Applied Intelligence
https://doi.org/10.1007/s10489-023-05101-6

Oriented transformer for infectious disease case prediction

Zhijin Wang[1] · Pesiong Zhang[2] · Yaohui Huang[3] · Guoqing Chao[4] · Xijiong Xie[5] · Yonggang Fu[1]

Accepted: 9 October 2023

Abstract
Accurate prediction of infectious disease cases plays a crucial role in achieving effective infection prevention and control. However, the inherent variability of incubation periods and progression dynamics of infectious diseases pose significant challenges to the accuracy of predicting multiple diseases. Multiple representation fusion (MRF) methods would improve the performance of prediction models, due to their capability to capture diverse temporal dependencies that reflect potential disease transmission patterns. But the traditional fusion approach for infectious disease prediction still faces many challenges, including the requirement of auxiliary data, vulnerability to disease evolution, and lack of intuitive explanation. To address these challenges, this paper proposes an oriented transformer (ORIT) for infectious diseases case prediction. Contrary to traditional MRF structures that integrate representations from multiple data sources, the MRF in the proposed ORIT combines multi-orientation context vectors solely by capturing multi-dimensional temporal relationships within disease case data. Furthermore, this paper considers the heterogeneity of the incubation period in the prediction of different infectious disease cases. Lastly, this paper conducts comprehensive experiments to evaluate the proposed method using two real datasets of infectious diseases, and compares it with 21 well-known prediction methods. The experimental results verify the effectiveness of the proposed method.

Keywords Infectious disease · Representation fusion · Oriented transformer · Attention mechanism · Time series

1 Introduction

Infectious diseases (IDs) persist as a substantial threat to human health [45]. Over one million people are newly infected by hand, foot, and mouth disease (HFMD) and hepatitis beta virus (HBV) per year in China [19]. Early warning systems play a pivotal role in managing infectious disease risk [32]. The predictive technique is the foundation of this system [42], which is instrumental in guiding healthcare decision-making processes and devising intervention strategies [8]. The development of efficient prediction methods for IDs has attracted considerable attention from both researchers and healthcare industries in recent years.

The majority of IDs cases prediction methods are divided into three categories: epidemic dynamics methods, statistical methods, and machine learning methods [4]. Epidemic dynamics methods are considerably more intuitive and comprehensible, largely due to their reliance on known biological and social factors that dictate disease propagation [35]. Predominantly, these techniques utilize mathematical or numerical models like the SIR (Susceptible, Infected, Recovered) model [36], SEIR (Susceptible, Exposed, Infected, Recovered) model [16], among others [10], to simulate the process of disease transmission. However, these models exhibit sensitivity to preset parameters [21] and are contingent upon oversimplified assumptions concerning disease transmission, such as population homogeneity and mixing, which may not consistently correspond with actual conditions. Statistical methods enable the exploration of trends and relationships within historical IDs cases data through a robust theoretical framework, thereby yielding trustworthy outcomes [15]. Nevertheless, their efficacy is constrained by data quality and struggle to capture complex relationships, time-step interactions, and nonlinear trends.

To complement these traditional methods of IDs cases prediction, machine learning approaches are receiving increased attention. Random Forest (RF), support vector regression (SVR), etc., have been actively studied for IDs cases prediction [42]. More recently, deep learning techniques, encompassing Recurrent Neural Networks (RNNs), Convolutional

✉ Zhijin Wang
zhijinecnu@gmail.com

Extended author information available on the last page of the article

Published online: 09 November 2023

Neural Networks (CNNs), Attention Mechanism (Attn), and Transformer models, have been leveraged for this task, demonstrating their capacity to extract complex, non-linear relationships within the data [49]. Despite these advancements, deploying IDs cases prediction models in real-world scenarios encounters several inherent challenges.

One challenge involves investigating the complex incubation period of IDs, a critical aspect for comprehending the transmission intensity and patterns of IDs [48], thereby enabling accurate predictions. The present approaches consider the temporal dependencies among historical observations, but have yet to effectively estimate the characteristic of the latent disease incubation period. The variability of disease pathogenesis, regional differences, and other related environmental factors further create a barrier to their practical application in the field [27]. Another notable hurdle involves handling intricate temporal patterns of historical records. The progression and dissemination dynamics of IDs are highly uncertain, which may stem from diverse sources such as environmental changes, human behavior shifts, and viral evolution [40]. While the success of the attention mechanism [39] prompting recent efforts to establish interrelationships between different time steps and periods [37], the complex and uncertain dynamics of IDs data still present learning challenges for models. This emphasizes the necessity for extracting more efficient and detailed representations of the progression of IDs cases.

To address the aforementioned challenges, it has become necessary to construct a multiple representation fusion (MRF) structure to maximize the value of the data and improve collaboration among different temporal characteristics at various scales. This strategy further augments the intrinsic characteristics of data and has achieved promising performance in diverse domains, including virtual/augmented reality [14], high-utility pattern mining [30], and the information extraction from large-scale databases [25]. Nevertheless, applying an MRF structure to IDs cases prediction has the following limitations. First, the current MRF structure relies on acquiring data from multiple sources or sensors. However, in this study, IDs cases data may not always be accessible or consistently reliable. For instance, reporting inadequacies due to insufficient medical facilities or issues like misreporting or delayed reporting could result in incorrect or incomplete data, reducing prediction accuracy. Secondly, the traditional MRF structure could potentially introduce more complex and heterogeneous feature associations from diverse data sources. This might necessitate additional preprocessing for alignment, resulting in a decline in model efficiency and stability. Therefore, models should obtain the most informative representation of the prediction target by utilizing as little additional data as possible.

To address the above issues, we propose a novel MRF model for IDs cases prediction, called oriented transformer (ORIT). This approach can be employed across different types of infectious diseases, relying solely on historical case observations. Compared to the conventional MRF model, the representation in ORIT is obtained from the temporal relationships across different time scales in IDs case data, generated by a designed multi-head oriented attention unit (MOAU). To consider the varying incubation periods across different diseases, we analyze the effect of distinct window sizes on the prediction model. Leveraging the optimally learned window sizes, MOAU extracts the temporal characteristics. Subsequently, a temporal fusion component is introduced to adaptively augment the beneficial patterns for the model, leading to the generation of predictions.

The main contributions of this paper are summarized as follows:

(1) The inherent diversity in the incubation periods and outbreak patterns for various infectious diseases has been investigated and incorporated into the construction of the prediction model.
(2) The impact of multi-oriented correlations among different scales of IDs progression dynamics has been validated using multiple MOAUs.
(3) Two real-world collections comprising HFMD and HBV cases are utilized to conduct a comprehensive evaluation of the proposed ORIT in comparison with 21 pertinent prediction models. The results demonstrate the superiority of our model in IDs cases prediction. A study on attention combination has validated the effectiveness of fusing the various latent relationship.

The remains of this paper are organized as follows. Section 2 addresses this research. Section 3 defines the prediction problem. Section 4 graphically illustrates the proposed ORIT. Section 5 gives descriptions of the collection and experiment setup. Section 6 presents experiments and analyses. Finally, a conclusion is drawn in Section 8.

2 Related work

This section discusses the related research from the perspectives of attention-based models and temporal representation learning methods.

2.1 Attention-based predictive methods

The attention mechanism has achieved success across various fields [11]. According to the difference in implementation, these methods can be categorized as additive attention and dot-product attention.

Additive attention is the earliest form of attention implementation, which typically works rely on the encoder-

decoder structure [12]. DA-RNN [29] and LSTNet [23] are representative methods that capture attention representations in the recurrent layer and highlight observations by augmenting attention representations to previous time steps. These methods effectively capture temporal dynamics from time series with conspicuous periodicity. However, the limitation of RNNs to handle various temporal dynamics restricts the prediction performance of these methods.

Dot product is currently a popular way to compute attention, which allows models to escape the constraints of the recurrent architecture [9]. DSANet [20] captures attention representations in dual-scaled convolutional layers from multiple time series. Informer [50] directly employed an attention mechanism to extract temporal maps from uni-variate time series and generate predictions by further emphasizing the temporal maps. In essence, these attention techniques offer the capability to highlight the observations of time intervals from the input time series. However, these forms of attention primarily consider local temporal dynamics as they underscore the weight of key time steps. Consequently, the temporal dynamics of inconspicuous periods, such as the incubation period of infectious diseases, would be overlooked [40].

This research assumes that attention is not solely limited to the dimensionality of time steps. The varied dimensionality of observations should be incorporated into attention as well.

2.2 Temporal representation learning methods

Temporal representation learning is broadly employed in time series prediction [24]. Depending on their strategies, these methods can be categorized as stacking and fusion methods.

Stacking methods repeat the network blocks on inputs or subsequent layers to extract temporal representations [46]. Temporal convolutional networks (TCN) [6] used stacked CNN layers to discover local temporal dynamics from time series. However, these methods frequently disregard long-term temporal dependencies, considerably reducing prediction accuracy. The stacking RNNs are proposed to tackle this challenge owing to their superiority in capturing decency relationships from the sequence [17]. For instance, Long Short-Term Memory Multi-Seasonal Net (LSTM-MSNet) [7] employed deep LSTM layers to capture long-term dependencies. Nevertheless, these methods struggle with capturing short-term dynamics effectively, weakening their ability to predict outbreak points. Even when combining the above techniques, such as CNNRNNRes [44], the problem persists.

Fusion methods typically capture several representations from various components or neural layers instead of stacking components on a time series. DGR [42] utilized recurrent layers to obtain temporal representations from dual-grained epidemic time series and achieved better performance when compared with solely stacking recurrent layers. Nevertheless, it merely considers the temporal dynamics from different time steps, whereas it ignores the effects of different periods. Mvt [43], and DGDR [49] leveraged different components to obtain several temporal representations from time series and fuses them to generate predictions. These methods regard the time series as a cube and consider the time series in different orientations. However, they cannot efficiently highlight the key information within temporal representations, which affects prediction accuracy.

