area shown in Fig. 3. Fig. 3 presents the fluctuation of daily dengue fever cases and other independent variables

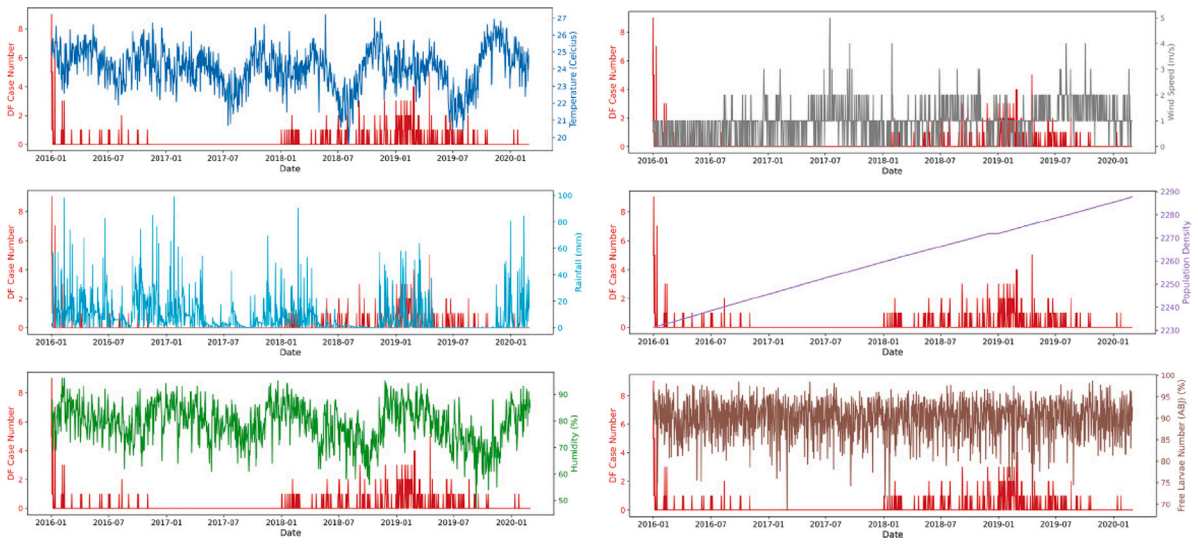


Fig. 4. Daily DF cases number and variables that affect the 2016–2020 period in Malang, each of which is: (a) DF Cases Number against Temperature, (b) DF Cases Number against Rainfall, (c) DF Cases Number against Humidity, (d) DF Cases Number against Wind Speed, (e) DF Cases Number against Population Density, (f) DF Cases Number against Mosquito Density.

during the study period, indicated by a seasonal pattern. In addition, it can be seen that the DF case number has an up and downtrend and a significant increase in December–January. This pattern repeats for the following years for the variables involved. Repetitive fluctuating patterns also occur in climate variables and mosquito density. Changes in the value of climate variables and mosquito density can affect the pattern of disease infection and risk of transmission, impacting the number of DF cases.

In this study, the experiments were conducted in 3 areas, namely lowland, middle land, and highland. The data used for the formation of the model is the selected district data on each land. The best models in each area are selected based on a combination of parameters in PSO and GRNN. The different values of GRNN parameter yield various forecasting performances as stated by Ghritlahre and Prasad (2018). Next, the model is applied to other regions on the same land with the same period. Furthermore, this model is also applied to the different areas with different periods and other secondary data. The experiment was conducted to test the robustness level of the model.

4.3.1. Experiment 1: Spearman correlation test between predictor variable and DF cases number

This experiment was conducted to determine the effect of the time lag of each variable predictor. The time lag used is 7, 10, 14, 20, and 30 days. The choice of time lag is adjusted to the incubation period for dengue disease and the mosquito's growth time that causes DF. This test is performed in all areas, and the results are shown in Table 2. Table 2 presents that the temperature variable has the highest positive correlation with DF cases number at seven-day time lag, the humidity at 14-day time lag, the wind speed at 20-day time lag, the population density at 14-day time lag. An inverse correlation, on the other hand, was observed between DF cases number and rainfall at ten days' lag and the mosquito density variable at 14 days. Each variable may have a different effect on a distant land, likewise with the most influential time lag. This study's time lag selection follows the Spearman correlation test results presented in Table 2.

4.3.2. Experiment 2: Empirical Mode Decomposition (EMD) process

Actual data will be decomposed into several different data patterns by EMD. This process is carried out on each predictor variable. Each variable has a different number of decomposition results. An example of an EMD result for variable mosquito density in one of the regions in the lowlands is shown in Fig. 4.

Fig. 4 shows that the mosquito density variable data pattern was successfully decomposed into 9 different data patterns called IMF 1 to IMF 9. This result is the extraction of data on its components, such as trends, seasonality, and others.

4.3.3. Experiment 3: Training data and testing data ratio

This experiment was designed to know the optimal proportion of training data and data testing. There are four training data and testing data ratios applied on each area, namely 50%:50%, 60%:40%, 70%:30%, and 80%:20%. In addition, we also divided training-testing data using 5-fold cross-validation. The results of the detailed experiment are referred to in Table 3. The performance testing data model for various ratios in each area in Table 3 shows that 80%:20% ratio gives the smallest average Mean Square Error. Then, this ratio is used in the following process. The best model parameters on the 80%:20% ratio in each area and the performance are shown in Table 4. For daily data, Table 4 shows that the model is not overfitting. The mean differences between MSE, MAE, and SMAPE for training data and testing data are only 0.0046, 0.0152, and 0.1425, respectively.

4.3.4. Experiment 4: Forecasting some periods ahead and performance comparison with other approaches

Forecasting for the following several periods is needed for resource planning to minimize the spread of disease (Albuquerque et al., 2022; Katris, 2021). The best model of each subsequent area is used to forecast the number of DF cases in the next periods. The next forecasting time horizon used is one month (31 days) and three months (92 days). The selection of this time horizon is based on the needs of the Region Public Health Office in the creation of action planning to prevent future outbreaks. The forecasting results performance on each time horizon between the proposed approach and several comparisons approaches in each area is shown in Table 5. The comparison methods involved here are the approaches commonly used in previous studies. These approaches include Poisson Regression (Jayaraj et al., 2019), GAM (Ebi & Nealon, 2016), ARIMAX (Albuquerque et al., 2022; Anggraeni & Aristiani, 2016; Katris, 2021), SARIMAX (Jayaraj et al., 2019; Wen et al., 2022), ARIMA-ANN (Chakraborty et al., 2019), ARIMA-SVM (Chakraborty et al., 2019), Propagation model (Abeyrathna et al., 2016), Kalman Filter (Hu et al., 2020; Jutinico et al., 2021), and GRNN as univariate methods. This metric can be calculated because fortunately, the Health Service provides the latest additional data outside the period of data used as mentioned in the material and

Table 2
Spearman correlation coefficients between DF cases and predictor variables at various time-lags in different land of Malang Regency.

