Journal of Computer Science and Cybernetics, V.40, N.4 (2024), 363–380
DOI no. 10.15625/1813-9663/21185

SPATIO-TEMPORAL GRAPH LEARNING WITH EPIDEMIOLOGICAL FACTORS FOR HIV EPIDEMIC SHORT-TERM PREDICTION

DAT PHAM THANH[1,*], DUONG NGUYEN VAN[2], THANH TRAN TAN[3]
VIET ANH NGUYEN[4]

[1] *Graduate University of Science and Technology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet Street, Cau Giay District, Ha Noi, Viet Nam*
[2] *Department of HIV/AIDS and chronic infectious diseases prevention, Center for Disease Control, 366A Au Duong Lan Street, 8 District, Ho Chi Minh City, Viet Nam*
[3] *Information Technology Faculty, Industrial University of Ho Chi Minh City, 12 Nguyen Van Bao Street, Go Vap District, Ho Chi Minh City, Viet Nam*
[4] *Institute of Information Technology, Viet Nam Academy of Science and Technology, 18 Hoang Quoc Viet Street, Cau Giau District, Ha Noi, Viet Nam*

Abstract. HIV/AIDS is a major epidemic in the 21^{st} century, with high mortality rates and no effective preventive vaccine. It significantly impacts the economy, mental well-being and health systems and shortens national lifespans. Early detection helps reduce transmission and allocate medical resources effectively. However, predicting outbreaks remains challenging due to the influence of temporal, spatial and epidemiological factors, which complicate the spread of the disease across regions and pose difficulties for predictive models. Very few studies use deep learning models to tackle the HIV epidemic. To address this gap, we suggest using a graph data structure to simulate HIV transmission between neighboring areas and integrate epidemiological factors into this framework. We develop a spatio-temporal graph neural network model to predict short-term infection trends. This model incorporates important factors from HIV modeling, including temporal dynamics, geographic regions, and epidemiological variables such as age groups, career groups, gender groups, risk population groups, and transmission routes within an area. Our approach uses self-attention in the graph architecture to gather node-level information across the infection graph at each step during time series processing. We employ a GRU mechanism to update the graph information over time, allowing for a comprehensive evaluation of transmission probabilities between regions and improving predictive accuracy. Our proposed model was tested on HIV datasets from districts in Ho Chi Minh City, Viet Nam, and demonstrated superior performance compared to existing spatio-temporal models applied to the same dataset.

Keywords. Epidemic forecasting, HIV forecasting, Spatio-temporal graph learning.

*Corresponding author.
E-mail addresses: datpt.nsc@ioit.ac.vn (D.P.Thanh); nvduong@hcdc.gov.vn (D.N.Van); trantanthanh@iuh.edu.vn (T.T.Tan); anhnv@ioit.ac.vn (V.A.Nguyen).

1. INTRODUCTION

Disease prediction plays a crucial role in mitigating the economic, social, and health impacts of infectious diseases. It enables governments to monitor outbreaks and develop effective control strategies. Healthcare facilities rely on accurate predictions to prepare their resources and equipment. HIV/AIDS, the deadliest pandemic of the 20^{th} and 21^{st} centuries, has resulted in millions of deaths and infections [1]. Unlike influenza and COVID-19, which spread primarily through respiratory routes [2,3], HIV is transmitted through blood, sexual contact, and from mother to child [4]. It has a longer incubation period and spreads more slowly within communities. HIV-positive cases are identified by confirmatory tests in HIV testing centers [5]. By predicting HIV rates early, health agencies can allocate resources more effectively and work to reduce transmission.

Recent studies on disease prediction are based on algorithms for time series analysis. Time-series prediction involves collecting observations from the past and using them to develop a mathematical model that captures the process of generating information. The algorithms then utilize these models to predict future outcomes. In simpler terms, time-series forecasting involves predicting the future values of a time series. Prediction models can be categorized based on different factors, such as the mathematical models used for univariate or multivariate time series or the algorithms for linear or nonlinear models.

Traditional forecasting models like ARIMA and exponential smoothing use statistical models to forecast based on time series data. These models have several advantages in time series forecasting. They are flexible in capturing seasonal patterns, cycles, or trends, and they are simple and easy to implement because they require few parameters and basic statistical assumptions. However, traditional models may not perform well when the data contains additional external information needed for prediction or lack historical information, which are common issues in the real world. In recent years, advanced deep learning techniques, such as long short-term memory (LSTM) and convolutional neural networks (CNN), are used in time series forecasting, demonstrating notable advancements compared to traditional models. Deep learning models can forecast complex high-dimensional data sets and effectively address the challenge of missing information during data collection. Furthermore, deep learning facilitates feature learning and extracting pertinent features or representations for forecasting tasks based on raw data.

Despite advances in deep learning, models still struggle to fully capture the complex and dynamic nature of disease spread. This is particularly true when considering the spatial and temporal dependencies and the unique epidemiological factors associated with different epidemics. Disease models are usually dynamic, as they involve the spread of disease between neighboring regions over time. For example, an outbreak may occur in a region the following month if neighboring regions have high infection rates, potentially spreading the infection to adjacent regions. Inspired by [6], the main factors that influence disease prediction are listed below:

Location graph factors. In the context of disease transmission, the spread of diseases can occur between neighboring locations, which can be depicted as a graph. In this graph, each location is a node, and the connections between them represent potential disease transmission routes. The strength of these connections is indicated by the weights of the edges, which represent the likelihood of disease transmission between locations.

Epidemiological factors play a crucial role in increasing or decreasing the transmission potential of diseases within and between different locations. The number of positive cases at a specific location is one of the most common indicators of the infection potential in an area. However, to assess the transmission potential of a location relative to its surrounding areas, additional epidemiological factors are necessary, such as the risk population group, the transmission route group, the age group, the gender group, and other relevant factors.

Temporal factors. During disease outbreaks, the number of cases in each area can increase or decrease at specific points in time. It is important to track the trend of the disease's spread in individual areas and across all areas over time. Therefore, understanding the temporal factor is crucial to understanding the dynamic nature of disease outbreaks.

Therefore, the challenge in disease forecasting, including HIV transmission, lies in integrating spatial, epidemiological, and temporal factors into the prediction process. Some recent studies on forecasting COVID-19 outbreaks propose spatio-temporal prediction models [7–9]. However, most of these models have not fully accounted for the transmission dynamics between different regions over time and the incorporation of epidemiological factors.

In our paper, we present an HIV disease forecasting model that uses HIV epidemic data collected according to three key factors: spatial, epidemiological, and temporal. We have compared this model with various previous disease prediction models. The primary contributions of our paper are summarized below.