This paper proposes a novel temporal representation learning method, known as the oriented transformer (ORIT). It learns the temporal representations by employing attention mechanisms from various orientations, enabling the capture of more key temporal dynamics from epidemic time series.

3 Problem definition

The observations of infection cases are sequential values within an identical time span. Hence, predicting infectious disease cases can be commonly regarded as a time series prediction problem. Let symbol $\boldsymbol{Z} \in \mathbb{R}^{N\times 1}$ denote the time series, where N is the number of total time steps. A look-back window is frequently used to reorganize the inputs to enhance the temporal characteristics of time series. Let T be the length of the look-back window, also called window size. An observation within a look-back window can be denoted as $\boldsymbol{Z}_{t+1:t+T,1} \in \mathbb{R}^{T\times 1}$, where t is a time step. And the infected cases prediction problem can be formulated as:

$$\hat{Y}_{t+T+1,1} = \mathcal{F}(\boldsymbol{Z}_{t+1:t+T,1}), \tag{1}$$

where $\hat{Y}_{t+T+1,1} \in \mathbb{R}$ is the prediction, and $\mathcal{F}(\cdot)$ is a mapping function denotes the prediction model.

To incorporate the attention mechanism to input time series, the formulation for infected cases prediction can be described as:

$$\hat{Y}_{t+T+1,1} = \mathcal{F}(\mathcal{A}^1(\boldsymbol{Z}_{t+1:t+T,1}), \mathcal{A}^2(\boldsymbol{Z}_{t+1:t+T,1}), \cdots), \tag{2}$$

where $\mathcal{A}^i(\cdot)$ is the attention mechanism of different orientations, each orientation denotes a type of temporal characteristic. It is worth noting that the inputs of the attention component could also be the outputs of proceeding neural networks. The main symbols are listed in Table 1.

4 The proposed ORIT

The schematic illustration of the proposed ORIT is plotted in Fig. 1. According to the workflow in subfigure 1(a), this

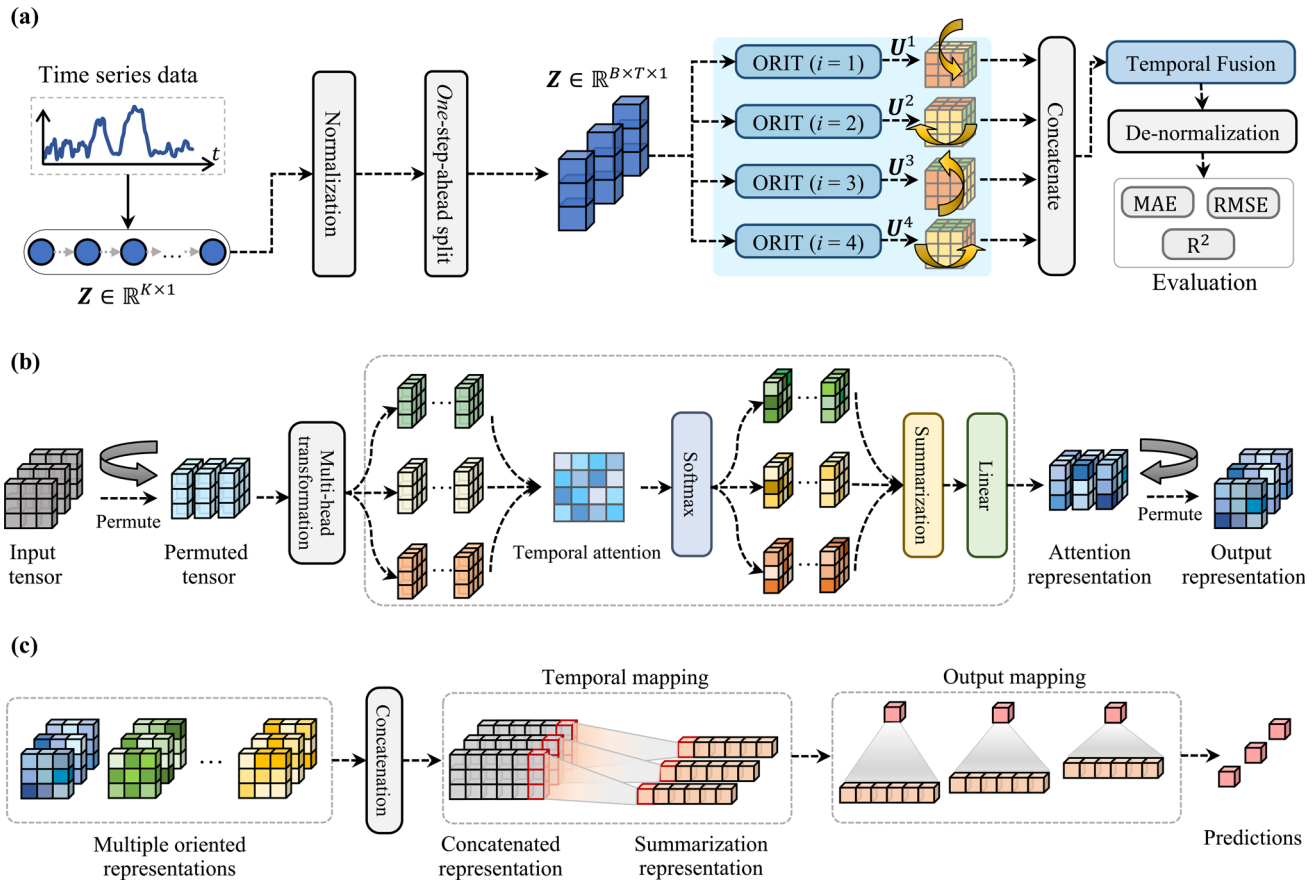


Fig. 1 The schematic illustration of the proposed oriented transformer (ORIT). (a) Workflow. (b) Multi-head oriented attention unit (MOAU). (c) Temporal fusion layer

model consists of three stages. The first stage is data preprocessing. The raw time series of infectious disease cases are normalized and spitted into supervised data for the prediction model. Subsequently, in the second stage, the proposed ORIT generates and combines the oriented attention from the processed inputs. Finally, the obtained combinations are fed into a temporal attention layer in the third stage. The model's outputs are subsequently de-normalized and evaluated using three distinct metrics. The pseudo-code describing the process of training ORIT is shown in Algorithm 1.

Table 1 Symbols and semantics

Symbol	Semantic
N	Total time steps number
T	Look-back window size
B	the number of consecutive look-back windows
$\boldsymbol{Z}$	Outpatient cases, $\boldsymbol{Z} \in \mathbb{R}^{N\times 1}$
$\boldsymbol{X}$	batched input tensor $\boldsymbol{X} \in \mathbb{R}^{B\times T\times 1}$
$\boldsymbol{Y}$	batched output tensor $\boldsymbol{Y} \in \mathbb{R}^{B\times T\times 1}$
K	Heads number
[;]	Concatenate operation
$\boldsymbol{C}$	Concatenated matrix
$\boldsymbol{P}$	Model outputs
$\mathcal{F}(\cdot)$	The predictive function
$\mathcal{A}(\cdot)$	The mapping function of attention
$\mathcal{A}^i(\cdot)$	The i-the orientation attention

4.1 Data preprocessing

Normalization To alleviate the impact of outliers on the model's learning process and foster faster convergence, we employ Min-Max normalization to scale the time series values into the [0, 1]. Compared to Z-Score normalization, Min-Max normalization offers a more straightforward computational process while preserving the original distribution of case data. The formulas for Min-Max normalization and de-normalization are as follows:

$$\boldsymbol{Z}' = \frac{\boldsymbol{Z} - \min(\boldsymbol{Z})}{\max(\boldsymbol{Z}) - \min(\boldsymbol{Z})}, \tag{3}$$

Algorithm 1 Pseudo-code for training the proposed ORIT model.

Input: The training set of historical case observations $\boldsymbol{Z}$, initialize model Φ
Output: The trained model Φ_{train}
// Feed forward and backward gradient updating
Trainer ($\boldsymbol{Z}$, Φ):
  $\boldsymbol{Z}' = \frac{\boldsymbol{Z}-\min(\boldsymbol{Z})}{\max(\boldsymbol{Z})-\min(\boldsymbol{Z})}$ ▷ normalization (3)
  $(\mathcal{X}, \mathcal{Y}) \leftarrow$ *one-step-forward* split $\boldsymbol{Z}'$ ▷ split (5)
  foreach *batch sample* $(\boldsymbol{X}, \boldsymbol{Y})$ *in* $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$ **do**
    $\boldsymbol{C} \leftarrow [\boldsymbol{X}]$
    foreach *orientation* $r \leftarrow 1$ *to* R **do**
      $\boldsymbol{X}^r \leftarrow$ transform the input x into r-th orientation
      $\boldsymbol{U}^r \leftarrow$ calculate the oriented attention representation using MOAU ($\boldsymbol{X}^r$)
      $\boldsymbol{C}_r \leftarrow [\boldsymbol{C}_{r-1}; \boldsymbol{U}^r]$ ▷ concatenation (11)
    end
    $p = (\sum^T \boldsymbol{W}_t * \boldsymbol{C}) \cdot \boldsymbol{W}_p + B_p$ ▷ prediction (12)
    Loss $\mathcal{L} \leftarrow p$ and y using MSE ▷ MSE error
    Backward using Adam optimizer
  end
  return Φ_{train}

$$\boldsymbol{Z} = \boldsymbol{Z}' * (\max(\boldsymbol{Z}) - \min(\boldsymbol{Z})) + \min(\boldsymbol{Z}), \tag{4}$$

where $\boldsymbol{Z}$ denotes the input samples, $\boldsymbol{Z}'$ is the normalized samples, $\min(\boldsymbol{Z})$ is the minimum value of the observed samples, and $\max(\boldsymbol{Z})$ is the maximum value of the observed samples.

