mathematics

MDPI

Article

Advanced Trans-BiGRU-QA Fusion Model for Atmospheric Mercury Prediction

Dong-Her Shih [1], Feng-I. Chung [2], Ting-Wei Wu [1,*], Bo-Hao Wang [1] and Ming-Hung Shih [3]

[1] Department of Information Management, National Yunlin University of Science and Technology, Douliu 64002, Taiwan; shihdh@yuntech.edu.tw (D.-H.S.); bohau888@gmail.com (B.-H.W.)
[2] Center for General Education, National Chung Cheng University, Chiayi 621301, Taiwan; albertchung@lib.ccu.edu.tw
[3] Department of Electrical and Computer Engineering, Iowa State University, 2520 Osborn Drive, Ames, IA 50011, USA; mshih@iastate.edu
* Correspondence: wutingw@yuntech.edu.tw

Abstract: With the deepening of the Industrial Revolution and the rapid development of the chemical industry, the large-scale emissions of corrosive dust and gases from numerous factories have become a significant source of air pollution. Mercury in the atmosphere, identified by the United Nations Environment Programme (UNEP) as one of the globally concerning air pollutants, has been proven to pose a threat to the human environment with potential carcinogenic risks. Therefore, accurately predicting atmospheric mercury concentration is of critical importance. This study proposes a novel advanced model—the Trans-BiGRU-QA hybrid—designed to predict the atmospheric mercury concentration accurately. Methodology includes feature engineering techniques to extract relevant features and applies a sliding window technique for time series data preprocessing. Furthermore, the proposed Trans-BiGRU-QA model is compared to other deep learning models, such as GRU, LSTM, RNN, Transformer, BiGRU, and Trans-BiGRU. This study utilizes air quality data from Vietnam to train and test the models, evaluating their performance in predicting atmospheric mercury concentration. The results show that the Trans-BiGRU-QA model performed exceptionally well in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2), demonstrating high accuracy and robustness. Compared to other deep learning models, the Trans-BiGRU-QA model exhibited significant advantages, indicating its broad potential for application in environmental pollution prediction.

Keywords: atmospheric mercury; air pollution; transformer; bidirectional gated recurrent unit (BiGRU); quick attention (QA)

MSC: 68T05; 68T35; 68Q32

Citation: Shih, D.-H.; Chung, F.-I.; Wu, T.-W.; Wang, B.-H.; Shih, M.-H. Advanced Trans-BiGRU-QA Fusion Model for Atmospheric Mercury Prediction. *Mathematics* **2024**, *12*, 3547. https://doi.org/10.3390/math12223547

Academic Editors: Jing Zhang and Chenyang Bu

Received: 25 October 2024
Revised: 10 November 2024
Accepted: 12 November 2024
Published: 13 November 2024

1. Introduction

Since the 20th century, the development of the Industrial Revolution has driven the rapid rise in the chemical industry, with its emissions of corrosive gases and dust becoming a significant source of air pollution. As industrialization and urbanization accelerate, air pollution has become increasingly severe in developing countries and cities [1], and the long-term emission of pollutants has had profound negative impacts on the human living environment [2]. Among these pollutants, the element mercury (Hg) has been listed by the United Nations Environment Programme (UNEP) as one pollutant that requires severe global attention. Yuan et al. [3] further pointed out that mercury is one of the most toxic metal elements. Its derivatives pose significant health risks to humans.

Mercury (Hg) exhibits potential carcinogenicity under certain conditions, with particularly pronounced neurotoxic effects on the human nervous system [4]. The long-range atmospheric dispersion of mercury emissions exacerbates the global pollution problem [5].

Therefore, predicting atmospheric mercury concentrations is significant for identifying risks in advance and reducing environmental impacts. Studies have analyzed the spatiotemporal variations, gaseous distributions, and potential sources of atmospheric mercury around the Taiwan Strait, revealing that mercury in the atmosphere, carried by airflows, may affect the air quality in Taiwan [3]. Wang et al. [6] conducted a two-year study on the spatiotemporal variations and long-distance transport of atmospheric mercury in northern coastal China, finding that the concentration of gaseous elemental mercury (GEM) exhibited significant seasonal variation, peaking in winter. In Ho Chi Minh City, Vietnam, the total gaseous mercury (TGM) concentration was significantly influenced by the monsoon, and local mercury pollution was closely related to urban areas [7]. Pang et al. [8] pointed out that small-scale gold mining significantly elevated atmospheric mercury concentrations in the surrounding environment and mining areas, causing severe health damage to residents and miners.

With the accelerated process of industrialization, air pollution problems in numerous cities have become increasingly severe. Previous studies have explored the correlation between mercury (Hg) and other air pollutants by integrating different air pollution components, revealing a close relationship between gaseous mercury (Hg) and carbon monoxide (CO), as well as positive correlations with $PM_{2.5}$, PM_{10}, and nitrogen dioxide (NO_2) [9]. Additionally, other research has successfully employed Long Short-Term Memory (LSTM) neural networks from deep learning to predict air quality in the Beijing–Tianjin–Hebei region of China, achieving significant results [10]. Wen et al. [11] effectively predicted dioxin concentrations by combining deep-learning models with time series data, noting that airflow, temperature, and temporal factors significantly impacted the concentrations. Samad et al. [2] accurately predicted $PM_{2.5}$, PM_{10}, and NO_2 concentrations using variables collected from monitoring stations combined with deep learning models. Yuan et al. [3] pointed out that mercury threatens human life. Since the 20th century, the public has gradually become aware of the significant impacts of mercury on the ecological environment and human health, prompting the adoption of preventive measures to mitigate mercury poisoning and emissions.

Regarding the uniqueness of atmospheric mercury data, the existing literature predominantly focuses on analyzing the environmental impact of various mercury forms and their distribution in the atmosphere. However, research on the interdependence between atmospheric mercury and other air pollution components is relatively limited, particularly in univariate and multivariate fusion. Studies that explore the predictive performance of various deep learning models in this regard are even scarcer. Additionally, due to the insufficient number and uneven distribution of air monitoring stations in Taiwan, relying solely on data from these stations to represent regional air pollution levels has certain limitations [12]. Therefore, the air pollution data prediction and simulation model proposed in this study is essential, providing a scientific basis for a more comprehensive understanding and prediction of atmospheric mercury concentrations.