1. We propose to construct an HIV transmission graph between neighboring regions at each time point based on epidemiological information, including age groups, occupation groups, gender groups, transmission target groups, and transmission routes in each region.

2. We propose a disease prediction model for the HIV epidemic called 3FPREDICT (3-Factors Prediction) based on three spatial, epidemiological, and temporal factors. Specifically, we utilize a graph self-attention model to compute the transmission likelihood between regions based on the HIV transmission graph and employ GRU to process temporal sequence information.

3. We implemented the model using real-world data on HIV transmission in Ho Chi Minh City, Viet Nam. We used a list of HIV-positive cases detected in 23 districts and counties in Ho Chi Minh City, Viet Nam, from January 2009 to December 2019 to create monthly infection graphs. This resulted in a total of 144 infection graphs. We then built prediction models based on these graphs. We used a graph self-attention model combined with GRU for graph-based learning, which yielded competitive results compared to previous research on disease prediction models.

The remaining sections of the paper are structured as follows. Section 2 concisely describes previous research on disease prediction and its associated limitations. The proposed model is elaborated on in Section 3. The experimental results are presented in Section 4. Finally, Section 5 concludes the paper and suggests avenues for future research.

2. RELATED WORK

2.1. Traditional approaches

Traditional epidemic prediction models are based on basic statistical methods. The simplest forecasting algorithm, Naive, predicts each time step using the value observed in the preceding time step [10]. Q. Chen [11] apply the SIR model, a compartmental model focusing on the mathematical modeling of population-level dynamics, for forecasting and analyzing SARs. The authors in [12–15] use ARIMA (Autoregressive Integrated Moving Average) models to forecast COVID-19 outbreaks and daily blood sample collection visits. Although ARIMA is widely used to forecast univariate time series, its main drawback lies in its inability to accommodate seasonal components.

Classical statistical methods are generally fast, have lower computational costs, offer higher interpretability, and provide better statistical guarantees. They remain widely adopted and have shown superior performance compared to machine learning solutions in the past. However, traditional approaches often have limitations, such as robustness to missing data, reliance on hand-crafted features, heavy preprocessing requirements, and rigid design choices that require experience and a strong theoretical foundation. Despite the availability of many tools to automate the creation of these models, their need for a robust theoretical background makes them impractical for predicting multiple time series.

2.2. Deep learning approaches

Deep learning models [16–20] allow extracting relevant features for forecasting by learning from large and informative datasets. V.Chimula *et al.* [21] employ the LSTM network to forecast the end time of the COVID-19 pandemic in Canada. LSTM, an improved model of RNN (Recurrent Neural Network) [22], enables learning from distant dependencies. This LSTM model yields promising results with COVID-19 data collected from Johns Hopkins University and Canadian health authorities, including confirmed positive cases up to March 31, 2020, and daily counts of deaths and recoveries. Although the model performs well on the collected dataset, some unaccounted cases may affect its forecasting results, including incomplete reporting of data and failure to consider potential outbreaks that could influence the model's predictions. Similarly, the authors at [23–27] apply recurrent neural network algorithms such as RNN, GRU, and LSTM to predict COVID-19 and other diseases over time series. Diqi et al. [28] employ a Convolutional Neural Network (CNN) architecture commonly used in image processing to predict COVID-19. When applied to time series, CNN allows for better feature extraction and aggregation for prediction activities. Effective forecasting improvements include combining ARIMA and LSTM [29], LSTM with the Markov method [30], and LSTM with optimization methods [31]. Although the prediction results align with time series data, a key limitation of these models is their inability to account for the possibilities of interregional infection related to temporal variations in prediction models.

Epidemic models typically incorporate spatial (location of the disease outbreak) and temporal (timing of disease outbreak) information. Disease surveillance data are often collected based on spatial and temporal information. Although deep learning models can predict disease datasets, they may not account for the potential outbreak or control of diseases in specific locations over a certain period. The outbreak or control of a disease in one area may affect other areas and the same location in the future. Combining CNN-LSTM [32] and RNN-CNN [33] models aims to extract relevant feature information and consider these for distant dependencies. However, they do not address the spatial and temporal features of epidemic models. Basic deep learning models fail to capture relationships regarding the potential of disease outbreak in an area with temporal changes.

2.3. Spatio-temporal graph learning

To address spatio-temporal relationships in graph models, recent studies focus on utilizing spatio-temporal graph deep learning models to tackle epidemic prediction problems, primarily concerning COVID-19. Y.Zheng *et al.* [7] construct a spatio-temporal COVID-19 prediction graph, combining an SEIR (susceptible-exposed-infectious-recovered) model to compute node features and an RNN (Recurrent Neural Network) to compute edge features. The RNN model aims to capture the neighboring effect and regularize the landscape of the loss function, ensuring an effective and robust contribution of local minima to prediction. A weakness of the model is its failure to account for the time-dependent influence on disease outbreaks in various regions.

The authors in [34] apply the GCN-LSTM model to predict the transmission of COVID-19 in Connecticut. In the constructed graph, the nodes represent 169 towns or cities in Connecticut, and the edges represent the distance between towns or cities in Connecticut. LSTM is used to update the graph snapshots for each time point. Although edge information is integrated, GCN (Graph Convolutional Network) [35] has the potential to cause over-smoothing [36] on infection graphs. Panagopoulos et al. [8] employ a graph neural network model to predict COVID-19 in regions of France, Italy, Spain and England. The authors propose constructing a COVID-19 infection graph between regions at each time point in the time series, where nodes represent regions in the graph, and edges denote the number of individuals moving from one area to another. The model utilizes MPNN (Message Passing Neural Network), a graph neural network architecture, to aggregate graph information at each time point and employs LSTM to update graph information over time. Improvements suggested for the model in the paper include incorporating additional epidemiological information such as age/gender groups and other external factors to improve model precision. However, the model does not consider the influence of infection in all regions.

A.Kapoor *et al.* [37] constructed a spatiotemporal graph for COVID-19 forecasting in the counties of the United States. In this graph, nodes represent the graph vertices. They create graphs with different types of edges to represent the dependence between space and time. In the spatial domain, edges represent the movement of individuals between two areas. The simple edges represented binary connections in the temporal domain over the past few days. The graph consists of approximately 100 combined layers, each representing the daily spread of infection between regions. It uses an MLP (Multi-Layer Perception) for forecasting, with the input being the last layer. Similarly to [8], the weakness of the model lies in its disregard of epidemiological factors and the failure to calculate the coefficient of influence of the infection between regions.