Data organization The *one-step-forward* splitting approach [21] is utilized to split the normalized time series into supervised data, represented as $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$. For a given time series, the process can be formulated as follows:

$$\begin{bmatrix} Z'_{1,1} & Z'_{2,1} & \cdots & Z'_{T,1} \\ Z'_{2,1} & Z'_{3,1} & \cdots & Z'_{T+1,1} \\ \cdots & \cdots & \cdots & \\ Z'_{N-T-1,1} & Z'_{N-T,1} & \cdots & Z'_{N-1,1} \end{bmatrix} \rightarrow \begin{bmatrix} Z_{T+1,1} \\ Z_{T+2,1} \\ \cdots \\ Z_{N,1} \end{bmatrix}, \tag{5}$$

where the left part is the model input $\boldsymbol{\mathcal{X}}$, and the right part is the forecast target $\boldsymbol{\mathcal{Y}}$.

4.2 Oriented transformer

The proposed ORIT comprises a multi-head oriented attention Unit (MOAU) and a temporal fusion layer. The detailed structure and pseudo-code of the proposed MOAU is displayed in Fig. 1(b) and Algorithm 2, respectively.

The progression of infected cases incorporates various patterns of information, including short-term trends, long-term trends, rare events, and more. Feeding the preprocessed data directly into the predictive model might impede the model's ability to efficiently extract valuable information from the sequences. To simplify data representations for predictors, we have incorporated the orientated transformation component and attention representation component within the MOAU. These components generate attention

Algorithm 2 Pseudo-code of for the learning process of the proposed MOAU.

Input: The processed observations of ID x^r
Output: The oriented attention representation $\boldsymbol{U}^r$
MOAU ($\boldsymbol{X}^r$):
  foreach *attention head* k *to* K **do**
    $\boldsymbol{h}_k^r \leftarrow \boldsymbol{w}_k^2 \cdot x^s$ ▷ mapping (6)
    $\boldsymbol{s}_k^r \leftarrow \boldsymbol{W}^l \cdot \boldsymbol{h}_k^2$ ▷ calculate attention score (7)
    foreach *time steps* t *to* T **do**
      $m_{k,t}^r \leftarrow \frac{\exp(s_{k,t}^2)}{\sum_{t=1}^T \exp(s_{k,t})}$ ▷ normalization (8)
    end
    $\boldsymbol{o}_k^r \leftarrow \boldsymbol{W}^r \cdot \boldsymbol{m}_k^2$ ▷ projection (9)
  end
  $\boldsymbol{U}^r = \sum_{k=1}^K \boldsymbol{O}_k^r * \boldsymbol{W}_o$ ▷ aggregation (10)
  return $\boldsymbol{U}^r$

representations encompassing various dimensions, such as time steps, periodicity, and the intricate interconnections among diverse feature dimensions.

There are multiple strategies to extract features from various orientations of model inputs. A prevalent approach is to transform the input tensor, which enhances the diversity of input data without necessitating additional parameters. Furthermore, these transformed representations contribute to a more diverse set of detailed patterns, enabling the attention mechanism to capture deeper, more complex features.

Given a traditional attention mechanism and the model input $\boldsymbol{X} \in \mathbb{R}^{B\times T\times N}$, the context-aware representations are computed by an attention mechanism operating on the final two dimensions (i.e., $T \times N$). To implement the proposed oriented attention mechanism, the input tensor X is transformed into four types: $\{\boldsymbol{X}^1 \in \mathbb{R}^{B\times T\times N}; \boldsymbol{X}^2 \in \mathbb{R}^{B\times N\times T}; \boldsymbol{X}^3 \in \mathbb{R}^{N\times B\times T}; \boldsymbol{X}^4 \in \mathbb{R}^{T\times B\times N}\}$. The purposes associated with these four orientations are described as follows:

(1) *Orientation* 1 is employed to assess the impact of diverse time steps by establishing an association between time steps and a specified time interval.
(2) *Orientation* 2 is utilized to distinguish the impacts of different time series by constructing a correlation between the windowed time series and the feature dimension.
(3) *Orientation* 3 is employed to underscore the significance of specific time segments by establishing the relationship between these time segments and partial time steps.
(4) *Orientation* 4 is applied to discern the importance of different time segments by comparing across different time series.

The multi-head attention representation component is a crucial structure within MOAU, which consists of a multi-head transformation layer, an attention layer, and a fully connected layer. A symmetrical structure for the attention mechanism is designed to enhance the flexibility of the fusion

process across various oriented attentions. This structure generates attention representations in an end-to-end fashion while maintaining shape consistency between the input and output representations.

Due to the consistent learning processes across different oriented representations, we provide an example of the attention representation on orientation 2 to elucidate the workflow of MOAU. Specifically, the MOAU first projected the input supervised data into multi-head representations through a linear layer, and the process can be formulated as follows:

$$\boldsymbol{h}_i^2 = \boldsymbol{w}_i^2 \cdot \boldsymbol{Z}^2 \quad i \in (1, 2, ..., K), \tag{6}$$

where $\boldsymbol{h}_i^2$ represents a state of $\boldsymbol{H}^2$, $\boldsymbol{H}^2 \in \mathbb{R}^{B\times K\times 1\times T}$ denotes the multi-head representations, $\boldsymbol{w}_i^2 \in \mathbb{R}^{1\times T}$ is the learnable weight matrix, and K represents the number of attention modules. $\boldsymbol{Z}^2$ denotes the transformed observations of the infected cases data with respect to orientation 2.

To enhance the extraction of latent patterns from the obtained multi-head representations, an adaptive matrix is integrated, leveraging batch matrix multiplication operations:

$$\boldsymbol{s}_i^2 = \boldsymbol{W}^l \cdot \boldsymbol{h}_i^2 \quad i \in (1, 2, ..., K), \tag{7}$$

where $\boldsymbol{s}_i^2$ denotes a state of $\boldsymbol{S}^2$, $\boldsymbol{S}^2 \in \mathbb{R}^{B\times K\times T\times T}$ is the attention score, $\boldsymbol{W}^l \in \mathbb{R}^{T\times 1}$ represents a weight matrix. From the perspective of orientation 2, the attention scores indicate the relative importance of various time series in predicting the target variable. A normalization layer is employed to emphasize the key elements of score representations and enhance the distinction among individual scores. The softmax function is employed within the normalization layer, serving to amplify the distinctions within a target dimension by considering its contributions. This process is formulated as follows:

$$m_{i,t}^2 = \frac{\exp(s_{i,t}^2)}{\sum_{t=1}^{T}\exp(s_{i,t})} \quad i \in (1, 2, ..., K), \tag{8}$$

where $\boldsymbol{m}_i^2$ is a state of $\boldsymbol{M}^2$, $\boldsymbol{M}^2 \in \mathbb{R}^{B\times K\times T\times T}$ is the normalized tensor.

Once the normalized scores are obtained, these results can be utilized to construct multi-head attention representations through a projection process:

$$\boldsymbol{o}_i^2 = \boldsymbol{W}^r \cdot \boldsymbol{m}_i^2, \quad i \in (1, 2, ..., K), \tag{9}$$

where $\boldsymbol{o}_i^2$ represents a state of $\boldsymbol{O}^2$, where $\boldsymbol{O}^2 \in \mathbb{R}^{B\times K\times 1\times T}$ is the multi-head attentions tensor, and $\boldsymbol{W}^r \in \mathbb{R}^{1\times T}$ is the linear weight matrix.

The tensors of multi-head attentions are passed through an aggregation layer to summarize the information of multi-head representations. The formulation of this process is as follows:

$$\boldsymbol{U}^2 = \sum_{k=1}^{K} \boldsymbol{O}_{b,k,i,t}^2 * \boldsymbol{W}_o, \tag{10}$$

where $\boldsymbol{U}^2 \in \mathbb{R}^{B\times 1\times T}$ is the oriented attention representations of orientation 2, $\boldsymbol{W}_o$ is a learnable weigh matrix, $*$ denotes the element-wise multiplication. The dimensions of the input representation and the output oriented attention representations are consistent.

4.3 Temporal fusion

To aggregate the context of the obtained oriented attention representations and produce predictions, a temporal fusion layer is incorporated within ORIT, as illustrated in Fig. 1(c). It is noteworthy that the original transformed data are also taken into account within the temporal fusion layer. This consideration is intended to prevent information loss and to enhance the propagation of gradients.

$$\boldsymbol{C} = [\boldsymbol{U}^1; \boldsymbol{U}^2; \boldsymbol{U}^3; \boldsymbol{U}^4; \boldsymbol{X}], \tag{11}$$

where $\boldsymbol{C} \in \mathbb{R}^{B\times T\times 5}$ is the concatenated representations, and [;] denotes the concatenation operation.

Lastly, the information regarding potential transmission patterns within the temporal dimension is summarized, enabling the prediction of the normalized progression of infected cases in the next interval. The process is formulated as follows:

$$\boldsymbol{P} = (\sum^{T} \boldsymbol{W}_t * \boldsymbol{C}) \cdot \boldsymbol{W}_p + B_p, \tag{12}$$

where $\boldsymbol{W}_t \in \mathbb{R}^{T\times 1}$ represents the learnable weight matrix responsible for projecting the temporal information. $\boldsymbol{W}_p \in \mathbb{R}^{5\times 1}$ denotes the weight matrix used to generate the prediction, and $B_p \in \mathbb{R}$ is a bias term. The predictions $\boldsymbol{P}$ are de-normalized using (4) to yield the final results.