Previous research results have shown that applying deep learning techniques enables more accurate predictions of regional air quality [2,9,10]. In air pollution studies, many researchers have collected and combined air pollution data with a single machine-learning model to assess environmental benefits. Studies have indicated that combining a single machine-learning model with others can further enhance the performance of predictive data [1]. Among these models, the Transformer architecture has demonstrated excellent performance in the data training and preprocessing phases, with its self-attention mechanism enabling the model to simultaneously process all positions in the input sequence and more effectively capture long-range dependencies, thereby understanding internal data relationships. Additionally, this model can perform data preprocessing and weight training through fine-tuning, avoiding repetitive training, and maintaining the stability of node information. The BiGRU model improves accuracy by considering both past and future data points, making predictions more precise. The Quick Attention model, due to

its high computational efficiency and simple structure, allows the model to focus more on extracting relevant features from the data.

Mercury has long been proven to pose a threat to human life. In the twentieth century, the public gradually became aware of its significant impact on ecosystems and living environments. However, the current number of monitoring stations still needs to be improved, making research on mercury prediction relatively scarce compared to other air pollutants. Therefore, this study aims to utilize relevant data collected from Vietnam to predict mercury levels, addressing the gap in mercury prediction research.

This study proposes a novel advanced Trans-BiGRU-QA (Transformer-bidirectional gated recurrent unit-Quick Attension) hybrid model for predicting atmospheric mercury concentration. To evaluate its training effectiveness, it compares its performance with those of various machine learning models, including GRU, LSTM, RNN, Transformer, BiGRU, and Trans-BiGRU. The performance of the models are assessed based on the Minimum Error Value (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2). The study utilizes an air variable dataset collected from Vietnam, which includes five features: total gaseous mercury (TGM), temperature, relative humidity, $PM_{2.5}$, and carbon dioxide. These features are used to predict the hourly atmospheric mercury concentration.

2. Literature Review

2.1. Air Pollution

In urban industrialization, the combustion of fossil fuels has not only had a severe impact on human health, but the greenhouse gases generated by industrial activities have also exacerbated the global warming trend [12]. Air pollution poses environmental challenges and exposes adults, pregnant women, school-aged children, and fetuses to potential health risks [13]. In addition to being affected by domestic sources such as vehicles, industrial zones, and factories, Taiwan's air quality is also impacted by the influx of pollutants from abroad, contributing to land pollution. Table 1 summarizes air pollution-related studies, covering research areas, methods, results, and author years. It classifies them by research field, discussing the significance of atmospheric mercury and the application of machine learning models in data prediction.

Table 1. Summary of research on air pollution.

Field	Method	Result	Refs.
Importance of Atmospheric Mercury	Interaction between different air pollution variables and mercury	Strong positive correlation between carbon monoxide (CO) and gaseous mercury (Hg)	[9]
Using Machine Learning Models for Data Prediction	LSTM model predicting air quality in the Beijing–Tianjin–Hebei region	MAE: 25.26 R^2: 0.37	[10]
	Predicting air quality using RIDGE, SVR, RFR, ETR, and XGBOOST	$PM_{2.5}$: 0.67 (R^2) PM_{10}: 0.54 (R^2) NO_2: 0.69 (R^2)	[2]
	Performance of LSTM, CNN, SVM, and RF models with non-temporal inputs (TSs)	CNN-TS: 0.412 (MAE) CNN-TS: 0.252 (MSE)	[11]
	Predicting Air Quality Index through LSTM-GRU	R^2: 0.69 MAE: 36.12 RMSE: 57.77	[14]
	Predict O_3 and NO_2 concentrations through Res-GCN-BiLSTM	O_3 R^2: 0.85 RMSE: 10.60 NO_2 R^2: 0.88 RMSE: 9.05	[15]

As shown in previous studies, Wang et al. [9] investigated the interaction between various air pollution variables and mercury, revealing a robust positive correlation between carbon monoxide (CO) and gaseous mercury (Hg), as well as positive correlations with nitrogen dioxide (NO_2), $PM_{2.5}$, and PM_{10}. Xu et al. [10] employed the LSTM model to predict air quality in the Beijing–Tianjin–Hebei region of China, with results showing that the LSTM model outperformed the ARIMA regression model. For the Shijiazhuang site, the prediction results indicated that the LSTM model achieved a Mean Absolute Error (MAE) of 25.26 and an R^2 of 0.37. Samad et al. [2] used machine learning techniques to simulate accurate monitoring station data to predict $PM_{2.5}$, PM_{10}, and NO_2 concentrations. The results demonstrated that the best model achieved an R^2 of 0.67 for $PM_{2.5}$, 0.54 for PM_{10}, and 0.69 for NO_2. Wen et al. [11] used a deep learning model combined with time series (TS) techniques to accurately predict dioxin concentrations, with their proposed CNN-TS hybrid model being the best-performing model, achieving the best MAE of 0.412 and the best MSE of 0.252. They also found significant correlations between temperature, airflow, temporal dimensions, and dioxin levels. In summary, the air pollution-related studies presented in Table 1 indicate that air pollution poses a threat to the environment in Taiwan and has severe health implications for humans. In studies on air pollution forecasting using hybrid models, Sarkar et al. [14] used an LSTM-GRU model to predict the Air Quality Index, while Wu et al. [15] developed a Res-GCN-BiLSTM model to forecast NO_2 and O_3 levels. Both achieved excellent predictive performance.

2.2. Atmospheric Mercury

Mercury (Hg) is a potent neurotoxin. Both Pang et al. [8] and Skalny et al. [4] pointed out that mercury can lead to cardiovascular diseases in adults and cause severe damage to fetal neurocognitive functions. Additionally, mercury is believed to increase the risk of cancer. Mercury affects daily human life and poses serious health threats [1,4]. In previous studies, it was revealed that the critical components of atmospheric mercury include gaseous elemental mercury (GEM), particle-bound mercury (PBM), and gaseous oxidized mercury (GOM). The sum of GEM and GOM is defined as total gaseous mercury (TGM) [16]. Gaseous elemental mercury (GEM) accounts for up to 90% of atmospheric mercury. Luo et al. [17] further indicated that mercury can combine with air to form toxic atmospheric mercury, which is transported globally through airflows and deposited in remote areas.

Table 2 summarizes studies related to atmospheric mercury, covering research areas, methods, results, and author years. The literature on atmospheric mercury is categorized into two areas: seasonal variations of mercury components and regional impacts. Wang et al. [6] explored the two-year spatiotemporal variations and long-distance transport of atmospheric mercury forms in northern coastal China, revealing that the concentration of gaseous elemental mercury (GEM) exhibited significant seasonal variation. Yuan et al. [3] collected mercury (Hgp) and total gaseous mercury (TGM) samples from six coastal and island regions on both sides of the central Taiwan Strait, indicating that atmospheric mercury could affect Taiwan's air quality. Pang et al. [8] simulated the global health impact of atmospheric mercury emissions from small-scale artisanal gold mining (ASGM). Nguyen et al. [7] monitored the concentration of total gaseous mercury (TGM) during the monsoon season in Ho Chi Minh City, Vietnam. Their results showed that tropical cyclone monsoons interacted with TGM concentrations, potentially altering the levels of TGM.