Current spatio-temporal forecasting models have several notable disadvantages when predicting disease outbreaks. Firstly, many of these models lack an attention mechanism, which means that they cannot identify and focus on important nodes in the graph. This leads to insufficient information aggregation in areas with varying infection rates. Secondly, these models often do not fully integrate epidemiological factors such as age, gender, and high-risk groups, which limits their ability to make accurate predictions in complex disease scenarios. In general, these disadvantages can significantly impact the effectiveness of disease forecasting efforts.

In this paper, we explore the use of spatio-temporal graph neural networks [38] to predict the spread of disease in spatial and temporal spaces. In our analysis, we incorporate epidemiological parameters for HIV in our analysis. We use spatial and temporal HIV disease data and epidemiological information to create a model input graph. This integrates spatial, temporal, and epidemiological factors, with nodes representing different regions characterized by epidemiological statistics. This includes the total number of positive cases, age groups, gender groups, career groups, transmission route groups, and high-risk population groups of positive cases in the districts. The edges of the graph represent the connections between adjacent regions. The weights indicate the distance between these areas, calculated based on the proximity of HIV testing facilities in each region and the detection of positive cases. We assume that each region has a single testing facility, which functions as the only location to detect HIV-positive cases. In the context of a city, the regions where positive cases are predicted are the districts within the city. Each district has an HIV testing center where people from surrounding areas who are suspected to have HIV can come to be tested. Suspected positive cases can be tested in the district where they live or in neighboring districts. As a result, the likelihood of transmission between two districts or counties is also related to the distance between the testing centers.

To aggregate information within the graph, we employ the Graph Self-Attention architecture, enabling the synthesis of infection information from one region with all regions in the graph, not just adjacent ones. The Graph Self-Attention aggregates feature information of graph nodes and edges to compute the infection influence of all nodes, considering both spatial and epidemiological factors. To handle temporal graph sequences, we use the GRU network [39]. Finally, we employ a Multi-Layer Perceptron (MLP) to predict the number of positive cases for all regions for the next month.

3. PROPOSED METHOD

Our task is to predict short-term HIV-positive cases across neighboring regions (such as districts within a city). This involves information on n regions at time h in the past and predicting the number of HIV-positive cases at time $h + z$. The problem can be described as follows

$$\mathrm{Npcase}_n^{h+h_1} = \mathrm{PREDICT}(R_h^n, R_{h-1}^n, ..., R_1^n), \tag{1}$$

where R_t^n represent profiles of n adjacency regions at time step t, $\forall t \in [1, h]$.

In our study, we construct a graph with n nodes representing n regions and an edge between two nodes representing the distance between two HIV testing centers in two regions at the time step t $\forall t \in [1, h]$. We develop the 3FPREDICT model for graph learning in the prediction of HIV-positive cases. The overall prediction process can be shown as follows

$$\mathrm{Vpcase}_n^{h+z} = \mathrm{3FPREDICT}(G_h^n, G_{h-1}^n, ..., G_1^n), \tag{2}$$

where G_t^n represents the infection graph among n regions at time step t, $\forall t \in [1, h]$. Vpcase_n^{h+z} is a vector that contains the prediction of the n regions in time $h + z$ based on the 3FPREDICT prediction model.

3.1. Location infection graph construction

The graph is a set of nodes (also known as vertices) and edges connecting between each two nodes [40]. The graph provides a structure for representing entities and relationships between entities. The graph's nodes or edges can contain values representing the features of the nodes or edges. The graph can be represented by an adjacency matrix, where each element of the matrix represents each relationship between nodes in the graph. The strength of graph problems lies in their ability to propagate information across the graph. This propagation property is suitable for studying the spread of diseases through regions. Therefore, we construct an infection spread graph among adjacent regions. Through the properties of the graph, we investigate how transmission between regions impacts the forecasting of HIV-positive cases.

We construct graphs $G^t = (V^t, E^t)$ at time step t, where $V^t = (v_1^t, ..., v_n^t)$ is the set of nodes with $n = |V^t|$ representing the number of nodes in the graph. Each node on the graph represents a region we aim to predict. Each node has p features $X^t = (x_{p1}^t, ..., x_{pn}^t)$ corresponding to epidemiological factors of regions, including **number of positive cases at time t, age groups, career groups, gender groups, risk population groups, transmission route groups** and any other epidemiological information if available. The edge of the graph e_{ij}^t between two nodes v_i^t and v_j^t represents the ability of infection transmission based on the distance between two regions. Specifically, we define the adjacency matrix A^t of each graph G^t as follows: $a_{ij}^t = D_{ij}^t$ if two regions are adjacent a_{ij}^t= 0 if other, where D_{ij}^t is the distance between two HIV testing center in two regions i and j

$$a_{ij}^t = \begin{cases} D_{ij}^t & \text{if region i and region j border} \\ \quad\quad 0 & \text{other} \end{cases} \tag{3}$$

where edge weight $e_{ij}^t = D_{ij}^t$ is measured by the distance between the testing facilities at the centers of the nearby regions, where testing and identification of positive cases occur. We assume that this distance represents the likelihood of infection between two adjacent regions. This implies that infected individuals are less likely to move between regions with a greater distance than regions with closer testing facilities.

Figure 1 illustrates the transmission of HIV infection between districts in Ho Chi Minh City. The large red circles on the map indicate testing centers in each district and county. These centers detect positive cases within their respective district or county and neighboring districts of the city. The lines connecting the red circles represent the distances between testing centers, with shorter distances indicating higher transmission probabilities between the districts.

3.2. Proposed model

To address the problem, we propose using an overall prediction model as depicted in Figure 2. To compute the influence between regions at a given time, we employ the self-attention mentioned in [40] for each infection graph G_t at each time step t. This technique enables the self-attention mechanism to compute the influence of infection between regions. The computing influence is based on the combination function of epidemiological feature information of the nodes in the graph (representing regions) and edge information of the graph representing the distance between adjacent regions. Subsequently, the graphs are updated using the Gated Recurrent Unit(GRU) for time-series processing.