5 Experimental settings

This section provides an overview of the experimental dataset, evaluation metrics, baseline methodologies, and model configuration.

5.1 Datasets

In total, 49,677 records of HFMD outpatient cases and 48,359 records of HB outpatient cases were collected to evaluate the proposed ORIT method and benchmark methods. The HB

outpatient cases were reported when the transaminase levels exceeded twice the standard. Consequently, there may be some errors in the HB outpatient case data. The distribution of the two datasets is illustrated in Fig. 2. These data were collected from January 5, 2011, to January 28, 2021, spanning a total of 2,184 days (312 weeks). The basic statistics of the two datasets are shown in Table 2.

The outpatient cases dataset, shared by the Xiamen Center for Disease Control and Prevention (XMCDC), is utilized to conduct experiments on both the proposed method and other benchmark methods. This dataset is bifurcated into two subsets for this purpose. The first subset, spanning the period from January 5, 2015, to January 23, 2020, is allocated for model training, with the last 31 samples employed as the validation set. The remaining data is utilized to test the trained models.

5.2 Baseline methods

To study the effectiveness of the proposed method, several methods have been developed and applied to the two real-world datasets. To study the benefits of the proposed attention mechanism, these models have been extended with an attention mechanism. Hence, those benefits can be observed by measuring the prediction performance of models. The comparable models are listed as follows:

(1) Auto-Regression (AR) [26] builds linear relationships between past observations and coming values.
(2) Long short-term memory (LSTM) [18] is a variation of the recurrent neural network (RNN), which exploits three gate units to capture the temporal dependency.
(3) Gated Recurrent Unit (GRU) [13] merged the hidden states and cell states of LSTM to reduce parameters.
(4) Encoder-Decoder (ED) [33] consisted of two RNN components in the encoder stage and decoder stage, respectively.
(5) Convolutional Neural Network (CNN1D) [44] extracts temporal patterns of sequential data and uses a fully connected layer to generate predictions.
(6) CNNRNN [44] extracted local sequential patterns to generate predictions by leveraging a connected CNN and RNN network.

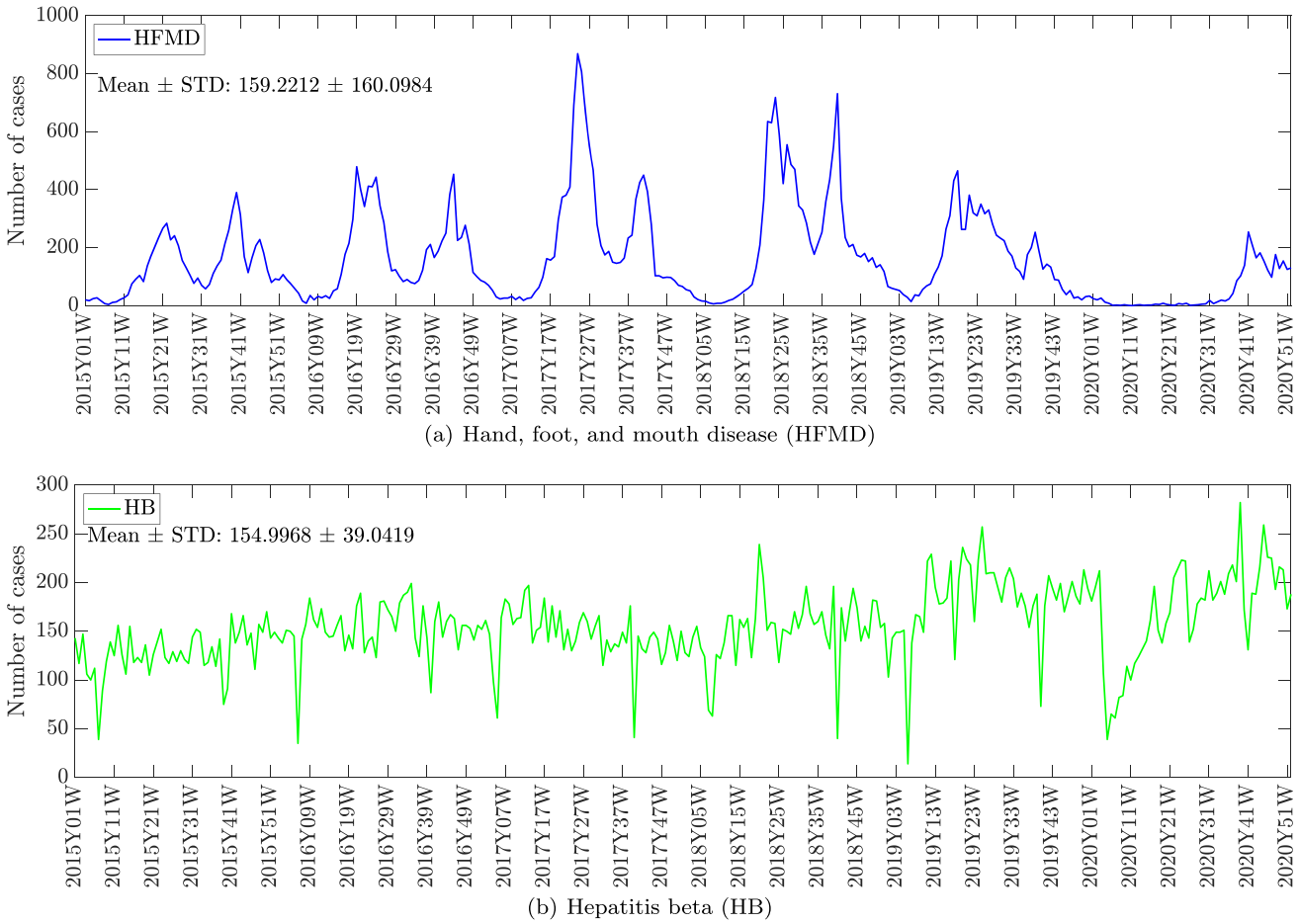


Fig. 2 The distributions of weekly hand, foot, and mouth disease (HFMD) outpatient counts and weekly hepatitis beta (HB) outpatient counts

Table 2 The basic description on HFMD and HB datasets. "STD" denotes standard deviation

Dataset	Training size	Validate size	Test size	Maximum	Average	Minimum	STD
HFMD	228	31	53	869	159.22	0	159.84
HB	228	31	53	282	154.99	14	38.98

(7) Oriented attention model (OAM) [48] generate predictions via consolidating attentions from several aspects. Moreover, the other score function integrated into the OAM, such as dot product, scaled dot product, MLP, and multi-head attention. And then we have OAM-SA(dot), OAM-SA(scaled dot), OAM-SA(score), and OAM(MA).

Moreover, the proposed symmetric attention mechanism has been integrated into comparable methods. As a result, we obtain variants such as LSTM-attn, GRU-attn, ED-attn, CNN-attn, and CNNRNN-attn.

5.3 Evaluation metrics

To appraise the prediction results from both the proposed ORIT and the benchmark methods, the mean absolute error (MAE), root mean square error ($RMSE$), and the coefficient of determination (R^2) are employed as our evaluation criteria. MAE quantifies the mean absolute deviation of the prediction errors, RMSE accounts for the variance in errors, and R^2 assesses the degree of fit between the predicted and actual data [41]. Each metric focuses on different aspects, thus offering a comprehensive evaluation. These criteria can be expressed in the following mathematical expressions:

$$MAE = \frac{1}{L}\sum_{i=1}^{L}|Z_i - \hat{Z}_i|, \tag{13}$$

$$RMSE = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(Z_i - \hat{Z}_i)^2}, \tag{14}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{L}(Z_i - \hat{Z}_i)^2}{\sum_{i=1}^{L}(Z_i - \bar{Z})^2}, \tag{15}$$

where Z_i and $\hat{Z}_i$ represent the actual and predicted values, respectively. $\bar{Z}$ signifies the average value of the test set, and L denotes the length of time steps within the test set. The best model is the performance with the smallest MAE, $RMSE$, and the largest R^2.

5.4 Configurations

To ensure a fair comparison, all training-related constant parameters are uniformly set across the various methods on a given dataset. Each method is optimized using the Adam optimizer [22], with the loss function defined as the mean squared error (MSE). The number of training epochs and the learning rate are tuned to achieve optimal performance for each method. The weights of prediction models are preserved upon achieving an optimal state and are subsequently loaded during testing procedures. To enhance the reliability and robustness of the presented results, each experiment is replicated five times, with the presented results reflecting the mean values. The grid search approach is employed to search for the relatively optimal hyperparameters, and detailed settings for the hyperparameters of each method are provided in Table 3. The models were implemented using PyTorch (v.1.12.1) [28]. All computations are performed on a server equipped with an Intel Xeon Gold 6226R CPU (2.90GHz) and 128GB of memory, with processing accelerated using two NVIDIA RTX A6000 GPUs.

6 Results

This section conducts comprehensive experiments to evaluate the effect of model parameters on the prediction

Table 3 Hyper-parameter settings

Model	Parameter	Option range
LSTM	Embedding hidden size	$\{2^5, 2^6\}$
GRU		
ED	Number of layers	1-3 (1 per step)
CNN1D	Out channel	$\{2^5, 2^6\}$
CNNRNN	Kernel size	3-9 (2 per step)
CNNRNN	Embedding hidden size	$\{2^5, 2^6\}$
	Number of layers	1-2 (1 per step)
OAM		Dot product
	Attention	Scaled dot product
	score function	MLP
		MA
ORIT	Numbers of heads	$\{2^1, 2^2, 2^3, 2^4\}$
	Embedding hidden size	$\{2^5, 2^6\}$
General	Learning rate	5e-04 – 5e-03
		(5e-04 per step)
	Learning epochs	100 – 1600
		(300 per step)

performance and to compare the proposed method with baseline approaches. The major intentions are listed as follows:

(1) How does the extent or range of temporal characteristics influence the accuracy of predictions within the model?
(2) How do different attention representations affect prediction accuracy?
(3) Could the proposed attention mechanism outperform the other implementations?
(4) Could the proposed ORIT outperform the other methods?