Variable	Lowland					Medium land					Highland				
	7	10	14	20	30	7	10	14	20	30	7	10	14	20	30
Humidity	0.038	0.03	**0.059**	0.048	0.054	0.045	**0.048**	0.033	0.042	0.035	0.039	**0.06**	0.049	0.002	0.057
Population density	0.204	0.201	**0.204**	0.203	0.201	0.177	0.176	0.182	0.185	**0.189**	0.123	0.127	0.133	0.134	**0.138**
Mosquito density	0.003	0.013	**−0.049**	0.02	0.036	−0.001	0.005	−0.024	**−0.027**	−0.007	−0.004	−0.027	**−0.027**	0.009	0.003
Temperature	**0.028**	0.022	0.002	0.013	0.018	**0.076**	0.043	0.045	0.027	0.012	0.019	0.002	0.034	**0.044**	0.002
Rainfall	−0.034	**−0.090**	−0.028	−0.025	−0.030	−0.063	**−0.071**	−0.069	−0.047	−0.061	**−0.018**	0.01	−0.008	0.004	−0.016
Wind speed	0.023	−0.012	0.054	**0.056**	0.001	−0.008	0.027	**0.043**	0.009	0.02	−0.004	0.01	**0.041**	0.027	−0.017

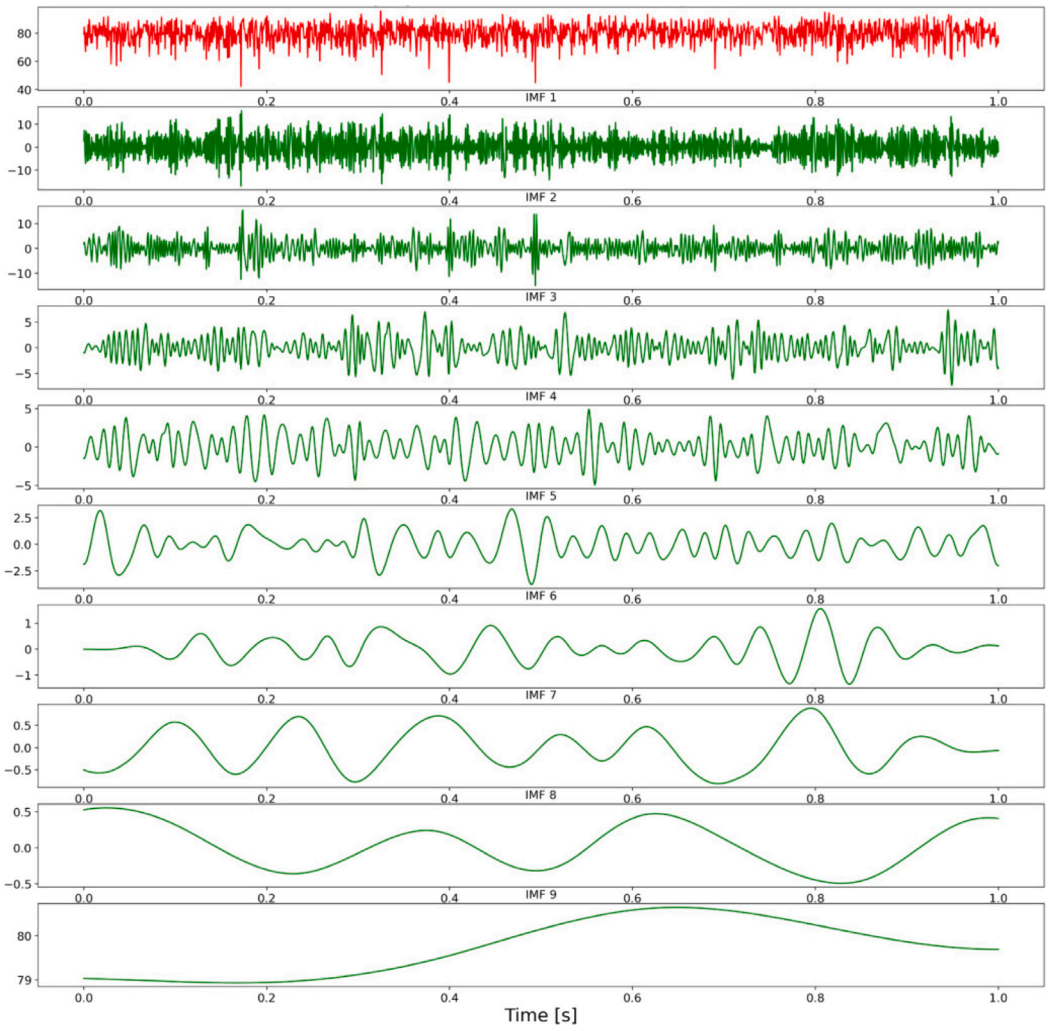


Fig. 5. The results of the empirical mode decomposition process for the mosquito density variable in the lowland area.

Table 3
The average model performance on various proportions of training data and testing data in each area. AMTr is the average Mean Square Error for data training, and AMTe is the average Mean Square Error for data testing.

Data proportion(%)	Lowland		Medium land		Highland	
Training:Testing	AMTr	AMTe	AMTr	AMTe	AMTr	AMTe
50 : 50	0.0069	0.0060	0.0529	0.2122	0.0030	0.0011
60 : 40	0.0078	0.0045	0.0022	0.0005	0.0020	0.0024
70 : 30	0.0001	0.0001	0.0006	0.0003	0.0006	0.0016
80 : 20	0.0000	**0.0001**	0.0049	**0.0000**	0.0005	**0.0001**
5-fold CV	0.0115	0.0045	0.0019	0.0008	0.0042	0.0011

method section. Finally, the additional data can be used to find out the performance metrics of forecasting results in the next periods.

Table 5 presents the performance metrics model for forecasting 1 and 3 months in advance using several models. For forecasting the next few months, the EMD-GRNN-PSO model has better performance than other models. This finding occurs in lowland, medium land, and highland. Table 5 also shows that the forecasting error averages in the next month are smaller than in the next three months. This condition indicates that the accuracy will decrease the more extended the forecasting period ahead. This finding is in line with the study conducted by Liu et al. (2021) and Albuquerque et al. (2022). The EMD-GRNN-PSO error forecast consisting of MSE and SMAPE appears smaller than other models. This smaller value shows that the EMD-GRNN-PSO has better performance than the comparison model. This statement is also supported by the results of the two-sided Diebold–Mariano test in

Table 4
The best model performance parameters for each Malang Regency area. MSE is the Mean Square Error, MAE is the Mean Absolute Error, and SMAPE is the Symmetric Mean Absolute Percentage Error.

Area	Best model		MSE		MAE		SMAPE	
	Sigma	Loss	Training	Testing	Training	Testing	Training	Testing
Low land	0.0002	14.5961	0.0264	0.0178	0.0516	0.0259	0.2823	0.1353
Medium land	−0.0003	2.7939	0.0049	0.0000	0.0166	0.0009	0.2334	0.0751
High land	−0.0002	0.3215	0.0005	0.0001	0.0054	0.0012	0.1565	0.0343

Table 5
Quantitative measure of performance for different forecasting model on daily period.