Graph Neural Network(GNN). Graph Neural Networks [41] is a set of deep learning techniques designed to process and analyze graph-structured data, particularly in node embedding tasks. This allows information aggregation from a node's features based on

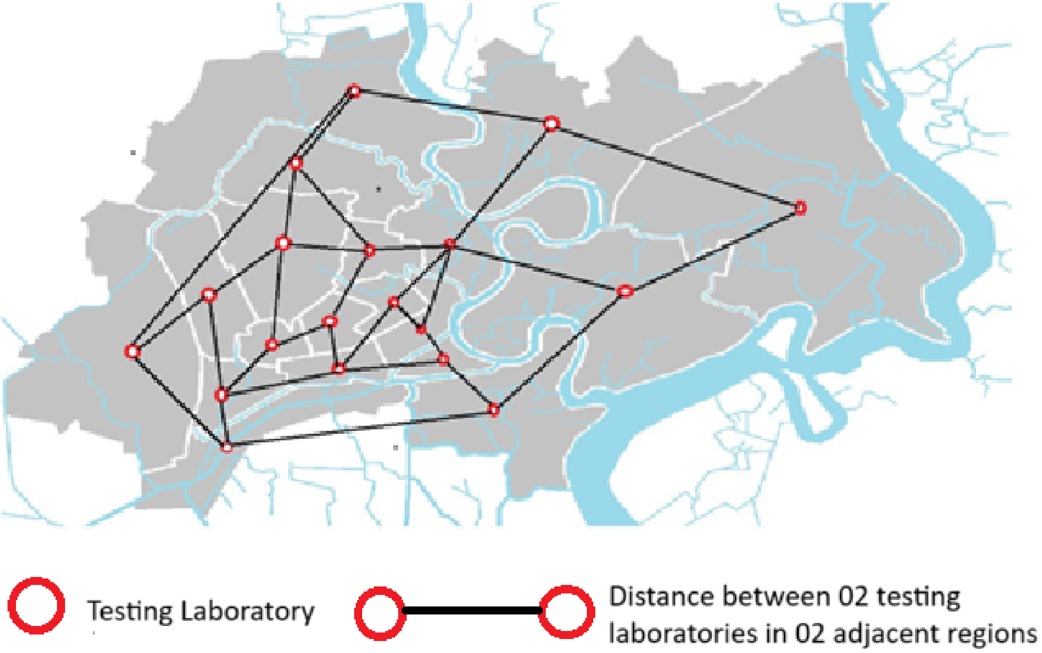
Testing Laboratory
Distance between 02 testing laboratories in 02 adjacent regions

Figure 1: HIV Infection map between areas in Ho Chi Minh City, Viet Nam

neighboring nodes' features. Two fundamental operations in GNN are message passing, which involves propagating information across the graph, aggregating neighborhood node features, and updating new node embedding. These processes enable a node in the graph to iteratively integrate information from its neighboring nodes. The propagation process [41] is formulated as follows

$$h_i^{'} = \gamma(h_i, \oplus_{j\in N_i}\phi(h_i, h_j, e_{ij})), \tag{4}$$

where h_i and h_j is node embedding of node vi, vj, e_{ij} is edge embedding of the edge between node v_i and node v_j and N_i is a set of neighboring nodes of node v_i, γ is update function, $\oplus$ is an aggregation function, and ϕ is an update function. Some typical GNN architectures are GCN (Graph Convolution Network) [35] and GAT (Graph Attention Network) [42].

Graph Self-attention(GSA) on Region infection graphs. Inspired by Vaswani et al. [43], we calculate the impact between the infected regions in the location infection graphs G^t at times step t by self-attention and multi-head attention. To calculate the influence coefficient between two nodes v_i^t and v_j^t, taking into account the weight of the edge e_{ij}^t we use the following formula

$$\alpha^t_{c,ij} = \frac{q^t_{c,i}, k^t_{c,i} + e^t_{c,ij}}{\sum_{u\epsilon N(i)} q^t_{c,i}, k^t_{c,u} + e^t_{c,iu}}, \tag{5}$$

where $q^t_{c,i} = W^t_{c,q}h^t_i + b^t_{c,q}$, $k^t_{c,j} = W^t_{c,k}h^t_j + b^t_{c,k}$, $e^t_{c,ij} = W^t_{c,e}e^t_{ij} + b^t_{c,e}$, and

$$(q^t, k^t) = \exp\left(\frac{q^t k^{t^T}}{\sqrt{d}}\right). \tag{6}$$

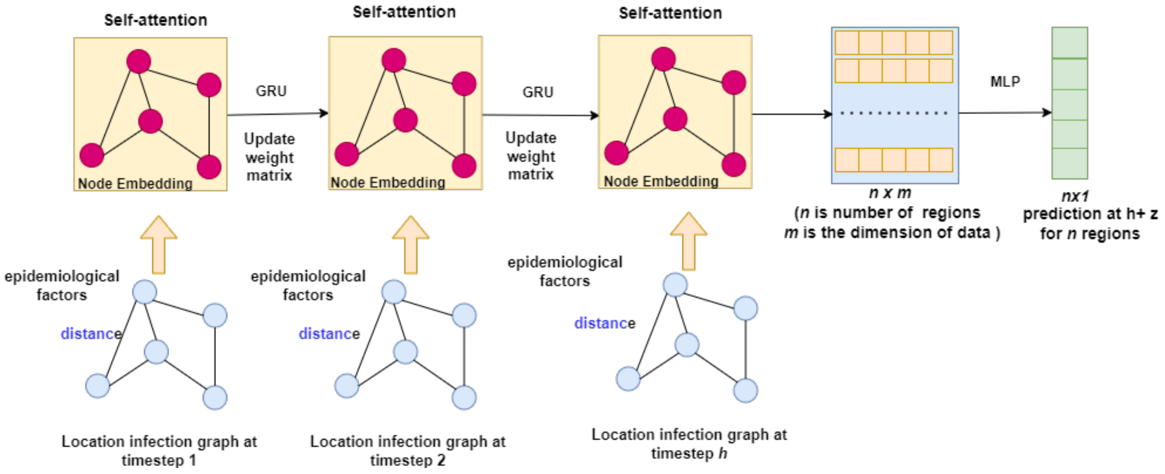

Figure 2: The overall prediction model **3FPREDICT**

Equation(6) represents the exponential scale dot product function, and d is the hidden dim of each head. With attention head c^{th}, we transform the features h_i^t of node v_i^t to the query vector $q_{c,i}^t \in \mathbb{R}^d$ and the feature h_j^t of node v_j^t to the key vector $k_{c,j}^t \in \mathbb{R}^d$ with the trainable parameter $W_{c,q}^t$, $b_{c,q}^t$, $W_{c,k}^t$, $b_{c,k}^t$. The edge feature value e_{ij}^t is aggregated with the key value as additional information.