6.1 Effects on parameter *B*

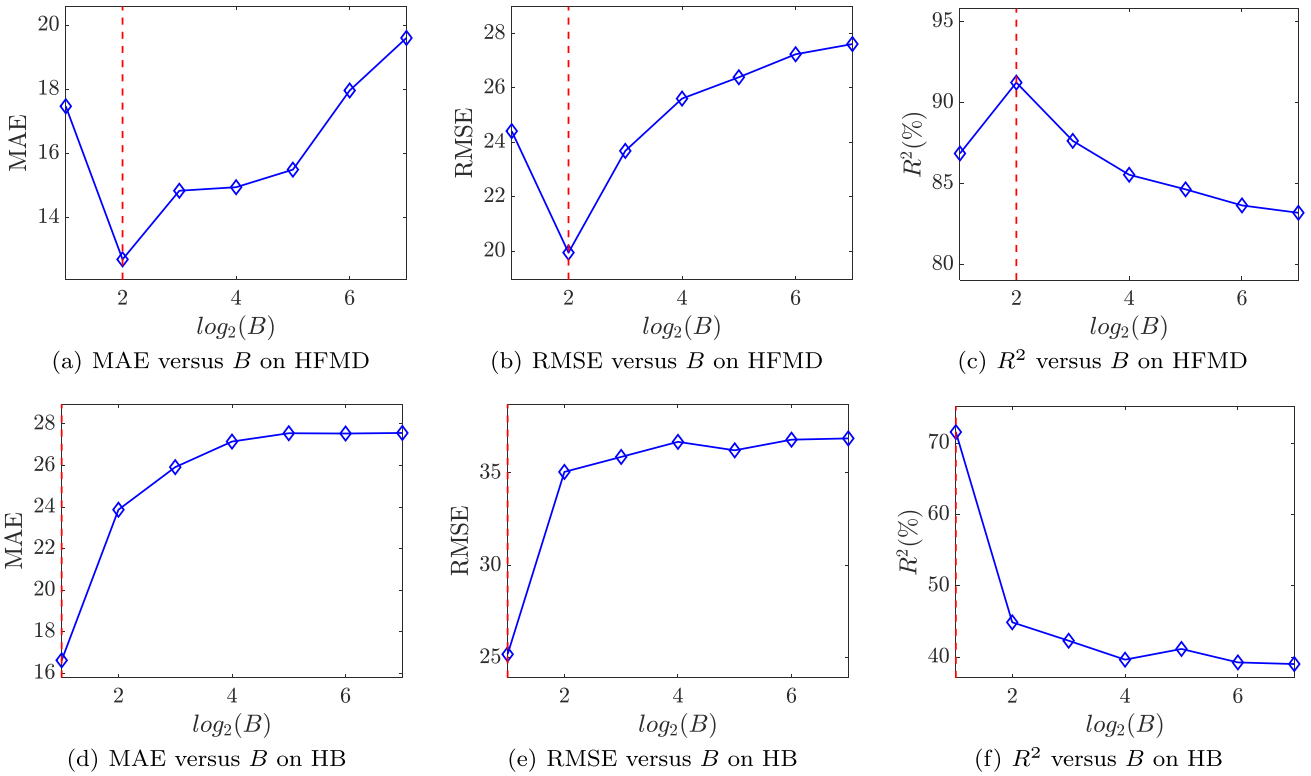
The performance of the prediction model is significantly influenced by the selection of appropriate parameters [2]. Therefore, it is essential to carefully evaluate and choose suitable hyperparameters B and T.

The term B is a constant parameter in the proposed ORIT, representing the number of time slices within a subset. Specifically, each time slice denotes a continuous period of disease progression. The parameter B defines the number of adjacent periods from which the predictive model can acquire knowledge. Insufficient information about period patterns can limit the fitting performance of a regression model, while excessive pattern detail may introduce redundant information, thereby limiting predictive accuracy. Hence, establishing an appropriate scope for the period is critical for the prediction model. While observing the performance of the prediction, the target parameter B is tuned while the others are held constant. For HFMD and HB datasets, parameter T is fixed at 13 and 4, respectively.

The investigation of the parameter B is plotted in Fig. 3. Several key observations are summarized as follows:

(1) For the HFMD dataset, optimal performance in terms of MAE, $RMSE$, and R^2 is achieved when $B = 4$.
(2) For the HB dataset, the optimal performance in terms of MAE, $RMSE$, and R^2 are found at $B = 2$.
(3) The prediction performance substantially deteriorates when the parameter B is either less or greater than the optimal value.

For the HFMD dataset, see Fig. 3(a)-(c), the optimal performance is found at $B = 4$. These results approximate the duration during which Enterovirus (EV), the primary pathogen causing HFMD, persists in infections [45]. For the HB dataset, as shown in Fig. 3(d)-(f), the optimal performance is found at $B = 2$. This indicates that the temporal dynamics across two consecutive periods provide the most significant correlation to the prediction target. The selected range for the period aligns closely with the typical duration



(a) MAE versus B on HFMD (b) RMSE versus B on HFMD (c) R^2 versus B on HFMD

(d) MAE versus B on HB (e) RMSE versus B on HB (f) R^2 versus B on HB

Fig. 3 The performance of the ORIT with varying the parameter B in terms of MAE, $RMSE$ and R^2. For each metric the optimal value is found at red dash line

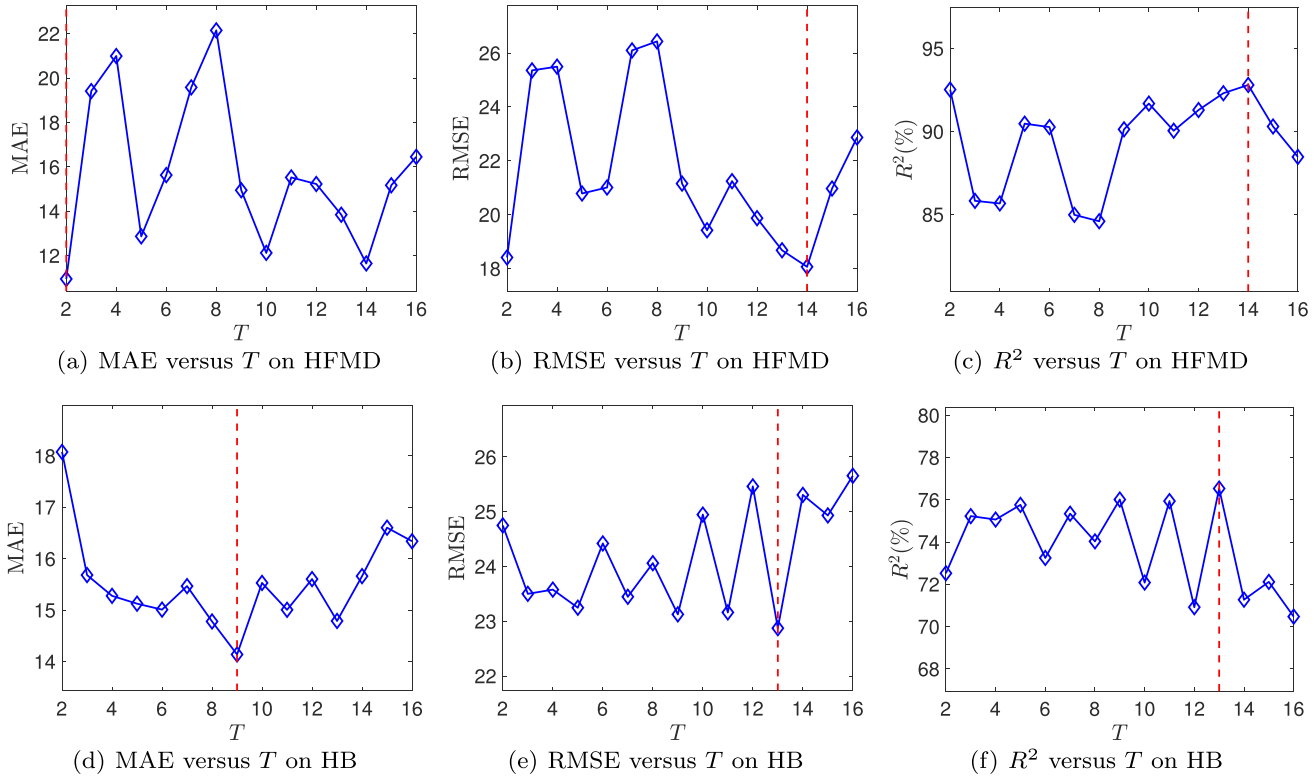


Fig. 4 The performance of the ORIT with varying the parameter T in terms of MAE, $RMSE$ and R^2. For each metric the optimal value is found at red dash line

from the onset of symptoms to the development of acute infection in patients [38].

6.2 Effects on parameter *T*

To further investigate the influence of the incubation period and seasonality on the prediction of infectious disease cases, the parameter T is also incorporated into the parameter analysis of this study. The experimental results are visualized in Fig. 4. For HFMD and HB datasets, parameter B is fixed at 4 and 2, respectively.