Previous studies, such as those by Wang et al. [6], Yuan et al. [3], and Nguyen et al. [7], primarily focused on collecting atmospheric mercury variables through sensing equipment and analyzing their impact on regional environments. However, relatively little literature has explored the combined prediction of atmospheric mercury concentration with other air pollutants. Since mercury has already infiltrated daily human life, predicted air pollution and analyzed the composition of mercury has become increasingly urgent.

Table 2. Summary of research on atmospheric mercury.

Field	Method	Result	Refs.
Seasonal Variation in Mercury Components	Collected gaseous oxidized mercury (GOM) and elemental mercury (Hg) for sampling analysis	Concentration of gaseous elemental mercury (GEM) varies seasonally	[6]
Regional Impact of Mercury Components	Collected mercury (Hgp) and TGM samples in the Taiwan Strait	Atmospheric mercury may affect air quality in Taiwan	[3]
	Simulation of ASGM impact on residents and miners	Global public faces 1.5 times the risk of ASGM miners	[8]
	Monitoring total gaseous mercury concentration (TGM) in Ho Chi Minh City, Vietnam	Tropical cyclone monsoons interact with TGM concentration and may alter TGM levels	[7]

3. Methodology

3.1. Transformer

The Transformer model is a powerful neural network architecture often used for analyzing sequential data. It uses a mechanism called "attention" to focus on the most relevant parts of input data across long sequences. This allows it to capture dependencies across distant time steps more effectively than traditional models, making it ideal for tasks that require understanding complex patterns in data. It has been widely applied in fields such as Natural Language Processing (NLP) [18], computer vision [19], and speech recognition [20], demonstrating unparalleled computational performance. In recent years, many studies have begun applying the Transformer model to time series data analysis with remarkable success [21]. The core technology of the Transformer model—Self-Attention—effectively avoids the issues of forgetting and redundant computations that Recurrent Neural Networks (RNNs) may encounter when processing long sequences. Additionally, the Transformer model offers the advantage of parallel computation, significantly enhancing computational speed, and its attention mechanism dramatically improves the model's interpretability [22].

Figure 1 illustrates the architecture of the Transformer model [23]. The model consists of several components: Input Embedding, Output Embedding, Positional Encoding, the Encoder (on the left), the Decoder (on the right), Linear transformations, and a SoftMax normalization layer. Each layer of the Transformer model contains vital elements, including the Multi-Head Attention mechanism, Feedforward Neural Network, and Masked Multi-Head Attention mechanism.

In the Transformer architecture, the encoder maps the input sequence into a continuous representation, which is then passed to the decoder. The decoder receives the output from the encoder and the output from the decoder at the previous time step, ultimately generating the output sequence. The input embedding layer maps each token in the input sequence to real-valued vectors, enabling the model to manipulate and learn the representation of the input sequence. Positional encoding introduces the positional information of the sequence into the model by adding positional encodings to the embedding vectors, ensuring the model can distinguish variables at different positions. The output embedding layer maps the continuous-valued vectors generated by the model back to discrete symbols, which are used in subsequent tasks, such as generating class labels in text classification and machine translation.

The Multi-Head Attention mechanism extends the standard attention mechanism, allowing the model to focus on different parts of the input sequence across various representation subspaces. After the Self-Attention layer, the representation at each position is passed through a Feedforward Neural Network (FFN), which includes a ReLU activation function. This helps the model capture the nonlinear relationships at each position. The Masked Multi-Head Attention mechanism operates similarly to Multi-Head Attention but is designed to prevent the model from attending to future information during certain computations, ensuring that future data are not leaked at the current time step.

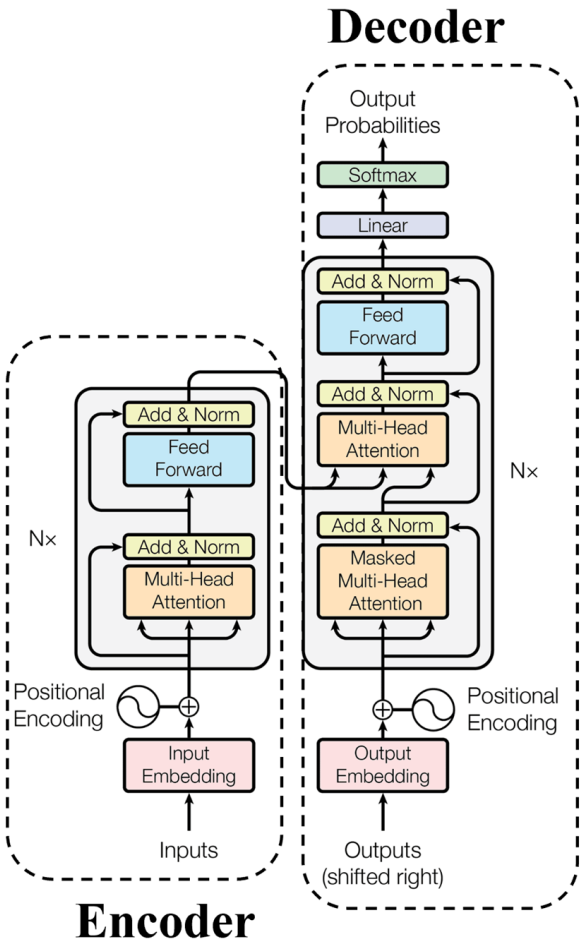


Figure 1. Architecture diagram of the transformer model.

Linear transformation is responsible for linearly projecting the input into a higher-dimensional space, enabling the network to learn features and representations more effectively. Finally, the crucial SoftMax normalization function is applied to classify the input into different categories by calculating the probability distribution across classes [23], providing a robust and reliable classification process.

The data undergo dimensional enhancement in the decoder before entering the Masked Multi-Head Attention layer in the decoder stage. This layer performs the decoding operation, where the query (Q) comes from the decoder's input, while the output of the encoder provides the key (K) and value (V). After training, the Feedforward Neural Network connects to and processes the data via the Add and norm layer. Before the final output, the data pass through another Add and norm layer, and the final computation result is produced [23].