After calculating the infection coefficient, we aggregate the information from node v_i^t to node v_j^t using the following formula

$$h_i^{'} = \|_{c=1}^{C} \left[\sum_{j \epsilon N(i)} \alpha_{c,ij}^t \left(v_{c,j}^t + e_{c,iu}^t \right) \right] \tag{7}$$

with

$$v_{c,j}^t = W_{c,v}^t h_j^t + b_{c,v}^t, \tag{8}$$

where C is the number of heads and h_j^t transforms to vector $v_{c,j}^t \in \mathbb{R}^d$ with trainable parameter $W_{c,v}^t$ and $b_{c,v}^t$.

Temporal representation. 3FPREDICT model employs Gated Recurrent Neural Networks (GRU) to handle G^t, $\forall t \in [1, h]$ in time series processing, preserving essential information while selectively discarding less important details. GRU consists of two gates: the reset gate and the update gate. The update gate helps the model determine how much information should be transferred. In contrast, the reset gate primarily decides how much past information should be discarded. GRU is formulated as follows

$$z^t = \sigma(W_z^t x^t + U_z^t h^{t-1}), \tag{9}$$

$$r^t = \sigma(W_r^t x^t + U_r^t h^{t-1}), \tag{10}$$

$$\tilde{h}^t = \text{Tanh}(W_h x^t + U_h(r^t \odot h^{t-1}), \tag{11}$$

$$h^t = (1 - z^t) \odot \tilde{h}^t + z^t \odot h^{t-1}, \tag{12}$$

where z^t is the update gate and r^t is the output gate, x^t is the input vector at time step t and h^{t-1} is the vector of the time step $t-1$. W_z^t,U_z^t,W_r^t,U_r^t and W_h^t,U_h^t is trainable weight matrices. $\tilde{h}^t$ is store past information using the reset gate, and h^t is the final result. To shorten, we can denote GRU as follows

$$h^t = \mathrm{GRU}(x^t, h^{t-1}). \tag{13}$$

Spatio-temporal representation. In spatio-temporal representation. GRU updates the Graph Self-attention weight matrix for layer l at time step t as follows

$$W^t = \mathrm{GRU}(h^t, W^{t-1}), \tag{14}$$

where h^t denotes the node embedding updated at time step t and W^{t-1} is the weight matrix in time step $t-1$. The result weight matrix is then used to calculate the embedding of the next layer node as follows

$$H^{t+1} = \mathrm{GSA}(A^t, H^t, W^t), \tag{15}$$

where A^t is the adjacency matrix of G^t.

Output layer. We utilize a Multi-layer Perceptron (MLP) to predict the number of HIV infections in the upcoming months. The input to the MLP is represented by the matrix $H \in \mathbb{R}^{n\times m}$, which contains information from n nodes in the infection graph at the prediction time, with m dimensions conveying the information. The output prediction is a vector $o \in \mathbb{R}^{n\times 1}$, which contains the predicted information for n nodes corresponding to n forecast regions.

4. EXPRIMENTAL RESULTS

We test the proposed model using a real HIV epidemic dataset and compare it with existing deep learning and spatio-temporal models previously examined in the same dataset.

4.1. Dataset and evaluation metrics

Dataset. We use the HIV case surveillance data set from Ho Chi Minh City, Viet Nam. The dataset contains aggregated data on HIV infection status in 23 districts from January 2009 to December 2019 (144 months). Each month, data are collected in each district and include the following features:

1. Number of positive cases in a month.

2. Age groups for positive cases (¡5, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34,35-39,40-44,45-49, ¿ 50).

3. Career groups of positive cases (service business staff susceptible to exploitation, such as drivers, fishermen, agricultural workers, soldiers, public employees, students, children, freelance workers, unemployed, and prisoners).

4. Gender groups of positive cases (male, female).

5. Risk population groups of positive cases (injecting drug users, female sex workers, pregnant women, blood donors, tuberculosis patients, people with sexually transmitted infections, youth undergoing military service examination, men who have sex with men, heterosexual individuals, prisoners, spouses/partners of injecting drug users, spouses/partners

of people living with HIV).

6. Transmission route groups of positive cases (via injection, homosexual transmission, heterosexual sexual transmission, from mother to child).

Graph Construction. We create monthly infection graphs for each district. In these graphs, districts are represented as nodes, and the node characteristics include the number of infections, age groups, occupation groups, gender groups, risk population groups, and transmission route groups of positive cases for each month. Unlike the COVID-19 pandemic, which can surge in a few days, HIV epidemic features involve a slower detection of infections compared to COVID-19. Thus, the infection graphs are constructed monthly instead of daily to accurately reflect the features of the HIV epidemic and forecast the number of infections for the next month in each district. If two districts are adjacent, there is an edge between their corresponding nodes. The edge weights are calculated based on the distance between the HIV testing centers of the two districts.

Baselines. In our study, we are comparing a new model with traditional time series forecasting methods such as LSTM [21,25], CNN [28], and previous spatio-temporal graph forecasting models. As there is very limited research on the use of spatio-temporal graph models for HIV epidemics, we decided to implement previously used COVID-19 epidemic forecasting models, including MPNN-LSTM [8] and GCN-LSTM [34], on the same HIV epidemic dataset.

Evaluation metrics. We apply the mean square error (MSE) and the coefficient of determination (R^2) to measure the performance of all models.

1. MSE is used to evaluate the similarity between actual and predicted values. A lower MSE indicates a better model. MSE is formulated as MSE $= \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2$ where y_i is the ground truth and $\hat{y_i}$ is the predicted value.

2. R^2 indicates the extent to which the model's independent variables explain the variance of the dependent variable. The R^2 value ranges from 0 to 1, with a value close to 0 indicating that the model explains the variance of the dependent variable poorly. A model with a higher R^2 value is generally considered better. R^2 is formulated as $R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}-y_i)^2}{\sum_{i=1}^{n}(\bar{y}-y_i)^2}$ where y_i is the ground truth and $\bar{y}$ is the mean value of grouth truth y_i.

4.2. Results

The 3FPREDICT model predicts the number of HIV-positive cases in 23 districts in Ho Chi Minh City. It achieved an MSE validation score of 29.25 and an R^2 validation score of 0.43; detailed results are presented in Table 1. The MSE indicator of 29.25 for the proposed 3FPREDICT model is lower than the MPNN-LSTM (35.11), GCN-LSTM (38.10), LSTM (48.51), and CNN (46.27) models. This suggests that the 3FPREDICT model is the best choice for forecasting HIV cases in Ho Chi Minh City among the models compared. The MSE indicates that the error between the predicted and actual values of 3FPREDICT is the smallest.