The major observations from experimental results are summarized as follows:

(1) For the HFMD dataset, the optimal value for MAE is observed when $T = 2$, whereas the optimal values for $RMSE$ and R^2 are observed when $T = 14$.
(2) The prediction performance demonstrates significant fluctuation with variations in the parameter T.
(3) For the HB dataset, the optimal value for MAE is observed when $T = 9$, while the optimal values for $RMSE$ and R^2 are observed when $T = 13$.
(4) The prediction performance exhibits a periodic fluctuation for approximately two weeks.

For the HFMD dataset, the optimal values of $RMSE$ and R^2 and the second-best value of MAE are found at $T = 14$. These results align with the observed cyclical and seasonal trends prevalent in real-world HFMD outbreaks. The best MAE value and the second best values for $RMSE$ and R^2 are observed at $T = 2$, which coincides with the HFMD incubation period typically within two weeks. However, these results show slight deviation from earlier empirical analyses [47]. This discrepancy could potentially be ascribed to the time lag between the onset of symptoms and medical consultation.

Notable fluctuations in performance are observed when varying the parameter T. This phenomenon may be related to changes in the outbreak period. Prior studies have highlighted that the HFMD infection outbreak period varies seasonally [40].

For the HB dataset, the optimal MAE value is observed at $T = 9$, whereas the best values for $RMSE$ and R^2 are obtained at $T = 13$. Since HBV exhibits no significant seasonality, the obtained results largely relate to the incubation periods of HB. The potential incubation periods for this dataset are 9 or 13 weeks. Both human behavior and environmental conditions can introduce uncertainty when attempting to identify the incubation period from historical records. Notably, the prediction performance exhibits

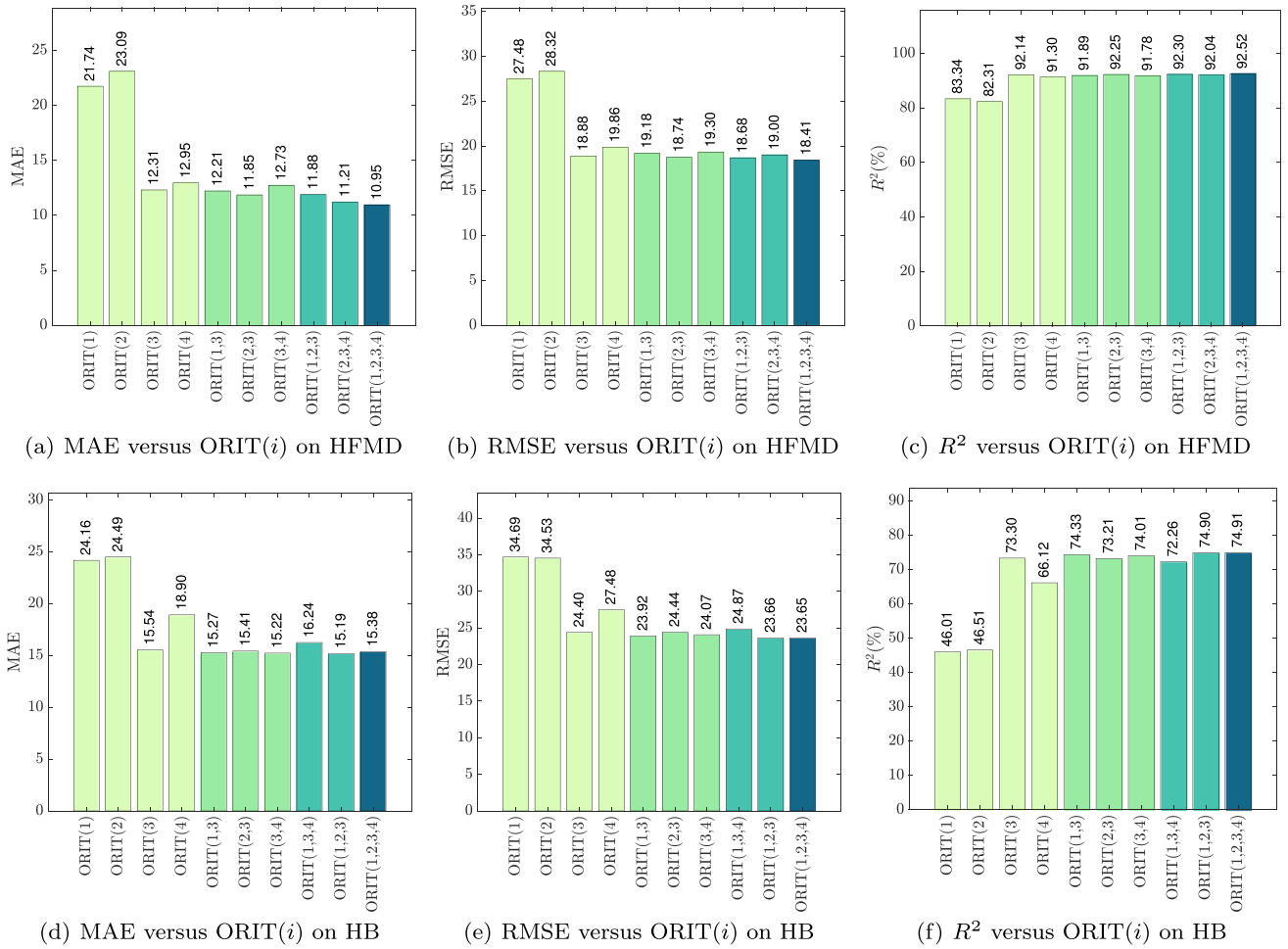
slight periodic fluctuations within a 2-week span, especially in terms of $RMSE$ and R^2. This may be attributed to the influence of the medical window period, which refers to the duration between the entry of a virus or bacteria into the body and its accurate detection by conventional medical testing methods.

6.3 Study on attention combination

To investigate the effects of different orientations of attention on prediction performance and explore the most effective combinations, comprehensive experiments were conducted using various combinations of MOAU. In order to avoid unnecessary experimentation with all possible MOAU combinations, the experiment employed a greedy search strategy. Firstly, the effectiveness of each individual oriented attention was analyzed. Subsequently, the optimal orientation was selected to combine with each of the other orientations. The current optimal combination was iteratively added with each remaining orientation until all orientations were included. For clarity in presentation, let ORIT(i) represent the observation in the orientation set i. The experimental results are plotted in Fig. 5. The key observations derived from Fig. 5 are summarized as follows:

(1) For both HFMD and HB datasets, ORIT(1, 2, 3, 4) achieves the optimal values for MAE, $RMSE$, and R^2.
(2) From the perspective of single oriented attention, ORIT(3) achieves the best performance, while ORIT(4) obtains the second-best performance.
(3) To integrate the ORIT(3) and ORIT(4), the prediction accuracy improve significantly.
(4) ORIT(1) and ORIT(2) have slight effects on improving the performance.

For both prediction tasks of the two infectious diseases, MRF demonstrates a promising approach to enhance prediction accuracy. ORIT(1, 2, 3, 4) achieves optimal performance in terms of MAE, $RMSE$, and R^2, thereby validating the benefits of incorporating multi-oriented correlations in prediction performance.



(a) MAE versus ORIT(i) on HFMD (b) RMSE versus ORIT(i) on HFMD (c) R^2 versus ORIT(i) on HFMD

(d) MAE versus ORIT(i) on HB (e) RMSE versus ORIT(i) on HB (f) R^2 versus ORIT(i) on HB

Fig. 5 The ORIT performance on combinations of the multi-head oriented attentions in terms of MAE, $RMSE$ and R^2

Each individual orientation attention exhibits varying predictive performance, indicating the diversity of the transformed input. ORIT(3) and ORIT(4) outperform ORIT(1) and ORIT(2) significantly. This can be attributed to the fact that ORIT(3) and ORIT(4) emphasize the significance of time segments for the prediction model, as they contain the context of consecutive time steps and furnish more detailed information about trends. In contrast, ORIT(1) and ORIT(2) primarily consider the correlation between different time series and specific time steps, enabling them to capture the local temporal dynamics of each time series or time step. However, they neglect the continuity of trend characteristics.

6.4 Comparison

Extensive experiments were conducted to compare the proposed ORIT method with 24 benchmark methods, as shown in Table 4. To ensure a fair comparison, the hyperparameters B and T were set to 4 and 2, respectively, for the HFMD dataset, and 2 and 13 for the HB dataset. The main observations are summarized as follows:

(1) The proposed ORIT achieves optimal performance in terms of three metrics for both datasets.
(2) The performance of all benchmark methods is further improved by the proposed attention mechanism.
(3) The number of hidden neurons does not significantly affect the performance of RNN-related models.
(4) The proposed symmetric structure attention mechanism outperforms other structures.
(5) AR exhibits the worst performance for the HFMD dataset, while CNN1D exhibits the lowest performance for the HB dataset.