The Transformer model is a Self-Attention structure comprising Scaled Dot-Product Attention and the Multi-Head Attention mechanism. Scaled Dot-Product Attention is the fundamental Self-Attention unit, where the input consists of queries (Q) and keys (K) with a dimension of d_k, and values (V) with a dimension of d_v. In the model proposed by Vaswani et al. [23], the Scaled Dot-Product Attention calculates the output by computing a weighted sum of all keys. Specifically, each key is scaled by $\sqrt{d_k}$, and then, the SoftMax normalization function is applied to compute the weight values.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

The Multi-Head Self-Attention mechanism involves multiple attention layers executing in parallel. Instead of applying a single attention function using keys, values, and queries with a dimensionality of d_{model}, the model benefits from using different learned linear projections of the queries, keys, and values into subspaces of dimensions d_k and d_k, and performing attention operations h times. Each query, key, and value operate in parallel within the attention function. The Multi-Head Self-Attention mechanism allows the model to simultaneously focus on different information and subspaces from various positions in the input sequence [23]. The equation for Multi-Head Self-Attention is as follows:

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^O,\ where\ head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (2)$$

where the set of Multi-Head Attention (projections) is the parameter matrix $W_i^Q \in R^{d_{model}*d_k}$, $W_i^K \in R^{d_{model}*d_k}$, $W_i^K \in R^{d_{model}*d_v}$, and $W^O \in R^{hd_v*d_{model}}$.

3.2. Gated Recurrent Unit (GRU) and Bidirectional Gated Recurrent Unit (BiGRU)

Figure 2 illustrates the architecture of the Gated Recurrent Unit (GRU) model [24]. As shown in Figure 2, the GRU is a variant of the Long Short-Term Memory (LSTM) model. This model addresses the issues of gradient explosion and gradient vanishing in Recurrent Neural Networks (RNNs) while retaining the predictive performance of LSTM. GRU's structure is more streamlined than that of LSTM, making training more accessible and improving the model training efficiency [25]. The GRU model has only two gates: the update gate and the reset gate, which reduces the computational complexity by eliminating one set of the matrix multiplications found in LSTM. As a result, the GRU model consumes less time when training with large datasets [26]. The following equations describe the architecture of the GRU model:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (3)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (4)$$

$$\widetilde{h_t} = tanh(W \cdot [r_t * h_{t-1}, x_t]) \quad (5)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t} \quad (6)$$

$$y_t = \sigma(W_o h_t) \quad (7)$$

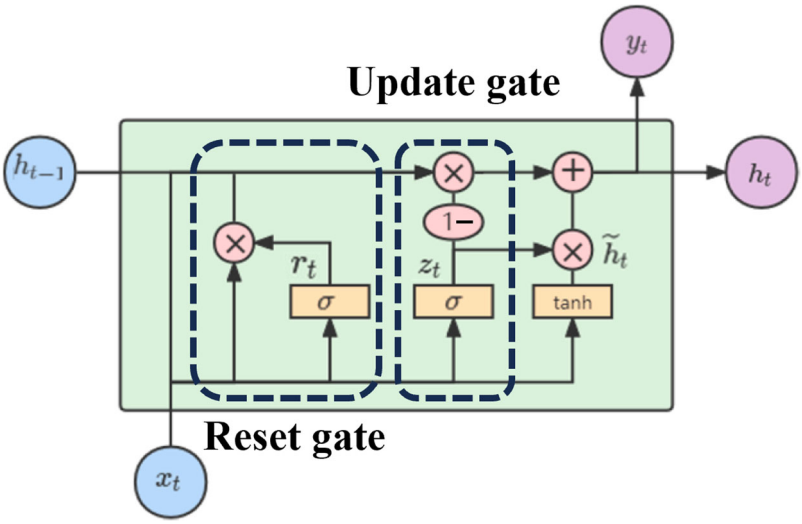


Figure 2. Architecture of the GRU model.

Equation (3) is the equation of the update gate. Update gates are used to control the ratio of past information to current information. $[h_{t-1}, X_t]$ uses the sigmoid function to determine z_t, and z_t allows for memory and forgetting. W_z is the weight matrix of the update gate [24,25]. σ is the sigmoid activation function, which compresses the output to the (0,1) interval and is used for gating structures (update gates and reset gates).

Equation (4) to Equation (6) are the equations of the reset gate. As shown in Equation (4) to Equation (6), tanh is the activation function, which compresses the output to the (−1,1) interval and is used to calculate candidate hidden states. The reset gate is used to forget the previous state information to obtain the post-reset data, and X_t and tanh are used in a scale of [−1,1] to retrieve $\widetilde{h}_t$. x_t and y_t are the input and output of time t, respectively [24,25]. h_{t-1} and h_t are the state output of the hidden layer unit at time $t-1$ and time t, respectively [24]. $(1-z_t) * h_{t-1}$ is used to forget the previous information in order to facilitate $z_t * \widetilde{h}_t$ to store new data. h_{t-1} is the neuron output at the previous moment, and x_t is the input at the current moment [25]. In the GRU model's architecture equation, r_t and z_t correspond to the outputs of the reset gate and update gate. W_r, W_z, W, and W_o are the weight matrices of the reset gate, update gate, hidden layer unit, and output layer, respectively [24].

In summary, the reset gate r_t allows the model to selectively ignore information from the previous time step, enabling a more flexible capture of short-term dependencies. The update gate z_t controls the proportion of past and new information in the hidden state at the current time step, allowing the model to capture both long-term and short-term dependencies flexibly. The candidate hidden state $\widetilde{h}_t$ is calculated using the current input and partial information from the previous time step, allowing the hidden state to reflect both current information and past influences.

Figure 3 illustrates the architecture of the Bidirectional Gated Recurrent Unit (BiGRU) model [22]. BiGRU is an advanced type of neural network that processes data in both directions, forward and backward. This bidirectional approach allows it to capture information from both the past and future of a data sequence, making it particularly effective for tasks like time series predictions where context from all sides is valuable. As shown in Figure 3, the BiGRU model is a machine-learning model composed of a forward GRU and a backward GRU. The BiGRU model effectively extracts deep features from data by simultaneously performing time series computations in both forward and backward directions. This bidirectional computation mitigates the disadvantage in the standard GRU model, where the input data sequence depends solely on the results from the final time step. Consequently, this approach significantly improves the model's prediction accuracy [24,27].