Model	Lowland		Middle land		Highland	
	1-month ahead forecast		1-month ahead forecast		1-month ahead forecast	
	MSE	SMAPE	MSE	SMAPE	MSE	SMAPE
Poisson Reg.	0.1631	0.9909	0.2282	0.9321	0.0900	0.9726
GAM	0.1081	0.9748	0.1681	0.8712	0.0770	0.9537
ARIMAX	0.1619	0.9834	0.3077	0.8912	0.1931	0.9997
SARIMAX	0.2269	0.9923	0.3072	0.8904	0.0923	0.9842
ARIMA-ANN	0.2360	0.9772	0.3459	0.8516	0.1930	0.9619
ARIMA-SVM	0.2346	0.9742	0.3431	0.8458	0.2191	0.9734
GRNN	0.1551	0.9860	0.2466	0.9526	0.0889	0.9775
Propagation	5.8616	1.5732	16.9525	0.8414	0.6945	0.3080
Kalman Filter	0.0140	0.2579	0.0150	0.3226	0.0033	0.6542
Proposed	**0.0001**	**0.6774**	**0.0001**	**0.7677**	**0.0001**	**0.8710**
Model	Lowland		Middle land		Highland	
	3-months ahead forecast		3-months ahead forecast		3-months ahead forecast	
	MSE	SMAPE	MSE	SMAPE	MSE	SMAPE
Poisson Reg.	0.1171	0.9872	0.1609	0.9553	0.0901	0.9789
GAM	0.0790	0.9574	0.1040	0.9017	0.0699	0.9423
ARIMAX	0.1151	0.9869	0.2078	0.9352	0.1934	0.9963
SARIMAX	0.1798	0.9967	0.2060	0.9345	0.0914	0.9871
ARIMA-ANN	0.1602	0.9695	0.2343	0.9033	0.2102	0.9723
ARIMA-SVM	0.1625	0.9652	0.2348	0.8981	0.2208	0.9731
GRNN	0.1128	0.9780	0.1689	0.9683	0.0892	0.9777
Propagation	15.0882	2.1725	16.9525	0.8414	3.5024	0.3080
Kalman Filter	0.0279	0.3695	0.0336	0.5190	0.0195	0.9487
Proposed	**0.0001**	**0.8045**	**0.0001**	**0.7283**	**0.0001**	**0.8261**

Table 6
Result of Diebold–Mariano (DM) test in each area.

Time Horizon Forecasting	Lowland			Middle land			Highland		
	Absolute DM Value			Absolute DM Value			Absolute DM Value		
	GAM	ARIMA-SVM	KALMAN FILTER	GAM	ARIMA-SVM	KALMAN FILTER	GAM	ARIMA-SVM	KALMAN FILTER
1 month ahead	1.6953	1.6000	1.3352	1.8918	1.2241	1.8918	0.4820	0.3305	0.5030
3 months ahead	1.7211	1.6817	1.3715	1.1868	1.2233	1.1499	0.2759	0.1951	0.2816

Table 6. The absolute of Diebold–Mariano value for some time horizons forecasting in each land is used to determine whether the EMD-GRNN-PSO forecast model results are different from GAM, ARIMA-SVM, and Kalman Filter model forecast results as the second, third, and fourth best models. Table 6 shows the DM test result between the EMD-GRNN-PSO model with the corresponding GAM, ARIMA-SVM, and Kalman Filter models. These absolute DM values in Table 6 are smaller than the critical value (1.96). This condition indicates that the performance of the proposed hybrid approach is significantly different from others.

In addition, the results of the paired-samples t-test are also shown in Table 7 to further strengthen the conclusions of the Diebold–Mariano test. This test is to show that being significantly different from others means better or not better. Table 7 shows that the *t*-value in one-month ahead and three-month ahead forecasting in all areas has a value that is bigger than the critical value (−1.6973 for one month and −1.6580 for three months ahead forecasting). This condition indicates that the null hypothesis is not rejected, which means that the EMD-GRNN-PSO hybrid approach is better than the other approaches.

Meanwhile, the forecasting results on each time horizon are matched with the three best comparison methods shown in Fig. 5. Fig. 5 displays a comparison of actual data and forecasting results using EMD-GRNN-PSO and the three best comparison methods after EMD-GRNN-PSO. As shown in Fig. 5, the blue line shows the actual data on the number of cases, while the other color line indicates the forecasted data using the EMD-GRNN-PSO, the ARIMA-SVM, the GAM model, and the Kalman Filter. In forecasting the next month and three months, forecasting results with EMD-GRNN-PSO models look more follow the actual data pattern than other models, despite the difference in value. Meanwhile the ARIMA-SVM model produces a forecast that looks similar but the period is too late. GAM models have forecasts with patterns that do not follow actual data patterns. Forecasting results with the Kalman Filter appear to be close to zero in the early period and it is difficult to follow the ups and downs pattern. However, in the last few periods it seems that it has been able to follow a pattern of ups and downs but there is a lag in periods (see Fig. 6).

4.3.5. Experiment 5: Robustness test model

The robustness test model also needs to be considered to determine how fit the model is applied to other data. This study uses two scenarios to observe the robustness of the model. The first scenario is to know how the EMD-GRNN-PSO performance in different periods of data. The data periods are daily, one week, two weeks, and one month. This scenario uses all areas. Furthermore, the second scenario is to test the model on the separate area in various land. This scenario is done on the

Table 7
Two means comparison test results.

Area	Pair	1-month ahead				3-months ahead			
		Std. Error Mean	95% Confidence Interval		t-value	Std. Error Mean	95% Confidence Interval		t-value
			Lower	Upper			Lower	Upper	
Lowland	M1 - M3	0,0812	0,0190	0,3504	2,276	0,0411	0,0009	0,1640	2,008
	M2 - M3	0,0819	−0,2435	0,0911	−0,930	0,0394	0,0421	0,1987	3,056
Middle land	M1 - M3	0,0688	−0,1386	0,1423	0,027	0,0587	0,0539	0,2870	2,905
	M2 - M3	0,0979	0,1200	0,5200	3,268	0,0595	−0,0520	0,1845	1,112
Highland	M1 - M3	0,0690	−0,1466	0,1351	−0,083	0,0421	−0,0684	0,0990	0,364
	M2 - M3	0,0856	−0,0470	0,3025	1,493	0,0473	0,0547	0,2425	3,143