Among the benchmark methods, the AR exhibits the lowest prediction accuracy for the HFMD dataset and the second-lowest performance for the HB dataset. This may be attributed to AR's limited capability in modeling non-linear patterns. The temporal dynamics of both infectious diseases are irregular and exhibit fluctuations. These characteristics may not be adequately captured by the AR method. The RNN variants, i.e., LSTM, GRU, and ED, incorporated with temporal learning units, demonstrate slightly better performance than the linear model. However, it should be noted that these

Table 4 Comparable results of twenty-two methods on the two datasets in terms of three metrics

Model	HFMD			HB		
	MAE	RMSE	R^2	MAE	RMSE	R^2
AR	22.6962	28.0246	0.8268	27.0550	37.1552	0.4550
LSTM-32	15.8524	25.6851	0.8545	26.1595	37.6786	0.4396
LSTM-attn-32	14.2515	24.5761	0.8668	23.3605	32.0211	0.5952
LSTM-64	15.9730	25.6855	0.8545	26.6344	37.7522	0.4374
LSTM-attn-64	13.5908	24.3641	0.8691	23.2384	32.0197	0.5953
GRU-32	19.0589	26.3326	0.8471	26.1140	36.8747	0.4632
GRU-attn-32	13.6181	24.3916	0.8688	22.7751	31.1019	0.6181
GRU-64	19.5851	26.5899	0.8441	26.1305	36.9358	0.4615
GRU-attn-64	13.2111	24.2959	0.8698	22.8632	31.1539	0.6169
ED-32	18.5562	26.2696	0.8478	26.2258	37.3032	0.4507
ED-attn-32	15.1897	25.4855	0.8568	23.1548	31.4827	0.6087
ED-64	17.4964	25.8206	0.8530	25.6824	37.0903	0.4569
ED-attn-64	14.8598	25.4567	0.8571	22.6006	30.6012	0.6300
CNN1D	20.9972	27.1071	0.8379	26.4590	37.2826	0.4513
CNN1D-attn	16.1284	25.9156	0.8519	26.0703	37.0805	0.4572
CNNRNN-32	19.5186	26.6080	0.8439	27.1456	36.5898	0.4715
CNNRNN-attn-32	16.1384	25.8667	0.8524	25.8596	37.1470	0.4553
CNNRNN-64	18.7608	26.3040	0.8474	26.1287	36.4199	0.4764
CNNRNN-attn-64	15.4331	25.6848	0.8545	26.0657	37.2316	0.4528
OAM-SA(dot)	23.2935	28.3903	0.8222	27.8437	37.5277	0.4441
OAM-SA(scale-dot)	22.9140	28.0769	0.8261	26.7775	37.0830	0.4572
OAM-SA(MLP)	19.3418	25.7491	0.8538	24.0594	35.0306	0.5156
OAM(MA)	14.3342	21.1728	0.9011	21.6811	29.2421	0.6625
OAM	10.9700	16.9735	0.9364	21.7509	28.9388	0.6692
ORIT	**10.2614**	**16.8450**	**0.9374**	**21.3500**	**28.2686**	**0.6846**

methods show insensitivity to the number of hidden neurons. A possible reason for this is that increasing the number of hidden neurons allows for more information storage, but due to the limited data scale, it does not lead to significant improvements.

The convolution component can capture high-level features by stacking multiple convolutional and pooling layers. However, an inherent limitation in extracting temporal dynamics leads to worse performance in predicting infectious disease cases. By integrating the RNN cell, the CNNRNN can model temporal dependencies and improve performance. However, the accuracy achieved is still unsatisfactory. These results can be attributed to the CNN's ability to disrupt the continuity of the input data, thereby posing a challenge for the recurrent component to capture meaningful information. With the assistance of the attention mechanism, the $RMSE$ values of benchmarks in the HFMD and HB datasets are reduced by up to 7.4% and 4.2%, respectively. This demonstrates the efficacy of the proposed attention mechanism.

Among the OAM and its variations, the OAM-SA(dot) exhibits the worst performance, while the OAM outperforms the other variations. A possible reason for this discrepancy is that the dot product is unsuitable for attention collaboration and fusion. The attention mechanism using the MLP score function obtains better performance, whereas the OAM has an obvious decrease by 19.8% and 1% in $RMSE$ value. These findings highlight the effectiveness of the symmetric structure attention mechanism in improving prediction accuracy.

The proposed ORIT is essentially a variant of OAM. Compared with OAM, it demonstrates a decrease of 6.46% and 0.8% in MAE and $RMSE$ values, respectively, on the HFMD dataset. Similarly, on the HB dataset, it shows a decrease of 1.84% and 2.3% in MAE and $RMSE$ values, respectively. These results suggest that the proposed MRF, with multiple orientation attentions, is more effective in capturing temporal dynamics.

7 Challenges and opportunities

After discussing the main advantages of utilizing the ORIT method for forecasting infectious cases, there are several potential directions to further enhance its effectiveness.

(1) Deep learning, shallow machine learning, and statistical learning method are widely applied in infectious disease case prediction. However, selecting the appropriate hyper-parameter combination is still an issue. Automated deep learning (ADL) strategies are powerful technologies that streamline the process of identifying optimal parameters [3]. The development of a hybrid system by combining the ADL framework with the ORIT method to enhance its accuracy could be an interesting avenue to explore. It would facilitate the ORIT performance on other scenarios, such as outbreak period tracking [34], and multi-area monitoring [8].

(2) The interpretation of results in ORIT can be challenging, as it relies on black-box models that do not provide clear explanations of the inference process [5]. To enhance the reliability and trustworthiness of the obtained results, the incorporation of explainable artificial intelligence (XAI) techniques can benefit the proposed ORIT method. This would enable a more accurate evaluation of ORIT's outputs and facilitate the exploration of its underlying mechanisms. Moreover, integrating XAI can facilitate the extension of ORIT to high-dimensional time series forecasting tasks.

(3) To enhance the robustness of the proposed ORIT method for different scenarios, it is necessary to consider relevant exogenous features. However, the data distribution of these exogenous features may be inconsistent with the target data, and their temporal characteristics may be asynchronous. Therefore, innovative feature selection and analysis methods need to be explored to address these challenges [1]. The effective inclusion of external features would introduce greater sample diversity, enabling the prediction model to adapt more efficiently to various scenarios and changing conditions.

(4) Real-time capability is essential for effective control and management of infectious diseases [31]. Online prediction methods offer significant advantages in terms of speed, real-time capability, dynamic updates, and consideration of real-time variables. By incorporating the proposed ORIT model into online monitoring systems for infectious diseases, decision-makers can promptly respond to the spread and control of multiple diseases.

8 Conclusion

This research proposed a multiple representation fusion model, which is named oriented transformer (ORIT), for infectious disease case prediction. To enrich the diversity of historical observations in the time series of infected cases, ORIT proposes an orientation representation structure, which considers the multi-dimensional correlations among temporal or spatial characteristics. These correlations are learned through an elaborated symmetrical attention unit, termed MOAU. Comprehensive experiments on real-world HFMD and HB datasets demonstrate the efficacy of the proposed ORIT. A combination study verifies the effectiveness of MRF within ORIT modules, while a sensitivity analysis is employed to assess the influence of key hyperparameters.

Acknowledgements This work was supported in part by the Natural Science Foundation of Fujian Province (CN) (nos. 2021J01857, 2021J01859, and 2022J01335). Thanks to the Xiamen City Center for Disease Control and Prevention (XMCDC) for sharing the data. We gratefully appreciate the editor and anonymous reviewers for their valuable insights and suggestions which enormously benefited this paper.

Data Availibility The data used to support the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of interest None.