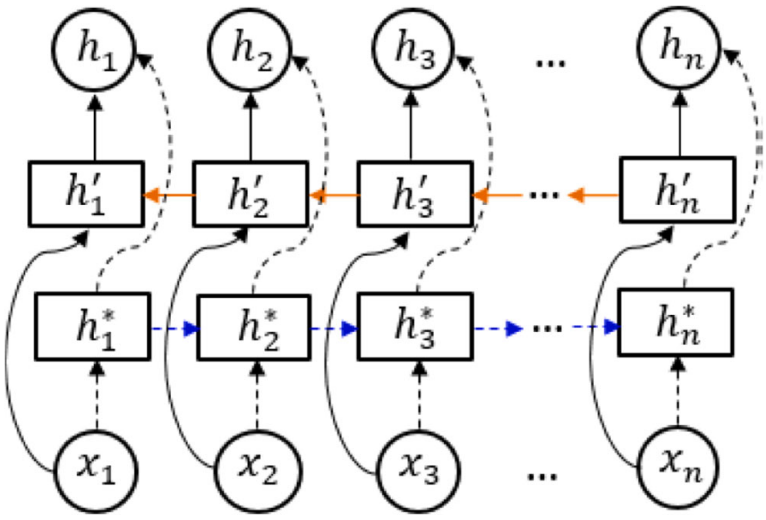


Figure 3. Architecture of the BiGRU model.

In summary, BiGRU processes sequences in both the forward and backward directions, enabling the model to extract more comprehensive information from context [28]. This is particularly effective for tasks that require contextual understanding, such as language comprehension and trend detection in time series forecasting. The forward and backward GRUs include reset and update gates, which control information filtering and updating. Consequently, BiGRU is well-suited for tasks that demand full contextual information, such as natural language processing, time series prediction, and speech recognition.

3.3. Quick Attention

Quick Attention (QA) is a simplified and efficient type of attention mechanism that helps the model prioritize important features in data without excessive computational demand. By quickly highlighting essential information, it enhances the model's ability to focus on the most relevant data points, improving prediction accuracy while keeping the processing time low [29]. Figure 4 illustrates its architecture. As shown in the figure, the Quick Attention model performs feature learning by applying a $1 \times 1 \times C$ convolution operation on the feature map, enabling the model to focus on extracting relevant features and enhancing global and local feature capture [29]. In the architecture of the Quick Attention model, the input feature map is duplicated twice, followed by $1 \times 1 \times C$ convolution operations. The results are then passed through a sigmoid function to compute the attention weights. Finally, the weighted features are added to the original input features, generating the final attention feature map as the output.

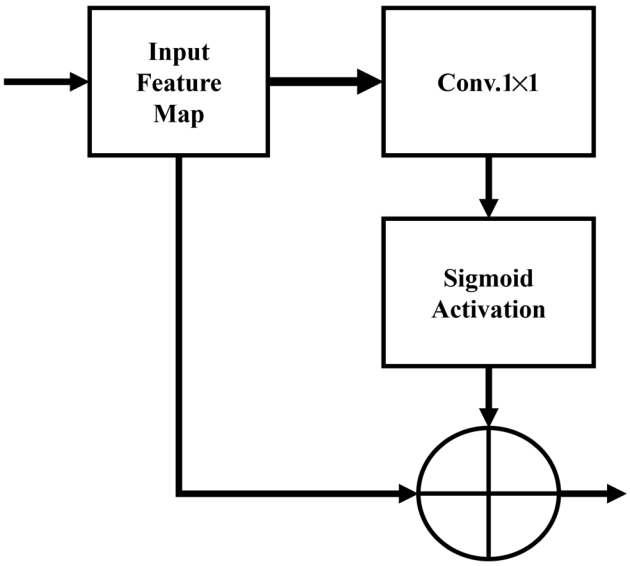


Figure 4. Architecture of the Quick Attention model.

The Quick Attention model focuses on learning essential features, effectively reducing the interference of noisy features and improving the model's stability and prediction accuracy [30]. The following equations describe the architecture of the Quick Attention model:

$$QA(x) = \sigma\left(f(x)^{1\times1}\right) + x \tag{8}$$

Equation (8) is the architectural equation of the Quick Attention model. As shown in Equation (8), x is the input feature image, and σ is the sigmoid function. In the Quick Attention model, the feature map $f(x)^{1\times1}$ is a 1×1 convolutional layer with a stride of 1 and the same number of filters as the input, ensuring that the output retains the same dimensionality as the input. The input feature image is structured as $W \times H \times C$ (width $\times$ height $\times$ channels), and it is duplicated twice before feature analysis begins [29].

Quick Attention simplifies the computational process compared to standard Multi-Head Self-Attention, significantly reducing resource consumption, especially for long sequences. By using attention weights to selectively focus on essential features within a sequence, Quick Attention minimizes the interference of irrelevant features, enhancing the model's predictive accuracy. It balances focused attention and lower computational cost, making it ideal for applications requiring rapid inference or where resources are limited.

3.4. The Proposed Trans-BiGRU-QA Hybrid Model

Figure 5 presents the architecture of the proposed Trans-BiGRU-QA hybrid model in this study. The orange section highlights the Transformer model. The Transformer model performs exceptionally well in data training and preprocessing, with its Self-Attention mechanism capable of simultaneously processing all positions within the input sequence. This mechanism effectively captures long-range dependencies, enabling a better understanding of the internal relationships within the data [23]. Since the original Transformer model's Input Embedding and Positional Encoding are designed to convert signals in Natural Language Processing (NLP) tasks into vectors, and the dataset in this study consists of floating-point data (float); these components were removed. Additionally, considering the scope of the Transformer model's application in this study, the output from the decoder (Transformer Outputs) is not utilized during the initial training phase.