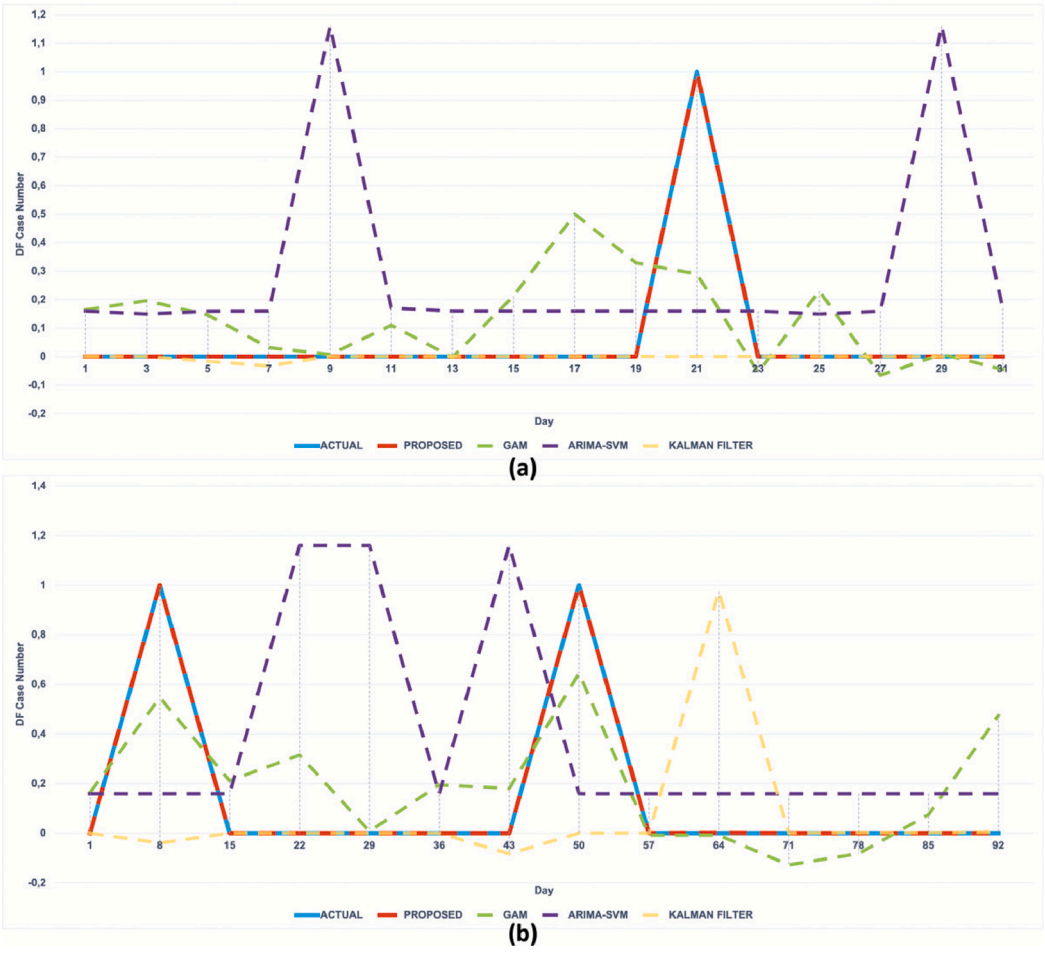


Fig. 6. The comparison of daily DF Cases Number and forecasting results with three best performance in Malang Regency, Indonesia, where (a) comparison in the next one month, (b) comparison in the next three months.

same data period as used in choosing the best model. These scenarios lead to differences in zero value proportions, data patterns, as well as data distribution.

Robustness testing in the first scenario with data for different periods was carried out in the sample areas on each land. It is necessary to adjust each region's lag, considering that the data period used differs from the best model in Table 4. The robustness test results of the best hybrid model for different data periods are presented in Table 8. Data with a longer period has a fewer proportion of zero values, as shown in column 3. Table 8 shows that models have different performances on different zero-value proportions. For the larger proportion of zero values and smaller standard deviations, it turns out that the model produces smaller MSE and SMAPE as well. It can be seen that the model has the best performance for data with a monthly period where the proportion of zero values is the smallest.

The second scenario was carried out on daily data for other areas with different zero values proportions. The performance results are presented in Table 9. Compared to the performance of the daily data forecasting results of the previous model in Table 8, the forecasting results in this other area in Table 9 have a tiny performance difference. Indeed, this forecasting performance in this scenario is better. It is evident from the smaller MSE, MAE, and SMAPE averages. In addition, each area in Table 9 has a different proportion of zero values for each variable. Column 2 shows the proportion of zero values in the DF Cases number variable. This variable has the most substantial proportion of zero values compared to other variables. Based on Table 9, Dau and Karangploso have a smaller proportion of zero values than others, but a greater standard deviation. This situation causes the MSE and SMAPE of this region to be larger than the Donomulyo and Jabung regions, which have a greater proportion of null values but smaller standard deviations. This finding is a bit different from the conditions shown in

Table 8
Performance model EMD-GRNN-PSO in different area and period.

Area	Period	Zero value proportion	Standard Deviation	MSE	MAE	SMAPE
Lowland	Daily	0.8940	0.5121	0.0388	0.0616	0.3065
	1 week	0.6517	0.5603	0.0007	0.0008	0.2247
	2 weeks	0.4608	3.7932	0.0005	0.0040	0.1026
	Monthly	0.0601	7.0443	0.0001	0.0014	0.0595
Medium land	Daily	0.8790	0.5972	0.0049	0.0168	0.2677
	1 week	0.5806	2.1621	0.0150	0.0161	0.1849
	2 weeks	0.4301	3.9521	0.0091	0.0321	0.2217
	Monthly	0.0621	4.9623	0.0009	0.0706	0.0819
Highland	Daily	0.9190	0.4224	0.0106	0.0160	0.2764
	1 week	0.6211	1.1952	0.0118	0.0271	0.2277
	2 weeks	0.4102	2.0571	0.0100	0.0153	0.1709
	Monthly	0.0001	2.2088	0.0021	0.0141	0.1700

Table 9
Performance model EMD-GRNN-PSO in daily data and different areas.

Area	Zero value proportion	Standard Deviation	MSE	MAE	SMAPE
Donomulyo	0.8604	0.4611	0.0011	0.0042	0.0029
Dau	0.6041	1.2690	0.0061	0.0178	0.1783
Karangploso	0.6761	0.6941	0.1112	0.0645	0.0631
Lawang	0.9125	0.2834	0.1279	0.0537	0.0188
Jabung	0.8638	0.3796	0.0001	0.0010	0.0004

Table 10
Performance EMD-GRNN-PSO model in different characteristic data. MSE is the Mean Square Error, MAE is the Mean Absolute Error, and SMAPE is the Symmetric Mean Absolute Percentage Error.

Dataset	Var. #	Min. value	Max. value	Zero value proportion	Standard Deviation	MSE	MAE	SMAPE
A	6	1343	1655	0.0000	173.8530	8.3086	0.4573	0.0003
B	6	−28	1045	0.0355	87.9220	0.0000	0.0000	0.0016
C	12	0	39650	0.0414	50.3750	0.7620	0.1086	0.0034

Table 8. However, if we examine the actual data more deeply, it turns out that the distribution of zero values in the Dau and Karangploso areas tends to accumulate in certain very long intervals. This very long series of zero values may cause the MSE and SMAPE values to be larger than other regional data where any zero value has a shorter series.

4.3.6. Experiment 6: Experiment on another secondary dataset

The next experiment was carried out on data outside the case study. This data has a proportion of zero values and a standard deviation that is different from the previous one. The three datasets used have a very small proportion of zero values (an average of less than 0.03) but a very large standard deviation on average (more than 100 on average). In addition, there are also negative values. This experiment aims to see how the performance of the proposed hybrid model forecasts data with different characteristics from the previous one. The model performance is shown in Table 10.