Table 1: Comparison of MSE and R^2 on Ho Chi Minh City HIV Dataset

	3FPRE	MPNNLSTM	GCNLSTM	LSTM	CNN
MSE	29.25	35.11	38.10	48.51	46.27
R^2	0.43	0.24	0.18	-0.04	0.003

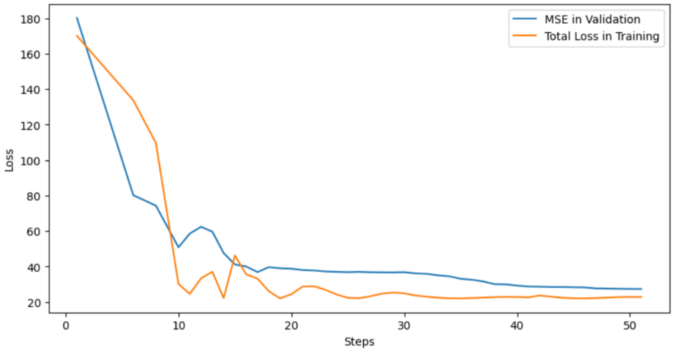

Figure 3: GSA-GRU MSE validation scores

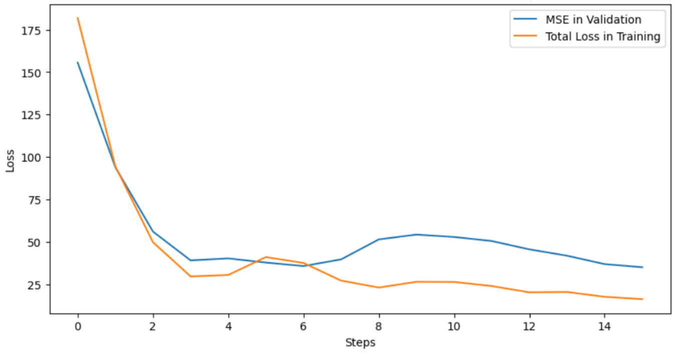

Figure 4: MPNN-LSTM MSE validation scores

The R^2 indicator of 0.43 for 3FPREDICT is the highest compared to the MPNN-LSTM (0.24), GCN-LSTM (0.17), LSTM (-0.04), and CNN (0.003) models. This indicates that based on the city's dataset, the 3FPREDICT model is the most suitable for making short-term predictions of HIV-positive cases in Ho Chi Minh City.

We have two charts to assess and compare the performance of the 3DPREDICT model and the MPNN-LSTM model based on their loss values during training and testing. Figure 3 displays the loss values of the 3DPREDICT model, while Figure 4 shows the loss values of the MPNN-LSTM model. The mean squared error (MSE) values and the total loss values in the 3FPREDICT model are observed to decrease during the training process, indicating that the model is learning and improving its prediction accuracy with each step. After about 20 steps, the MSE and total loss values stabilize and fluctuate between 20 to 30. This stability contrasts with the MPNN-LSTM model, which fluctuates between 20 and 50. The stabilization at this point indicates that the model has learned the main features of the data and has reached a certain level of convergence. The smaller difference between the MSE values and the training loss values, compared to the MPNN-LSTM model, suggests that the model is not overfitting as much.

Figures 5 and 6 compare the R^2 values of the 3FPREDICT model and the MPNN-LSTM model during validation. Figure 5 presents the R^2 values of the Graph Self-attention GRU model, while Figure 6 demonstrates the R^2 values of the MPNN-LSTM model. This com-

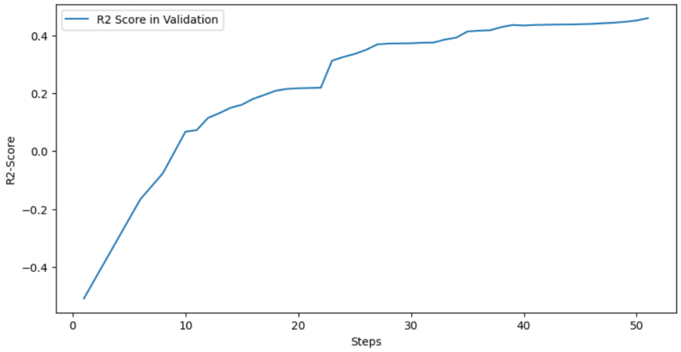

Figure 5: GSA-GRU R^2 validation scores

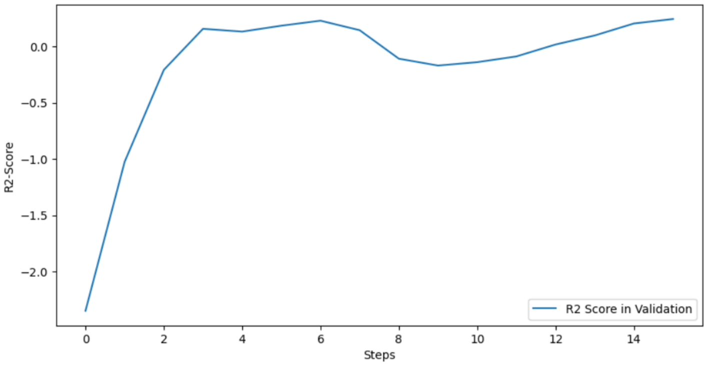

Figure 6: MPNN-LSTM R^2 validation scores

parison allows us to assess the suitability of each model for the actual data. The Graph Self-attention GRU model exhibits a steady increase in R^2 values, reaching a higher level. In contrast, the MPNN-LSTM model displays more fluctuations and does not achieve as high R^2 values as the Graph Self-attention GRU model. This suggests that the graph self-attention - GRU model possesses better predictive and explanatory power for the data.

5. CONCLUSION

In this study, we propose an HIV disease prediction model called 3FPREDICT, where 3F denotes 03 spatial, temporal, and HIV epidemiological factors. The model combines graph self-attention for processing infection graphs by region at each time point and a GRU for handling the time series data. It is used for short-term prediction in Ho Chi Minh City, Vietnam.

In 3FPREDICT, at each time step in the time series, a location infection graph is created, connecting neighboring districts. In this graph, the nodes represent the districts and their features include various epidemiological factors such as age groups, career groups, gender groups, risk population groups, and transmission routes groups. The edges of the graph connect neighboring regions. They are weighted according to the distance between HIV testing centers in each district, where tests are performed and positive case detection occurs. To aggregate node information at each time step, we utilize Graph Self-attention, allowing the synthesis of information across the entire graph for each node representing a district in the city using a self-attention mechanism. For temporal processing, we used a GRU. The inputs are the weight matrices of the graph after passing through the self-attention layer, and the outputs are the weight matrices of the subsequent GNN layer. We analyzed real HIV infection data in 23 districts and communes in Ho Chi Minh City from January 2009 to December 2019, which produced more accurate results than previously utilized spatio-temporal disease prediction models.