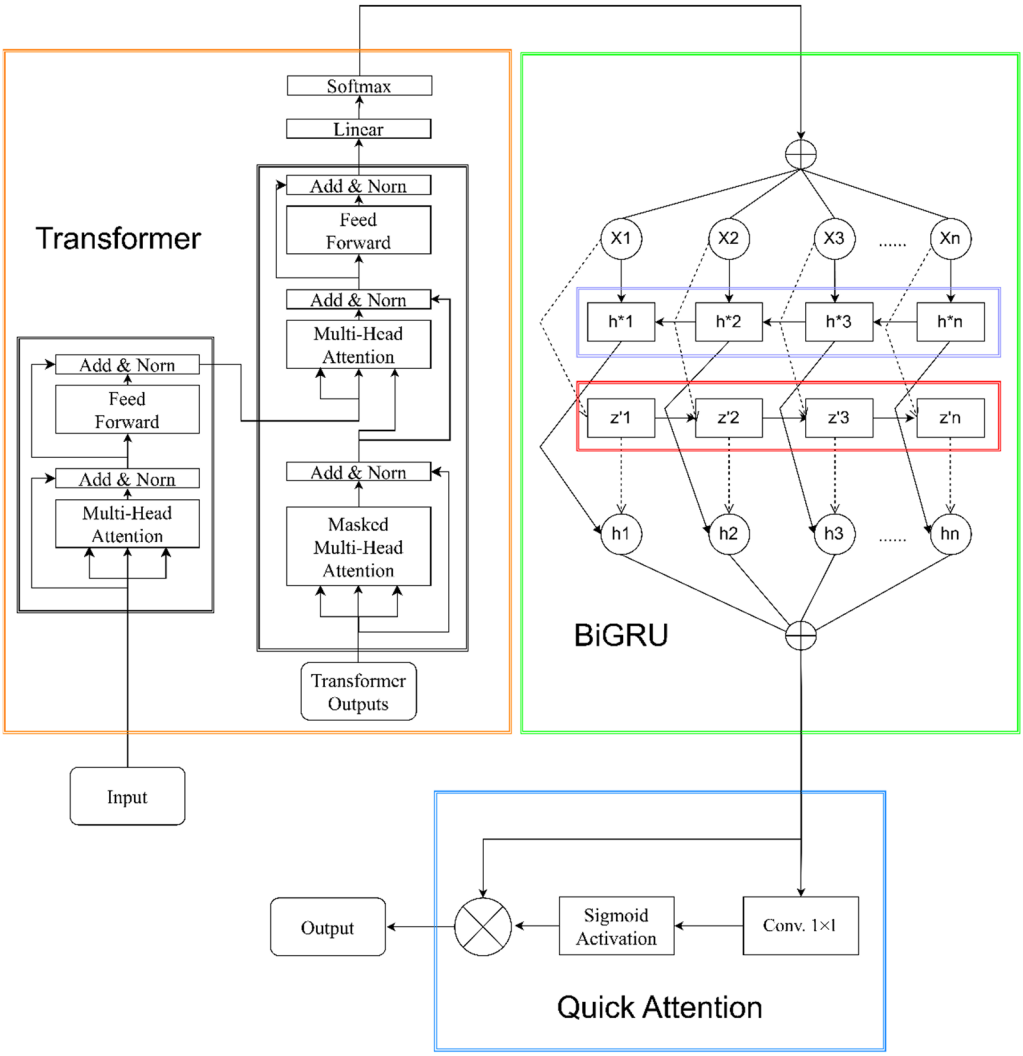


Figure 5. Architecture of the proposed Trans-BiGRU-QA hybrid model.

The hybrid model proposed in this study aims to leverage the advantages of the Multi-Head Attention mechanism to capture long-range dependencies better, thus selecting the Transformer model as the first part of the hybrid structure. The BiGRU model, represented in the green box, is integrated as the middle block. The BiGRU extracts both forward and backward deep features simultaneously, enabling the model to capture deeper data characteristics comprehensively [24,27]. Finally, the Quick Attention mechanism, shown in the blue box, is incorporated at the end of the model to enhance the learning of essential features, thereby improving prediction accuracy and model stability. Since the data predictions in this study are not related to image prediction, the final addition operation in the Quick Attention model is modified to a multiplication-sum operation. By combining the Self-Attention mechanism of the Transformer model, the BiGRU model, and the Quick Attention mechanism, this study aims to improve training speed and prediction accuracy, validating the superiority of the proposed Trans-BiGRU-QA hybrid model.

The hybrid model proposed in this study consists of an input layer (Input) and an output layer (Output), internally integrating the Transformer model, BiGRU model, and Quick Attention mechanism. First, data enter through the input layer and are immediately passed to the encoder part of the Transformer model. Within the encoder, the Multi-Head Attention mechanism processes the data, which learns the internal relationships within the input sequence. The training effectiveness in deep learning is further enhanced through residual connections and normalization (Add & Norm) combined with the Feedforward Neural Network. Once the Feedforward Neural Network training is complete, the data are transmitted via the Add & Norm layer to the decoder. During decoding, the decoder compares the optimal features from the previous time step with the current features, decoding data through the Masked Multi-Head Attention layer. Subsequently, the data pass again through the Add & Norm layer and are processed by the Feedforward Neural Network. Before generating the final output, a linear transformation (Linear) and normalization (SoftMax) are applied to improve the accuracy of the data training further.

Subsequently, the data undergo a summation operation and are passed to the BiGRU model. In the BiGRU model, the input (X) is processed through both forward and backward layers, ultimately generating the output (h). The hidden layers of the BiGRU model consist of X1, X2, X3, . . ., up to Xn. The forward GRU layer (red box) and the backward GRU layer (purple box) extract deep features from the current and the next time step, respectively. For example, taking X1 as a reference, the input X1 is simultaneously passed to the forward GRU layer (z'1) and the backward GRU layer (h1). Since X1 is the first step in the sequence, the forward GRU layer z'1 only receives information from X1, and the backward GRU layer h1 only processes information from X1. These two GRU layers calculate their respective hidden states and combine them to generate the final output h1, which includes the optimal deep features from both the forward and backward GRU layers. This process continues through the sequence, and the final hidden states h1, h2, h3, . . ., up to hn are summed and then passed into the Quick Attention mechanism for further processing.

In the Quick Attention (QA) mechanism, the input feature data are duplicated twice and processed through a $1 \times 1 \times C$ convolution for feature learning. Following this, the sigmoid function is applied to compute the attention weights. The original input features are then combined with the generated results through an element-wise multiplication operation, ultimately producing the Attention data as the final output of the model. The Quick Attention model, through the $1 \times 1 \times C$ convolutional feature learning, focuses on extracting key features, thereby enhancing the model's ability to capture essential data characteristics [29]. QA is a lightweight and efficient attention mechanism that maintains computational efficiency while focusing on a sequence's most representative and relevant parts. BiGRU inherently has bidirectional memory, allowing it to consider a sequence's preceding and succeeding context. However, more than BiGRU alone may still be required in practical applications to capture key features. By integrating QA, the model can quickly focus on critical instantaneous or periodic information within time series data, enhancing

BiGRU's feature-learning capabilities and enabling it to capture essential patterns and characteristics in the data more effectively.

In the Trans-BiGRU-QA model architecture proposed in this study, each component plays a distinct and crucial role in predicting atmospheric mercury concentrations: Transformer: Leveraging its self-attention mechanism, the Transformer effectively captures long-range dependencies within time series data. This is especially important for predicting atmospheric mercury concentrations, as fluctuations in concentration may be related to specific events at past time points. Its Multi-Head Attention mechanism allows the model to learn various patterns or trends in the data from multiple perspectives. For mercury concentration prediction, this means the model can capture both short-term fluctuations and long-term trends, thereby enhancing prediction accuracy.