Table 10 shows that the hybrid model is still able to predict data with better performance than the previous one, even for data with negative values. This is indicated by the very small SMAPE value of 0.03% for dataset A, 0.16% for dataset B, and 0.34% for dataset C.

5. Discussion

DF is a case that mostly occurs in the rainy season (Lee et al., 2017; Malang, 2020). These cases will increase over the rainy season and then fall again when entering the dry season. It could even be that during the dry season, there are no cases (Malang, 2020). The observation in Fig. 3 shows that the average number of DF cases increased very rapidly in December-January and peaked every January with an average increase of about 15 times from the previous month. The peak DF cases' timing follows that presented by Husnina et al. (2019). The number of cases that increased in January could be caused by climatic conditions that support this disease's growth. Temperature conditions in December-January have an average of 25.92 degrees Celsius. This temperature is consistent with those mentioned by Husnina et al. (2019), who say that the best temperature range for mosquitoes to survive is between 18–27 °C. It is also consistent with Tsai et al. (2018) and Ebi and Nealon (2016) who state that the best temperature range is between 20–30 °C, and 21.6–32.9 °C according to Xiang et al. (2017). The average humidity in December–January of 85.42% exceeds the optimal temperature at which mosquitoes can survive. It is said that the optimal humidity is 60%–80% (Husnina et al., 2019). Rainfall average condition is 13.87 mm, 0.95 m/s for average wind speed, mosquito density is 90.11%, and average population density is 2,204.48 people/km^2. In this January, the temperature conditions decreased by 1.87% from the previous month's average. Meanwhile, the average humidity increased by 2.62%, rainfall increased by 60.82%, wind speed increased by 22.23%, and mosquito density increased by 2.48%.

Based on Table 2, the temperature has a positive correlation with the DF case number. This result follows what was stated by Husnina et al. (2019) and Lee et al. (2017). Although the effect is not the highest, the temperature is also very influential on mosquitoes' life cycle (Naish et al., 2014). This mean temperature can increase the larval growth rate, cut the appearance of adult mosquitoes, intensify the biting rate, and decrease the time it takes to reproduce (Naish et al., 2014). Population density and humidity have a greater influence than other variables on the DF case number in all lands. These variables provide suitable conditions for mosquitoes' development and survival, as suggested in Carvajal et al. (2018). The combination of humidity and temperature affects the amount of blood food and can affect mosquitoes' survival rate (Naish et al., 2014). Larvae free index that was represented as mosquito density and rainfall have a negative correlation on the DF case number, although the correlation value is small. Heavy rain can remove eggs, larvae, and pupae from the water, but residual water can create long-term breeding habitats (Naish et al., 2014). Wind speed has a small correlation close to zero with the DF case number. However, combining wind speed with other climatic variables can affect the duration and intensity of DF spread as defined in Naish et al. (2014).

Time lag or delay effect of climate variable and mosquito density on DF cases also need to be considered. This time lag variable affects the occurrence of DF. It is related to the dynamics of the life cycle of mosquitoes and viruses. Starting from the hatching of mosquitoes, the development of larvae and cocoons, adult mosquitoes, and virus amplification and incubation in humans (Carvajal et al., 2018; Ebi & Nealon, 2016; Naish et al., 2014). The time lag for each variable is distinct (Carvajal et al., 2018; Naish et al., 2014). Based on this finding, we selected the time lag of 7, 10, 14, 20, and 30 days in this study. The best time lag for each land is shown in Table 2. In all areas, it can be seen that the temperature lag is shorter than the humidity lag. This condition is in line with what was conveyed by Tsai et al. (2018) and Naish et al. (2014).

On the whole land level, the hybrid EMD-GRNN-PSO model can reduce the average of SMAPE. Modification of the GRNN inputs in the EMD-GRNN-PSO model has been proven to improve forecasting accuracy, as stated by Ghritlahre and Prasad (2018). Likewise, the addition of PSO, accelerates the discovery of optimal GRNN parameters so that optimal forecasting results are obtained. Furthermore, a comparison between the forecasting results for the next periods using EMD-GRNN-PSO and those using the other methods is presented in Fig. 5 and Table 5. Fig. 5 compares the forecasting results for the next 1 and 3-month periods using the best three approaches. Fig. 5 shows that the EMD-GRNN-PSO model appears to be able to present the best forecasting in terms of magnitude. However, the ARIMA-SVM model produces forecasts that at a glance seem to follow the actual data pattern, but the period is too late. If this is about to be applied to the real world, it will be pointless. Also, in terms of magnitude, the ARIMA-SVM forecasting result is different from the actual data. GAM appears to be able to follow the fluctuating pattern of actual data at the right time. Nonetheless, GAM produces negative forecasting over several periods. This is also the same as the Kalman Filter. It is also a bit slow in responding the fluctuation patterns. This finding is consistent with Abeyrathna et al. (2016) who states that the Kalman Filter cannot predict accurately if there are sudden peaks or drops in the actual data. However, the best model is EMD-GRNN-PSO. This condition is as shown by the performance comparison in Table 5. The EMD-GRNN-PSO model looks perfect in the forecast for the next 31 and 92 days. On forecasting the next 3 months, the benchmark models have a lower performance, especially in periods where there are very long periods of zero values. In addition, Table 5 shows that the EMD-GRNN-PSO model performs better than GRNN in all lands. It indicates that the use of EMD result as GRNN inputs has an impact on GRNN performance. This finding supports the discovery submitted by Rooki (2016) which states the need to modify GRNN inputs.

The feasibility of applying EMD-GRNN-PSO in forecasting the dengue cases number shows that this model is more competitive than other models in this study's empirical case. This comparison is made based on forecasting for the following several periods in 3 land levels. The comparison results for each land are presented in Fig. 5 and Table 5. The hybrid model can improve the performance of forecasting at each land level. This statement is also supported by the results of the two-sided Diebold–Mariano test in Table 6 and paired samples t-test in Table 7. In comparing the EMD-GRNN-PSO model with GAM and ARIMA-SVM models, the values of absolute DM statistics are all less than the critical value at the 5% significance level. In addition, the critical value of the 5% level is 1.645. Therefore the null hypothesis is rejected as stated in Wang et al. (2020). It means that the forecast result of the EMD-GRNN-PSO model is significantly different from the GAM and ARIMA-SVM model. Likewise, the paired sample t-test, shows that the EMD-GRNN-PSO model produces better performance than the comparison model.

In the same case data, testing on data with various time intervals may result in different forecasting performances, as presented in Table 8. Forecasting using a longer time interval, weekly or monthly, results in a larger standard deviation than daily. It happens because the weekly or monthly data is an aggregate of daily data with shorter periods. However, this is different from the SMAPE, where the value is smaller when the data period is longer. This situation happens because the lag used gets smaller when the data period is longer. Table 8 also shows that for smaller standard deviation values, the error goes to more minor. Instead of increasing the proportion of zero value was followed by expanding the SMAPE. This increase indicates that the forecasting performance has decreased. These conditions strengthen (Mussumeci & Codeço, 2020) which states that getting natural results with lots of zero values with good performance is difficult. In addition, data with a more extended period has a smaller proportion of zero values, which allows the model to get better performance. The statement submitted by Mussumeci and Codeço (2020) supports this condition.