References

1. Abbaszadeh Shahri A, Asheghi R, Khorsand Zak M (2021) A hybridized intelligence model to improve the predictability level of strength index parameters of rocks. Neural Comput Appl 33:3841–3854. https://doi.org/10.1007/s00521-020-05223-9
2. Abbaszadeh Shahri A, Shan C, Zäll E, Larsson S (2021) Spatial distribution modeling of subsurface bedrock using a developed automated intelligence deep learning procedure: A case study in sweden. J Rock Mech Geotechnical Eng 13(6):1300–1310. https://doi.org/10.1016/j.jrmge.2021.07.006
3. Abbaszadeh Shahri A, Shan C, Larsson S (2022) A novel approach to uncertainty quantification in groundwater table modeling by automated predictive deep learning. Nat Resour Res 31(3):1351–1373. https://doi.org/10.1007/s11053-022-10051-w
4. Alamo T, Reina DG, Gata PM, Preciado VM, Giordano G (2021) Data-driven methods for present and future pandemics: Monitoring, modelling and managing. Annu Rev Control 52:448–464. https://doi.org/10.1016/j.arcontrol.2021.05.003
5. Asheghi R, Hosseini SA, Saneie M, Shahri AA (2020) Updating the neural network sediment load models using different sensitivity analysis methods: a regional application. J Hydroinf 22(3):562–577. https://doi.org/10.2166/hydro.2020.098
6. Bai S, Kolter JZ, Koltun V (2018) An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271
7. Bandara K, Bergmeir C, Hewamalage H (2020) LSTM-MSNet: Leveraging Forecasts on Sets of Related Time Series With Multiple Seasonal Patterns. IEEE Trans Neural Netw Learning Syst pp 1–14, https://doi.org/10.1109/TNNLS.2020.2985720
8. Bracher J, Held L (2022) Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. Int J Forecast 38(3):1221–1233
9. Brauwers G, Frasincar F (2023) A general survey on attention mechanisms in deep learning. IEEE Trans Knowl Data Eng 35(4):3279–3298. https://doi.org/10.1109/TKDE.2021.3126456
10. Chang SL, Piraveenan M, Pattison P, Prokopenko M (2020) Game theoretic modelling of infectious disease dynamics and intervention methods: a review. J Biol Dyn 14(1):57–89. https://doi.org/10.1080/17513758.2020.1720322
11. Chaudhari S, Mithal V, Polatkan G, Ramanath R (2021a) An attentive survey of attention models. ACM Trans Intell Syst Technol 12(5):53:1 – 53:32, https://doi.org/10.1145/3465055
12. Chaudhari S, Mithal V, Polatkan G, Ramanath R (2021) An attentive survey of attention models. ACM Trans Intell Syst Technol 12(5):1–32. https://doi.org/10.1145/3465055
13. Cho K, van Merrienboer B, Gülçehre Ç, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, ACL, Doha, Qatar, pp 1724–1734, https://doi.org/10.3115/v1/d14-1179
14. Djenouri Y, Belhadi A, Srivastava G, Lin JCW (2021) Secure collaborative augmented reality framework for biomedical informatics. IEEE J Biomed Health Inform 26(6):2417–2424. https://doi.org/10.1109/JBHI.2021.3139575
15. Du L, Gao R, Suganthan PN, Wang DZ (2022) Bayesian optimization based dynamic ensemble for time series forecasting. Inf Sci 591:155–175. https://doi.org/10.1016/j.ins.2022.01.010
16. Efimov D, Ushirobira R (2021) On an interval prediction of covid-19 development based on a seir epidemic model. Annu Rev Control 51:477–487. https://doi.org/10.1016/j.arcontrol.2021.01.006
17. Hewamalage H, Bergmeir C, Bandara K (2021) Recurrent neural networks for time series forecasting: Current status and future directions. Int J Forecast 37(1):388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008
18. Hochreiter S, Schmidhuber J (1997) Long Short-term Memory. Neural Comput 9:1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
19. Hong J, Liu F, Qi H, Tu W, Ward MP, Ren M, Zhao Z, Su Q, Huang J, Chen X, Le J, Ren X, Hu Y, Cowling B, Li Z, Chang Z, Zhang Z (2022) Changing epidemiology of hand, foot, and mouth disease in china, 2013 2019: a population-based study. The Lancet Regional Health - Western Pacific 20:100370. https://doi.org/10.1016/j.lanwpc.2021.100370
20. Huang S, Wang D, Wu X, Tang A (2019) Dsanet: Dual self-attention network for multivariate time series forecasting. In: Proceedings of the 28th International Conference on Information and Knowledge Management, ACM, Beijing, China, pp 2129 – 2132, https://doi.org/10.1145/3357384.3358132
21. Huang Y, Zhang P, Wang Z, Lu Z, Wang Z (2023) HFMD cases prediction using transfer one-step-ahead learning. Neural Process Lett 55(3):2321–2339. https://doi.org/10.1007/s11063-022-10795-9
22. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations, OpenReview.net, San Diego, CA, USA
23. Lai G, Chang W, Yang Y, Liu H (2018) Modeling long- and short-term temporal patterns with deep neural networks. In: Proceedings of the 41st International Conference on Research and Development in Information Retrieval, ACM, Ann Arbor, MI, USA, pp 95 – 104, https://doi.org/10.1145/3209978.3210006
24. Lim B, Zohren S (2021) Time-series forecasting with deep learning: a survey. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 379(2194):20200209. https://doi.org/10.1098/rsta.2020.0209
25. Lin JCW, Djenouri Y, Srivastava G, Fourier-Viger P (2022) Efficient evolutionary computation model of closed high-utility itemset mining. Appl Intell 52(9):10604–10616. https://doi.org/10.1007/s10489-021-03134-3
26. Mabrouk AB, Abdallah NB, Dhifaoui Z (2008) Wavelet decomposition and autoregressive model for time series prediction. Appl Math Comput 199(1):334–340. https://doi.org/10.1016/j.amc.2007.09.067
27. Mora C, McKenzie T, Gaw IM, Dean JM, von Hammerstein H, Knudson TA, Setter RO, Smith CZ, Webster KM, Patz JA et al (2022) Over half of known human pathogenic diseases can be aggravated by climate change. Nat Clim Chang 12(9):869–875. https://doi.org/10.1038/s41558-022-01426-1
28. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang EZ, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: An imperative style, high-performance deep learning library. In: Proceddings of the 33rd

Annual Conference on Neural Information Processing Systems, Vancouver, BC, Canada, vol 32, pp 8024–8035
29. Qin Y, Song D, Chen H, Cheng W, Jiang G, Cottrell GW (2017) A dual-stage attention-based recurrent neural network for time series prediction. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, ijcai.org, Melbourne, Australia, pp 2627 – 2633, https://doi.org/10.24963/ijcai.2017/366
30. Srivastava G, Lin JCW, Pirouz M, Li Y, Yun U (2020) A pre-large weighted-fusion system of sensed high-utility patterns. IEEE Sens J 21(14):15626–15634. https://doi.org/10.1109/JSEN.2020.2991045
31. Stockdale JE, Liu P, Colijn C (2022) The potential of genomics for infectious disease forecasting. Nat Microbiol 7(11):1736–1743
32. Subissi L, von Gottberg A, Thukral L, Worp N, Oude Munnink BB, Rathore S, Abu-Raddad LJ, Aguilera X, Alm E, Archer BN et al (2022) An early warning system for emerging sars-cov-2 variants. Nat Med 28(6):1110–1115. https://doi.org/10.1038/s41591-022-01836-w
33. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Proceedings of the 28th Annual Conference on Neural Information Processing Systems, MIT Press, Montreal, Quebec, Canada, vol 27, pp 3104 – 3112
34. Sweileh WM (2022) Global research activity on mathematical modeling of transmission and control of 23 selected infectious disease outbreak. Glob Health 18(1):1–14
35. Tadić B, Melnik R, (2021) Microscopic dynamics modeling unravels the role of asymptomatic virus carriers in sars-cov-2 epidemics at the interplay between biological and social factors. Comput Biol Med 133:104422. https://doi.org/10.1016/j.compbiomed.2021.104422
36. Thurner S, Klimek P, Hanel R (2020) A network-based explanation of why most covid-19 infection curves are linear. Proc Natl Acad Sci 117(37):22684–22689. https://doi.org/10.1073/pnas.2010398117
37. Tiwari S, Chanak P, Singh SK (2023) A review of the machine learning algorithms for covid-19 case analysis. IEEE Transactions on Artificial Intelligence 4(1):44–59. https://doi.org/10.1109/TAI.2022.3142241
38. Tsukuda S, Watashi K (2020) Hepatitis b virus biology and life cycle. Antiviral Res 182:104925. https://doi.org/10.1016/j.antiviral.2020.104925
39. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st International conference on Neural Information Processing Systems, Red Hook, NY, USA, pp 5998–6008
40. Wang Z, Cai B (2022) COVID-19 cases prediction in multiple areas via shapelet learning. Appl Intell 52(1):595–606. https://doi.org/10.1007/s10489-021-02391-6
41. Wang Z, Huang Y, He B, Luo T, Wang Y, Lin Y (2019) TDDF: HFMD outpatients prediction based on time series decomposition and heterogenous data fusion in Xiamen, China. In: Proceedings of the 15th International Conference Advanced Data Miningand Applications, Springer, Dalian, China, pp 658 – 667, https://doi.org/10.1007/978-3-030-35231-8_48
42. Wang Z, Huang Y, He B (2021) Dual-grained representation for hand, foot, and mouth disease prediction within public health cyber-physical systems. Software: Practice and Experience 51(11):2290 – 2305, https://doi.org/10.1002/spe.2940
43. Wang Z, Su Q, Chao G, Cai B, Huang Y, Fu Y (2022) A multi-view time series model for share turnover prediction. Appl Intell 52:14595–14606. https://doi.org/10.1007/s10489-021-02979-y
44. Wu Y, Yang Y, Nishiura H, Saitoh M (2018) Deep learning for epidemiological predictions. In: Proceedings of the 41st International Conference on Research and Development in Information Retrieval, ACM, Ann Arbor, MI, USA, pp 1085 – 1088, https://doi.org/10.1145/3209978.3210077
45. Xiao J, Zhu Q, Yang F, Zeng S, Zhu Z, Gong D, Li Y, Zhang L, Li B, Zeng W, Li X, Rong Z, Hu J, He G, Sun J, Lu J, Liu T, Ma W, Sun L (2022) The impact of enterovirus a71 vaccination program on hand, foot, and mouth disease in guangdong, china: A longitudinal surveillance study. J Infect 85(4):428–435. https://doi.org/10.1016/j.jinf.2022.06.020
46. Xu X, Ren W (2022) A hybrid model of stacked autoencoder and modified particle swarm optimization for multivariate chaotic time series forecasting. Appl Soft Comput 116:108321. https://doi.org/10.1016/j.asoc.2021.108321
47. Yang Z, Zhang Q, Cowling BJ, Lau EH (2017) Estimating the incubation period of hand, foot and mouth disease for children in different age groups. Sci Rep 7(1):16464
48. Zhang P, Wang Z, Chao G, Huang Y, Yan J (2022a) An oriented attention model for infectious disease cases prediction. In: Proceedings of the 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Springer, Kitakyushu, Japan, vol 13343, https://doi.org/10.1007/978-3-031-08530-7_11
49. Zhang P, Wang Z, Huang Y, Wang M (2022) Dual-grained directional representation for infectious disease case prediction. Knowl-Based Syst 256:109806. https://doi.org/10.1016/j.knosys.2022.109806
50. Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H, Zhang W (2021) Informer: Beyond efficient transformer for long sequence time-series forecasting. In: Proceedings of the 35th AAAI Conference on Artificial Intelligence, AAAI Press

Authors and Affiliations

Zhijin Wang[1] · Pesiong Zhang[2] · Yaohui Huang[3] · Guoqing Chao[4] · Xijiong Xie[5] · Yonggang Fu[1]

Pesiong Zhang
penci1007123@gmail.com

Yaohui Huang
yhhuang5212@gmail.com

Guoqing Chao
guoqingchao@hit.edu.cn

Xijiong Xie
xjxie11@gmail.com

Yonggang Fu
yonggangfu@jmu.edu.cn

[1] College of Computer Engineering, Jimei University, Yinjiang Road 185, Xiamen 361021, China

[2] School of Science, Jimei University, Yinjiang Road 185, Xiamen 361021, China

[3] College of Electronic Information, Guangxi Minzu University, Daxue East Road 188, Nanning 530006, China

[4] School of Computer Science and Technology, Harbin Institute of Technology, 2 West Culture Road, Weihai 264209, People's Republic of China

[5] School of Information Science and Engineering, Ningbo University, Fenghua Road 818, Ningbo 315211, China