BiGRU considers both the preceding and succeeding relationships within a sequence, making it particularly effective at capturing contextual information related to changes in mercury concentration. Compared to traditional LSTM, the GRU structure is more straightforward and has fewer parameters, yet it maintains sequential solid modeling capabilities. This makes BiGRU an efficient and stable choice, helping to improve model efficiency, reduce the risk of overfitting, and enhance stability in sequence prediction tasks.

In the Trans-BiGRU-QA model, QA filters and enhances vital features. The QA mechanism focuses on critical information, such as significant variation points or periodic patterns based on the feature distribution of atmospheric mercury concentrations, rather than processing all data points equally. Its lightweight design enables faster attention operations, avoiding the high computational cost of traditional Self-Attention mechanisms. This efficiency advantage of QA is precious when handling large-scale environmental data.

3.5. Evaluation Metrics

To evaluate the accuracy of the prediction models, this study employed multiple evaluation metrics to determine the precision and relative advantages of the experimental models. Specifically, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2) were used to assess the performance of deep learning models in predicting atmospheric mercury levels. RMSE provides an immediate reflection of the deviation between actual and predicted values, while MAE measures the absolute difference between the actual and predicted values [31]. Additionally, the R^2 coefficient is used to evaluate the model's and the dataset's fit, exploring the correlation between the actual and predicted values [31,32].

MAE directly measures the model's prediction accuracy, assessing the degree of deviation between predicted and actual observed values. In atmospheric mercury concentration prediction, MAE helps to understand the model's average error in daily data fluctuations, reflecting its accuracy under typical conditions. RMSE, as the square root of the Mean Squared Error (MSE), is more sensitive to larger error values. When the model exhibits significant prediction deviations for specific samples, RMSE magnifies these outlier errors, indicating the model's performance in handling high-error cases. R^2 provides an intuitive understanding of the model's fit to the data. In complex and variable data such as mercury concentration, R^2 helps determine whether the model effectively captures the main trends in concentration changes

Through the combined evaluation of MAE, RMSE, and R^2, the model's effectiveness in predicting atmospheric mercury concentration can be assessed from various perspectives. MAE and RMSE provide insight into the range of errors, while R^2 reflects the model's ability to explain the variance in the data. Using these metrics together allows for a more comprehensive understanding of the model's strengths and weaknesses, facilitating targeted adjustments and optimizations to enhance reliability in practical applications.

The following are the equations for calculating the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2):

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{n}|\hat{y} - y| \tag{9}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(\hat{y} - y)^2} \tag{10}$$

4. Data Description

The data used in this study came from the Vietnam air pollution dataset. Table 3 summarizes the attributes of this dataset. As shown in Table 3, the dataset contains 1272 data entries spanning from 13 September 2022 to 21 June 2023. The dataset includes six input variables, $X_1(t)$ to $X_6(t)$, which are in the following order: Temperature (Temp), Relative Humidity (RH), Fine Particulate Matter ($PM_{2.5}$), Carbon Dioxide (CO_2), Total Gaseous Mercury (TGM), and the current Atmospheric Mercury Concentration (Hg(t)). The output variable, Y_{t+1}, represents the atmospheric mercury concentration at the next time step (Hg(t + 1)).

Table 3. Summary of data attributes.

Variable	Name	Description	Data Type	Update Frequency	Unit
$X_1(t)$	Temp	Physical quantity representing air temperature	Float	Hourly	°C
$X_2(t)$	RH	Predicted probability of rainfall in meteorology	Float	Hourly	%
$X_3(t)$	$PM_{2.5}$	Fine particulate matter harmful to human health	Float	Hourly	ug/m^3
$X_4(t)$	CO_2	Major greenhouse gas contributing to global warming	Float	Hourly	ppm
$X_5(t)$	TGM	Chemical component affecting human health	Float	Hourly	ng/m^3
$X_6(t)$	Hg(t)	Current atmospheric mercury index	Float	Hourly	ng/m^3
Y_{t+1}	Hg(t + 1)	Predicted atmospheric mercury index for the next hour	Float	Hourly	ng/m^3

Additionally, this study summarized the air variables used in previous research in Table 3 to further explore the relationship between the air prediction variables and atmospheric mercury index in this study. Such analysis will help reveal the impact of different air pollutants on the atmospheric mercury concentration and enhance the accuracy of the model's predictions.

The data type used in this study was floating-point (Float), with the data being updated hourly daily. According to the publicly available Air Quality Index data provided by Taiwan's Environmental Protection Agency's Air Quality Monitoring Network (https://airtw.moenv.gov.tw/, accessed on 1 July 2023), the existing publicly available data are updated hourly and daily. Therefore, this study utilized the existing input variables $X_1(t)$ to $X_6(t)$ to predict the atmospheric mercury index for the next hour using the model, allowing for an analysis of the distribution of mercury in the air. Some indirect factors can also directly influence the concentration of mercury in the atmosphere and air pollution, including:

- Volatility of Mercury: Mercury exists in various forms in the environment and can enter the human body through air, food, and water. The primary sources of mercury emissions are industrial activities, waste incineration, and coal-fired power generation. Mercury typically remains in the air as particulate-bound mercury and

gaseous mercury [33]. The temperature is one of the critical factors influencing mercury volatility. Under high-temperature conditions, mercury tends to volatilize more easily [34]. Higher temperatures may increase the volatility of certain mercury compounds, thereby exacerbating mercury pollution [34,35]. Mercury's high volatility and low water solubility significantly contribute to air pollution [36].

- Effect of Humidity in the Atmosphere: Mercury is a heavy metal element with intense volatility [36]. The relative humidity can influence the volatility of mercury, as higher humidity levels cause water molecules in the air to bind with mercury atoms, forming mercury hydrates. This process reduces the volatility of mercury [37,38]. In environments with high relative humidity, mercury pollution can worsen due to the decreased ability of mercury to volatilize effectively, leading to its accumulation in the atmosphere.
- Mercury and Fine Particulate Matter: Mercury pollution threatens the environment and human health [3,4]. Fine particulate matter ($PM_{2.5}$) is one of the central air pollutants and can bind with mercury, prolonging the time mercury remains suspended in the air. In environments with higher concentrations of fine particulate matter, the pollution level of mercury increases as the particles enhance mercury's persistence in the atmosphere [37,38].
- Mercury and Carbon Dioxide: The interaction between mercury and carbon dioxide is an often-unnoticed environmental issue. Carbon dioxide can combine with mercury to form compounds, increasing the amount of mercury in the environment. Mercury–carbon dioxide compounds are insoluble in water, and their presence raises mercury levels in the environment, exacerbating mercury pollution [37]. With global climate change, mercury pollution may pose even more significant environmental hazards [39,40].