Table 8 show that the decrease in the average SMAPE is 7.11% for 1-week data, 11.85% for 2-week data, and 17.98% for monthly data. The decline in SMAPE for the lowland model is 8.18% for 1-week data, 20.39% for 2-week data, and 24.70% for monthly data. The medium terrain model reduces the SMAPE by 8.28% for 1-week data, 4.60% for 2-week period data, and 18.58% for monthly period data. Finally, SMAPE decreased by 4.88% for 1-week data for the highland model, 10.55% for 2-week period data, and 10.64% for monthly data. The decline in SMAPE occurred because when the data had weekly or monthly periods, the zero value proportion was will decrease. It causes the model to be better able to adapt to actual data. This finding is consistent with Mussumeci and Codeço (2020) stating that the model will have decreased performance if the data used increasingly contain zero values. In addition, Mussumeci and Codeço (2020) also state that forecasting with many zero values is difficult to get a good forecasting result also. Intermittent data forecasting is not easy to do. This statement is also shown in Table 8. The larger SMAPE value followed the increasing proportion of zeros in Table 8. In addition, performance results of the daily data model with different proportions of zero values from other areas in each land show that the EMD-GRNN-PSO model has not significantly different performance from daily best model performance in Table 9. Therefore, based on the results in scenario 1 and scenario 2, it is shown that the EMD-GRNN-PSO model is still robust and can produce better SMAPE as stated in Rooki (2016). In addition, the EMD-GRNN-PSO model is proven to be feasible for forecasting data by involving many zero values.

Model testing on data with very different characteristics is also presented in Table 10. It shows that each model can still produce excellent performance, even better than the previous DF data. All models have a SMAPE average of 0.178%, an average MSE of 3.024, and an average MAE of 0.189. The experiments of each land against other secondary data show that the hybrid model is still robust. Even for data containing negative values and larger interval data. In addition, SMAPE values also seem much smaller than SMAPE values in Tables 8 and 9. This situation is because the zero value in Table 10 is much smaller than the data in the previous experiment. Moreover, dataset B has an MSE and MAE that are smaller than other data even if there is much negative value in one of the variables, and the range of data in one variable is vast. Actually, this value can be considered as noise which can cause difficulties in getting forecasting values that are close to the actual data. In addition, negative values can add noise, just like zero values.

6. Conclusion

An excellent forecasting performance using data with unique characteristics is not easy to achieve. Such data characteristics include a combination of multiple zero values, small values, and heteroscedasticity. The hybrid model of empirical mode decomposition, generalized regression neural network, and particle swarm optimization successfully predicted the data with reliable performance. The use of empirical mode decomposition results as input successfully decreased the forecasting errors. Moreover, this input also enabled the forecasting data

pattern to follow the actual data pattern. Other methods may produce errors that are not too different from our experimental result, but in reality, their forecast data patterns tend to differ from actual data patterns. In addition to the type of input, the difference in the proportion of zeros significantly affects the performance of the model. As the proportion of zero values increases, this hybrid model produces lower SMAPE.

There is a limitation to this study. The limitation is the problem of data reporting restrictions. Data on the case number used in this study are under the case number registered at the Malang Regency Public Health Office. In reality, there may be cases that are not reported. For this reason, the subsequent development can be considered to involve social media data as a complement to conventional data. Apart from these limitations, the EMD-GRNN-PSO model can forecast data with different periods and characteristics. EMD-GRNN-PSO was successful in predicting the dengue cases number in the following few periods. Policymakers can use these results to regulate resource use and determine mitigation measures to reduce the impact of future dengue fever outbreaks.

CRediT authorship contribution statement

Wiwik Anggraeni: Conceptualization, Methodology, Formal analysis, Writing – original draft preparation and editing. **Eko Mulyanto Yuniarno:** Conceptualization, Validation, Formal analysis. **Reza Fuad Rachmadi:** Data curation, Methodology, Validation. **Surya Sumpeno:** Supervision, Writing – review. **Pujiadi Pujiadi:** Conceptualization, Data, Validation. **Sugiyanto Sugiyanto:** Validation, Writing – review & editing. **Joan Santoso:** Formal analysis, Software, Writing – review. **Mauridhi Hery Purnomo:** Supervision, Formal analysis, Writing – review.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work is supported by the Ministry of Education, Culture, Research, and Technology Indonesia for funding under the Excellent and Applied Research of Universities grant scheme under contract number 008/E5/PG.02.00.PT/2022, implementation support from the Public Health Office of Malang Regency and the University Center of Excellence on Artificial Intelligence for Healthcare and Society, Institut Teknologi Sepuluh Nopember.

References

Abeyrathna, M., Abeygunawrdane, D., Wijesundara, R., Mudalige, V., Bandara, M., Perera, S., Maldeniya, D., Madhawa, K., & Locknathan, S. (2016). Dengue propagation prediction using human mobility. In *2016 Moratuwa engineering research conference* (pp. 156–161). http://dx.doi.org/10.1109/MERCon.2016.7480132.

Albuquerque, P. C., Cajueiro, D. O., & Rossi, M. D. (2022). Machine learning models for forecasting power electricity consumption using a high dimensional dataset. *Expert Systems with Applications*, *187*, Article 115917. http://dx.doi.org/10.1016/j.eswa.2021.115917.

Anggraeni, W., Abdillah, A., Pujiadi, Trikoratno, L., Wibowo, R., Purnomo, M., & Sudiarti, Y. (2019). Modelling and forecasting the dengue hemorrhagic fever cases number using hybrid fuzzy-ARIMA. In *2019 IEEE 7th international conference on serious games and applications for health* (pp. 1–8). IEEE, http://dx.doi.org/10.1109/SeGAH.2019.8882433.

Anggraeni, W., & Aristiani, L. (2016). Using google trend data in forecasting number of dengue fever cases with ARIMAX method case study: Surabaya, Indonesia. In *2016 International conference on information communication technology and systems* (pp. 114–118). http://dx.doi.org/10.1109/ICTS.2016.7910283.

Atchadé, M., & Sokadjo, Y. (2021). Overview and cross-validation of COVID-19 forecasting univariate models. *Alexandria Engineering Journal*, http://dx.doi.org/10.1016/j.aej.2021.08.028.

Boudraa, J. (2007). EMD-based signal filtering. *IEEE Transactions on Instrumentation and Measurement*, *56*, 2196–2202. http://dx.doi.org/10.1109/TIM.2007.907967.

Carvajal, T. M., Viacrusis, K., Hernandez, L., Ho, H., Amalin, D., & Watanabe, K. (2018). Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. *BMC Infectious Diseases*, *18*, 183. http://dx.doi.org/10.1186/s12879-018-3066-0.

Chakraborty, T., Chattopadhyay, S., & Ghosh, I. (2019). Forecasting dengue epidemics using a hybrid methodology. *Physica A. Statistical Mechanics and its Applications*, *527*, Article 121266. http://dx.doi.org/10.1016/j.physa.2019.121266.