The use of spatio-temporal graph models for HIV forecasting represents a new approach in this field. This research provides useful information on the potential of advanced machine learning techniques in predicting epidemics. It emphasizes integrating spatial, temporal, and epidemiological factors in the data to improve prediction accuracy.

‘ Our study has some limitations. The model was trained and tested using a specific HIV epidemic dataset from Ho Chi Minh City, Viet Nam. This may restrict the generalizability of the results for HIV epidemic forecasting in other regions and limit the model's application to other infectious diseases. Improving the model's performance in the context of the HIV epidemic could benefit from integrating more diverse datasets within HIV. This could involve incorporating treatment data, such as treatment participation rates or the rates of treated cases with viral load test results below the threshold. Although we have enhanced graph learning with Graph Self-attention, our model has not improved in handling time-series data, as we relied on the basic GRU algorithm.

The next steps of our research should focus on overcoming these limitations. This can be achieved by testing our models using different HIV datasets and diseases and refining them to improve their reliability and accuracy. It would also be beneficial to investigate further how additional treatment data, such as the rates of treated cases with viral load test results below a certain threshold, can offer more comprehensive insights into the dynamics of the HIV epidemic. Additionally, we should consider expanding the GRU architecture with an attention-GRU, as this could enhance accuracy and provide insights into the variation coefficient across all regions at different times compared to other data points in the prediction sequence.

In conclusion, our study shows that combining spatio-temporal graph models with epidemiological factors can improve the accuracy and effectiveness of HIV forecasting, leading to more effective public health responses. We believe this research can pave the way for future studies and applications of this model in combatting HIV and other infectious diseases.

REFERENCES

[1] UNAIDS, “Global HIV & AIDS statistics,” 2023, accessed: 01-Jun-2024. [Online]. Available: https://www.unaids.org/en

[2] USCDC, “COVID-19,” 2024, accessed: Jun. 30, 2024. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/index.html

[3] ——, “Influenza (flu): About flu,” 2024, accessed: Jun. 30, 2024. [Online]. Available: https://www.cdc.gov/flu/about/index.html

[4] WHO, “HIV/AIDS,” 2024, accessed: 30-Jun-2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/hiv-aids

[5] USCDC, “HIV: Causes of HIV,” 2024, accessed: Jun. 30, 2024. [Online]. Available: https://www.cdc.gov/hiv/causes/index.html

[6] S. Zhong and L. Bian, “What drives disease flows between locations?” *Transactions in GIS*, vol. 24, no. 6, pp. 1740–1755, Dec 2020. [Online]. Available: https://doi.org/10.1111/tgis.12675

[7] Y. Zheng, Z. Li, J. Xin, and G. Zhou, “A spatial-temporal graph based hybrid infectious disease model with application to COVID-19,” *Proceedings of the 11th International Conference on Bioinformatics and Biomedical Technology (ICBBT '21)*, pp. 357–364, 2021. [Online]. Available: https://doi.org/10.5220/0010349003570364

[8] G. Panagopoulos, G. Nikolentzos, and M. Vazirgiannis, “Transfer graph neural networks for pandemic forecasting,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 4838–4845, 2021. [Online]. Available: https://doi.org/10.1609/aaai.v35i6.16616

[9] S. Berkani, B. Guermah, M. Zakroum, and M. Ghogho, “Spatio-temporal forecasting: A survey of data-driven models using exogenous data,” *IEEE Access*, vol. 11, pp. 75 191–75 214, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3282545

[10] M. Langarizadeh and F. Moghbeli, “Applying Naive Bayesian networks to disease prediction: a systematic review,” *Acta Inform Med*, vol. 24, no. 5, pp. 364–369, Oct 2016. [Online]. Available: https://doi.org/10.5455/aim.2016.24.364-369

[11] Q. Chen, “Application of SIR model in forecasting and analyzing for SARS,” *Beijing da xue xue bao. Yi xue ban = Journal of Peking University. Health sciences*, vol. 35 Suppl, pp. 75–80, 2003.

[12] G. Wang, W. Wei, J. Jiang, C. Ning, H. Chen, J. Huang, B. Liang, N. Zang, Y. Liao, R. Chen, J. Lai, O. Zhou, J. Han, H. Liang, and L. Ye, “Appburgeoning method of deep learning in forecasting HIV incidence in Guangxi, China,” *Epidemiol Infect*, vol. 147, p. e194, 2019. [Online]. Available: https://doi.org/10.1017/S095026881900075X

[13] S. Claris and N. Peter, “ARIMA model in predicting of COVID-19 epidemic for the Southern Africa region,” *African Journal of Infectious Diseases*, vol. 17, no. 1, pp. 1–9, Dec 2022. [Online]. Available: https://doi.org/10.21010/Ajidv17i1.1

[14] N. Duong, L. Thao, D. Quynh, L. Binh, C. Loan, and P. Diem, “Predicting the pandemic COVID-19 using ARIMA model,” *VNU Journal of Science: Mathematics - Physics*, vol. 36, 2020. [Online]. Available: https://doi.org/10.25073/2588-1124/vnumap.4492

[15] X. Zhang, Y. Yu, F. Xiong, and L. Luo, “Prediction of daily blood sampling room visits based on ARIMA and SES model,” *Computational and Mathematical Methods in Medicine*, vol. 2020, p. 1720134, 2020. [Online]. Available: https://doi.org/10.1155/2020/1720134

[16] A. Fokas, N. Dikaios, and G. Kastis, "Mathematical models and deep learning for predicting the number of individuals reported to be infected with SARS-CoV-2," *Journal of the Royal Society Interface*, vol. 17, no. 169, p. 20200494, Aug 2020. [Online]. Available: https://doi.org/10.1098/rsif.2020.0494

[17] S. Shastri, K. Singh, S. Kumar, P. Kour, and V. Mansotra, "Time series forecasting of COVID-19 using deep learning models: India-USA comparative case study," *Chaos Solitons Fractals*, vol. 140, p. 110227, Nov 2020. [Online]. Available: https://doi.org/10.1016/j.chaos.2020.110227

[18] P. Arora, H. Kumar, and B. K. Panigrahi, "Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India," *Chaos, Solitons Fractals*, vol. 139, p. 110017, 2020. [Online]. Available: https://doi.org/10.1016/j.chaos.2020.110017