Chen, Y., Chu, C. W., Chen, M. I., & Cook, A. R. (2018). The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison. *Journal of Biomedical Informatics*, *81*, 16–30. http://dx.doi.org/10.1016/j.jbi.2018.02.014.

Chen, S., Xu, J., Wu, Y., Wang, X., Fang, S., Cheng, J., Ma, H., Zhang, R., Liu, Y., Zhang, L., Zhang, X., Chen, L., & Liu, X. (2019). Predicting temporal propagation of seasonal influenza using improved gaussian process model. *Journal of Biomedical Informatics*, *93*, Article 103144. http://dx.doi.org/10.1016/j.jbi.2019.103144.

Chowell, G., Cazelles, B., & Broutin, C. (2011). The influence of geographic and climate factors on the timing of dengue epidemics in Perú, 1994–2008. *BMC Infectious Diseases*, *11*, 1–15. http://dx.doi.org/10.1186/1471-2334-11-164.

Cortes, F., Turchi Martelli, C., Arraes de Alencar Ximenes, R., Montarroyos, U., Siqueira Junior, J., Gonçalves Cruz, O., Alexander, N., & Vieira de Souza, W. (2018). Time series analysis of dengue surveillance data in two Brazilian cities. *Acta Tropica*, *182*, 190–197. http://dx.doi.org/10.1016/j.actatropica.2018.03.006.

Dhiman, R., Pahwa, S., Dhillon, G. P. S., & Dash, A. (2010). Climate change and threat of vector-borne diseases in India: are we prepared? *Parasitology Research*, *106*, 763–773. http://dx.doi.org/10.1007/s00436-010-1767-4.

Dos Santos, B. S., Steiner, M. T. A., Fenerich, A. T., & Lima, R. H. P. (2019). Data mining and machine learning techniques applied to public health problems: A bibliometric analysis from 2009 to 2018. *Computers & Industrial Engineering*, *138*, Article 106120. http://dx.doi.org/10.1016/j.cie.2019.106120.

Du, P., Wang, J., Yang, W., & Niu, T. (2019). A novel hybrid model for short-term wind power forecasting. *Applied Soft Computing*, *80*, 93–106. http://dx.doi.org/10.1016/j.asoc.2019.03.035.

Ebi, K., & Nealon, J. (2016). Dengue in a changing climate. *Environmental Research*, *151*, 115–123. http://dx.doi.org/10.1016/j.envres.2016.07.026.

Gabriel, A., Alencar, A., & Miraglia, S. (2019). Dengue outbreaks: unpredictable incidence time series. *Epidemiology and Infection*, *147*, 116. http://dx.doi.org/10.1017/S0950268819000311.

Ghritlahre, H., & Prasad, R. (2018). Investigation of thermal performance of unidirectional flow porous bed solar air heater using MLP, GRNN, and RBF models of ANN technique. *Thermal Science and Engineering Progress*, *6*, 226–235. http://dx.doi.org/10.1016/j.tsep.2018.04.006.

Guo, P., Zhang, Q., Chen, Y., Xiao, J., He, J., Zhang, Y., Wang, L., Liu, T., & Ma, W. (2019). An ensemble forecast model of dengue in guangzhou, China using climate and social media surveillance data. *Science of the Total Environment*, *647*, 752–762. http://dx.doi.org/10.1016/j.scitotenv.2018.08.044.

Hoang Cao, T., Duy Nguyen, A., Quang Dinh, T., Chan Luong, Q., & Thanh Diep, H. (2018). Forecasting dengue incidences: Statistical and dynamic models. *The Oxford Journal of Intelligent Decision and Data Science*, *2018*, 1–13. http://dx.doi.org/10.5899/2018/ojids-00017.

Hu, B., Qiu, W., Xu, C., & Wang, J. (2020). Integration of a Kalman filter in the geographically weighted regression for modeling the transmission of hand, foot and mouth disease. *BMC Public Health*, *20*(479), http://dx.doi.org/10.1186/s12889-020-08607-7.

Husnina, Z., Clements, A., & Wangdi, K. (2019). Forest cover and climate as potential drivers for dengue fever in sumatra and kalimantan 2006–2016: A spatiotemporal analysis. *Tropical Medicine & International Health*, tmi.13248. http://dx.doi.org/10.1111/tmi.13248.

Jain, R., Sontisirikit, S., Iamsirithaworn, S., & Prendinger, H. (2019). Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. *BMC Infectious Diseases*, *19*, 272. http://dx.doi.org/10.1186/s12879-019-3874-x.

Jayaraj, V., Avoi, R., Gopalakrishnan, N., Raja, D., & Umasa, Y. (2019). Developing a dengue prediction model based on climate in Tawau, Malaysia. *Acta Tropica*, *197*, Article 105055. http://dx.doi.org/10.1016/j.actatropica.2019.105055.

Jutinico, A., Vergara, E., Garcia, C., Palencia, M., & Orjuela-Cañon, A. (2021). Robust Kalman filter for tuberculosis incidence time series forecasting. In *IFAC-papersonline, Vol. 54* (pp. 424–429). http://dx.doi.org/10.1016/j.ifacol.2021.10.293.

Katris, C. (2021). A time series-based statistical approach for outbreak spread forecasting: Application of COVID-19 in Greece. *Expert Systems with Applications*, *166*, Article 114077. http://dx.doi.org/10.1016/j.eswa.2020.114077.

Kuranga, C., Muwani, T. S., & Ranganai, N. (2023). A multi-population particle swarm optimization-based time series predictive technique. *Expert Systems with Applications*, *233*, Article 120935. http://dx.doi.org/10.1016/j.eswa.2023.120935.

Kuranga, C., & Pillay, N. (2022). A comparative study of nonlinear regression and auto-regressive techniques in hybrid with particle swarm optimization for time-series forecasting. *Expert Systems with Applications*, *190*, Article 116163. http://dx.doi.org/10.1016/j.eswa.2021.116163.

Lee, H., Nguyen-Viet, H., Nam, V., Lee, M., Won, S., Duc, P., & Grace, D. (2017). Seasonal patterns of dengue fever and associated climate factors in 4 provinces in Vietnam from 1994 to 2013. *BMC Infectious Diseases*, *17*, 218. http://dx.doi.org/10.1186/s12879-017-2326-8.

Li, G., Liu, Z., Li, J., Fang, Y., Liu, T., Mei, Y., & Wang, Z. (2018). Application of general regression neural network to model a novel integrated fluidized bed gasifier. *International Journal of Hydrocarbon Engineering*, *43*, 5512–5521. http://dx.doi.org/10.1016/j.ijhydene.2018.01.130.

Liu, Y., Feng, G., Tsui, K., & Sun, S. (2021). Forecasting influenza epidemics in Hong Kong using google search queries data: A new integrated approach. *Expert Systems with Applications*, *185*, Article 115604. http://dx.doi.org/10.1016/j.eswa.2021.115604.

Liu, J., Wang, P., Chen, H., & Zhu, J. (2022). A combination forecasting model based on hybrid interval multi-scale decomposition: Application to interval-valued carbon price forecasting. *Expert Systems with Applications*, *191*, Article 116267. http://dx.doi.org/10.1016/j.eswa.2021.116267.