[19] L. Wang, J. Chen, and M. Marathe, "DEFSI: Deep learning based epidemic forecasting with synthetic information," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9607–9612. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33019607

[20] Y. He, Y. Zhao, Y. Chen, H.-Y. Yuan, and K.-L. Tsui, "Nowcasting influenza-like illness (ILI) via a deep learning approach using Google search data: An empirical study on Taiwan ILI," *International Journal of Intelligent Systems*, vol. 37, 2021. [Online]. Available: https://doi.org/10.1002/int.22788

[21] V. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos Solitons Fractals*, vol. 135, p. 109864, Jun 2020. [Online]. Available: https://doi.org/10.1016/j.chaos.2020.109864

[22] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020. [Online]. Available: https://doi.org/10.1016/j.physd.2019.132306

[23] K. ArunKumar, D. Kalaga, C. Kumar, M. Kawaji, and T. Brenza, "Forecasting of COVID-19 using deep layer recurrent Neural Networks (RNNs) with gated recurrent units (GRUs) and long short-term memory (LSTM) cells," *Chaos Solitons Fractals*, vol. 146, p. 110861, May 2021. [Online]. Available: https://doi.org/10.1016/j.chaos.2021.110861

[24] Shahid, Farah and Zameer, Aneela and Muneeb, Muhammad, "Predictions for covid-19 with deep learning models of LSTM, GRU and Bi-LSTM," *Chaos, Solitons Fractals*, vol. 140, p. 110212, 2020. [Online]. Available: https://doi.org/10.1016/j.chaos.2020.110212

[25] R. Pathan, M. Biswas, and M. Khandaker, "Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based LSTM model," *Chaos Solitons Fractals*, vol. 138, p. 110018, Sep 2020. [Online]. Available: https://doi.org/10.1016/j.chaos.2020.110018

[26] N. T. Pham, C. T. Nguyen, and M. R. B. Pineda-Cortel, "Time-series modeling of dengue incidence in the Mekong delta region of Viet Nam using remote sensing data," *Western Pacification of a Long Short-Term Memory Neural Network: A Surveill Response J*, vol. 11, no. 1, pp. 13–21, 2020. [Online]. Available: https://doi.org/10.1038/s41598-018-34552-6

[27] S. Aribe Jr, B. Gerardo, and R. Medina, "Time Series Forecasting of HIV/AIDS in the Philippines Using Deep Learning: Does COVID-19 Epidemic Matter?" *International Journal of Emerging Technology and Advanced Engineering*, vol. 12, pp. 144–157, 2022. [Online]. Available: https://www.ijetae.com/files/IJETAE-12-1/IJETAE-12-1-144.pdf

[28] M. Diqi, S. Mulyani, and R. Pradila, "DeepCov: Effective prediction model of COVID-19 using CNN algorithm," *SN Computer Science*, vol. 4, no. 4, p. 396, May 2023. [Online]. Available: https://doi.org/10.1007/s42979-023-01834-w

[29] X. Guo, Y. Yang, S. P. Tseng, Z. Yin, and C. Wang, "Research on HIV/AIDS Epidemic Trend Prediction Model Based on ARIMA-LSTM," in *2022 10th International Conference on Orange Technology (ICOT)*, 2022, pp. 1–4. [Online]. Available: https://doi.org/10.1109/ICOT56925.2022.10008185

[30] R. Ma, X. Zheng, and P. Wang, "The prediction and analysis of COVID-19 epidemic trend by combining LSTM and Markov method," *Scientific Reports*, vol. 11, p. 17421, 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-97037-5

[31] L. Zhou, C. Zhao, N. Liu, X. Yao, and Z. Cheng, "Improved LSTM-based deep learning model for COVID-19 prediction using optimized approach," *Engineering Applications of Artificial Intelligence*, vol. 122, p. 106157, 2023. [Online]. Available: https://doi.org/10.1016/j.engappai.2023.106157

[32] Z. M. Zain and N. Alturki, "Covid-19 pandemic forecasting using CNN-LSTM: a hybrid approach," *Journal of Control Science and Engineering*, vol. 2021, pp. 8 785 636:1–8 785 636:23, 2021. [Online]. Available: https://doi.org/10.1155/2021/8785636

[33] M. Fakhfakh, B. Bouaziz, F. Gargouri, and L. Chaari, "ProgNet: COVID-19 prognosis using recurrent and convolutional Neural Networks," *The Open Medical Imaging Journal*, vol. 12, pp. 11–12, 2020. [Online]. Available: https://doi.org/10.2174/1874347102012010011

[34] N. Choi, "A deep learning model for predicting covid-19 transmission in connecticut," *Journal of Physics: Conference Series*, vol. 1972, no. 1, p. 012108, 2021. [Online]. Available: https://doi.org/10.1088/1742-6596/1972/1/012108

[35] S. Zhang, H. Tong, and J. Xu, "Graph convolutional networks: a comprehensive review," *Computational Social Networks*, vol. 6, p. 11, 2019. [Online]. Available: https://doi.org/10.1186/s40649-019-0069-y

[36] T. Rusch, M. Bronstein, and S. Mishra, "A survey on oversmoothing in graph neural networks," *arXiv preprint arXiv:2301.00156*, 2023. [Online]. Available: https://arxiv.org/abs/2301.00156

[37] A. Kapoor, X. Ben, L. Liu, B. Perozzi, M. Barnes, M. Blais, and S. O'Banion, "Examining COVID-19 forecasting using spatio-temporal graph neural networks," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020. [Online]. Available: https://doi.org/10.1145/3394486.3403222

[38] Z. Al Sahili and M. Awad, "Spatio-temporal graph neural networks:a survey," *arXiv preprint arXiv:2301.10569*, 2023. [Online]. Available: https://arxiv.org/abs/2301.10569

[39] Z. Li, X. Li, and Y. Huang, "Research on accurate location algorithm of optimized multi-source data fusion based on improved GRU network," in *Journal of Physics: Conference Series*, vol. 2113, 2021, p. 012073. [Online]. Available: https://doi.org/10.1088/1742-6596/2113/1/012073

[40] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: a review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020. [Online]. Available: https://doi.org/10.1016/j.aiopen.2021.01.001

[41] B. Khemani, S. Patil, K. Kotecha, and S. Tanwar, "A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions," *Journal of Big Data*, vol. 11, p. 18, 2024. [Online]. Available: https://doi.org/10.1186/s40537-023-00876-4

[42] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1710.10903

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [Online]. Available: https://doi.org/10.5555/3295222.3295349

Received on July 20, 2024
Accepted on October 18, 2024