Lu, S., Zhang, Q., Chen, G., & Seng, W. (2021). A combined method for short-term traffic flow prediction based on recurrent neural network. *Alexandria Engineering Journal*, *60*, 87–94. http://dx.doi.org/10.1016/j.aej.2020.06.008.

Lv, S., & Wang, L. (2022). Deep learning combined wind speed forecasting with hybrid time series decomposition and multi-objective parameter optimization. *Applied Energy*, *311*, Article 118674. http://dx.doi.org/10.1016/j.apenergy.2022.118674.

Mala, S., & Jat, M. (2019). Implications of meteorological and physiographical parameters on dengue fever occurrences in Delhi. *Science of the Total Environment*, *650*, 2267–2283. http://dx.doi.org/10.1016/j.scitotenv.2018.09.357.

Malang (2020). Geographical condition -malang one data (kondisi geografis -kabupaten malang satu data). URL http://malangkab.go.id/uploads/dokumen/malangkab-KondisiGeografis.pdf, online (accessed July 30, 2020).

Malang, P. K. (2021). Center for data and information - health ministry of the Republic of Indonesia (Pusat Data dan Informasi - Kementerian Kesehatan Republik Indonesia. URL https://pusdatin.kemkes.go.id/folder/view/01/structure-publikasi-data-pusat-data-dan-informasi.html, online (accessed March 5, 2021).

Masikos, M., Demestichas, K., Adamopoulou, E., & Theologou, M. (2015). Mesoscopic forecasting of vehicular consumption using neural networks. *Soft Computing*, *19*, 145–156. http://dx.doi.org/10.1007/s00500-014-1238-4.

Murray, N., Quam, M., & Wilder-Smith, A. (2013). Epidemiology of dengue: Past, present and future prospects. *Clinical Epidemiology*, 299. http://dx.doi.org/10.2147/CLEP.S34440.

Mussumeci, E., & Codeço, C. (2020). Large-scale multivariate forecasting models for dengue - LSTM versus random forest regression. *Spatial Spatio-temporal Epidemiology*, http://dx.doi.org/10.1016/j.sste.2020.100372.

Naish, S., Dale, P., Mackenzie, J., McBride, J., Mengersen, K., & Tong, S. (2014). Climate change and dengue: A critical and systematic review of quantitative modelling approaches. *BMC Infectious Diseases*, *14*, 167. http://dx.doi.org/10.1186/1471-2334-14-167.

Nur Adli Zakaria, M., Abdul Malek, M., Zolkepli, M., & Najah Ahmed, A. (2021). Application of artificial intelligence algorithms for hourly river level forecast: A case study of Muda river, Malaysia. *Alexandria Engineering Journal*, *60*, 4015–4028. http://dx.doi.org/10.1016/j.aej.2021.02.046.

Ozer, I., Efe, S., & Ozbay, H. (2021). A combined deep learning application for short term load forecasting. *Alexandria Engineering Journal*, *60*, 3807–3818. http://dx.doi.org/10.1016/j.aej.2021.02.050.

Rooki, R. (2016). Application of general regression neural network (GRNN) for indirect measuring pressure loss of Herschel–Bulkley drilling fluids in oil drilling. *Measurement*, *85*, 184–191. http://dx.doi.org/10.1016/j.measurement.2016.02.037.

Shashvat, K., Basu, R., & Bhondekar, A. (2019). Application of time series methods for dengue cases in north India (Chandigarh). *Journal of Public Health*, *29*, 433–441. http://dx.doi.org/10.1007/s10389-019-01136-7.

Siriyasatien, P., Chadsuthi, S., Jampachaisri, K., & Kesorn, K. (2018). Dengue epidemics prediction: A survey of the state-of-the-art based on data science processes. *IEEE Access*, 53757–53795. http://dx.doi.org/10.1109/ACCESS.2018.2871241.

Siriyasatien, P., Phumee, A., Ongruk, P., Jampachaisri, K., & Kesorn, K. (2016). Analysis of significant factors for dengue fever incidence prediction. *BMC Bioinformatics*, *17*, http://dx.doi.org/10.1186/s12859-016-1034-5.

Tsai, C., Yeh, T.-G., & Hsiao, Y.-R. (2018). Evaluation of hydrologic and meteorological impacts on dengue fever incidences in southern Taiwan using time-frequency analysis methods. *Ecological Informatics*, *46*, 166–178. http://dx.doi.org/10.1016/j.ecoinf.2018.05.002.

Wang, S., Li, Y., & Yang, H. (2019). Self-adaptive mutation differential evolution algorithm based on particle swarm optimization. *Applied Soft Computing*, *81*, Article 105496. http://dx.doi.org/10.1016/j.asoc.2019.105496.

Wang, J., Niu, X., Liu, Z., & Zhang, L. (2020). Analysis of the influence of international benchmark oil price on China's real exchange rate forecasting. *Engineering Applications of Artificial Intelligence*, *94*, Article 103783. http://dx.doi.org/10.1016/j.engappai.2020.103783.

Wen, K., Zhao, G., He, B., Ma, J., & Zhang, H. (2022). A decomposition-based forecasting method with transfer learning for railway short-term passenger flow in holidays. *Expert Systems with Applications*, *189*, Article 116102. http://dx.doi.org/10.1016/j.eswa.2021.116102.

WHO (2016). Weekly epidemiology report. URL http://www.who.int/wer/2016/wer9130/en/, online (accessed June 23, 2020).

WHO (2017). Dengue guidelines for diagnosis, treatment, prevention and control: new edition. URL https://www.who.int/rpc/guidelines/9789241547871/en, online (accessed June 23, 2020).

WHO (2020). Dengue and severe dengue. URL https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue, online (accessed June 23, 2020).

Xiang, J., Hansen, A., Liu, Q., Liu, X., Tong, M. X., Sun, Y., Cameron, S., Hanson-Easey, S., Han, G., Williams, C., Weinstein, P., & Bi, P. (2017). Association between dengue fever incidence and meteorological factors in Guangzhou, China, 2005–2014. *Environmental Research*, *153*, 17–26. http://dx.doi.org/10.1016/j.envres.2016.11.009.

Yang, X., & Li, H. (2023). Evolutionary-state-driven multi-swarm cooperation particle warm optimization for complex optimization problem. *Information Sciences*, *646*, Article 119302. http://dx.doi.org/10.1016/j.ins.2023.119302.

Zhang, K., Cao, H., Thé, J., & Yu, H. (2022). A hybrid model for multi-step coal price forecasting using decomposition technique and deep learning algorithms. *Applied Energy*, *306*, Article 118011. http://dx.doi.org/10.1016/j.apenergy.2021.118011.

Zhu, G., Liu, T., Xiao, J., Zhang, B., Song, T., Zhang, Y., Lin, L., Peng, Z., Deng, A., Ma, W., & Hao, Y. (2019). Effects of human mobility, temperature and mosquito control on the spatiotemporal transmission of dengue. *Science of the Total Environment*, *651*, 969–978. http://dx.doi.org/10.1016/j.scitotenv.2018.09.182.