The interactions between mercury and factors such as volatility, humidity effects, fine particulate matter, and carbon dioxide reveal that mercury can indirectly influence atmospheric mercury concentrations through these air variables. This, in turn, contributes to air pollution and poses risks to human health and the living environment [33,36–38].

4.1. Feature Engineering

Feature engineering is constructing features (input variables), selecting them based on domain knowledge, modifying them, or enhancing the efficiency of machine learning methods [41]. The feature engineering process in this study consists of five stages: feature adjustment, feature transformation, feature normalization, dimensionality reduction, and feature importance assessment.

This study employed feature engineering methods, including feature adjustment, feature normalization, and feature importance assessment, to conduct data preprocessing and feature comparison. Feature adjustment aims to modify the data to reduce skewness appropriately [41] and eliminate abnormal data, such as missing or erroneous values, which may distort training performance [42]. Feature normalization enhances the model's training efficiency and convergence speed by scaling all features to a uniform numerical range [41]. Finally, feature importance assessment is used to measure the contribution of each feature in the machine learning model's prediction process and evaluate its impact on the model's predictive results.

4.1.1. Feature Adjustment

Feature adjustment involves the appropriate transformation or modification of data to reduce skewness. This process includes mathematically correcting heavily skewed or unevenly distributed data, bringing them closer to a normal distribution. Thus, positive or negative skewness in the distribution can be effectively corrected, making the data more symmetric [41]. Incorrect or missing data and other anomalies may distort the training performance [42]. Even if the dataset does not contain explicit errors, there may still be anomalous elements that differ from other patterns within the dataset. Advanced methods

for detecting outliers in data samples extend beyond measuring average time intervals; they can fit the observed data within a set period to a curve, allowing for direct comparisons [43]. In this study, the dataset's missing data handling and outlier detection were performed using SPSS 25. Missing values were imputed through linear interpolation, while outlier detection commonly utilized Cook's Distance to assess anomalies in the data. Cook's Distance is a statistical measure used to evaluate the degree of change in the parameter estimates of a regression model when a particular observation is removed.

Cook's Distance is calculated as follows:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{y}_j - \hat{y}_{j(i)} \right)^2}{\rho \cdot MSE} \tag{11}$$

D_i represents the Cook's Distance for the i-th observation, where $\hat{y}_j$ is the predicted value of the j-th observation when all data points are included, and $\hat{y}_{j(i)}$ is the predicted value of the j-th observation after removing the i-th observation. ρ denotes the number of estimated parameters in the model (including the intercept). The Mean Squared Error (MSE) is used to standardize Cook's Distance, allowing for the comparability of distances across different models. Cook's Distance can be viewed as a measure of the overall change in model predictions after removing a particular observation, with standardization by MSE to ensure consistency across models.

Before conducting deep learning training, this study performed outlier detection and data cleaning. During the data cleaning process, missing values, anomalies, and duplicate data were removed to ensure that the data fell within normal ranges. This process also minimized the impact of outliers on the model's prediction accuracy, ensuring that the dataset was clean and reliable for training.

4.1.2. Feature Normalization

Feature normalization involves scaling all features to the exact numerical range to improve model training effectiveness and speed up convergence [41]. Data normalization transforms data into a specific range, such as scaling it between -1 and 1, or between 0 and 1. Normalization becomes especially important when the numerical ranges of different features in the dataset vary significantly. When there are no missing values or anomalies in the dataset, Min-max scaling is a beneficial normalization technique [44]. This study applied Min-max scaling to normalize the air pollution dataset. The equation for Min-max normalization is as follows:

$$X' = \frac{x - min}{max - min} \tag{12}$$

Equation (12) is the equation for data normalization. X' is the normalized data, x is the denormalized data, and min and max have the same values used previously in the normalization process.

4.1.3. Feature Importance Evaluation

Feature importance assessment measures the contribution of each feature during the model prediction process and evaluates its impact on the machine learning model's predictive outcomes. Through this process, researchers can identify the features that significantly influence model performance, enabling model optimization and selecting the best features [45]. This study used Pearson correlation coefficients and SHAP values to assess feature importance. First, the Pearson correlation coefficient was used to observe the relationships among the five variables: Atmospheric Mercury (TGM), Temperature (Temp), Relative Humidity (RH), $PM_{2.5}$, and Carbon Dioxide (CO_2). Second, SHAP values were applied to quantify each feature's contribution to the prediction outcomes, helping to identify which features are most critical to the model's overall performance.

4.2. Sliding Window

In this study, the variables of the data attributes in Table 3 were added to $\boldsymbol{X}_t = \{X_1(t), X_2(t), \ldots, X_6(t)\}$. The details of the variables in the sliding window of this study are shown in Figure 6.

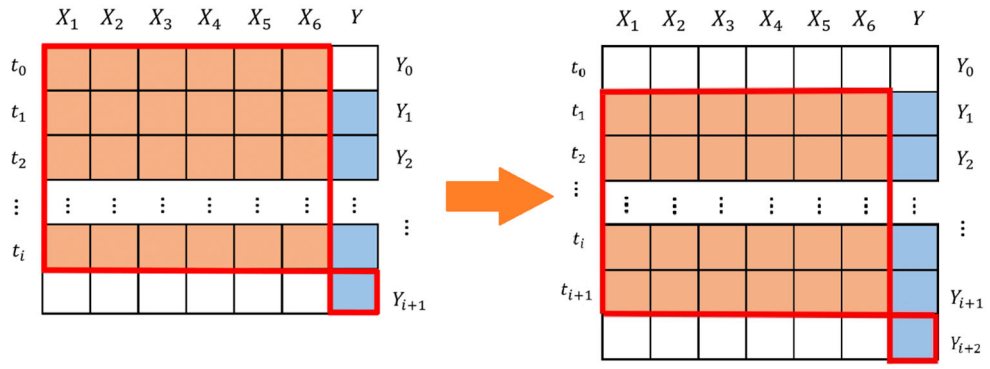


Figure 6. Sliding window with a width of $i + 1$.

Figure 6 illustrates the sliding window equation with a width of $i + 1$. As shown in Figure 6, the process begins from the left side, where there is no output data for the initial value t_0; so, it is excluded, and the first output is set as Y_{i+1}. The input variables, X_1 to X_6, are the predictive features, including Temperature (Temp), Relative Humidity (RH), Fine Particulate Matter ($PM_{2.5}$), Carbon Dioxide (CO_2), Total Gaseous Mercury (TGM), and Current Atmospheric Mercury (Hg(t)). The output variable, Y, represents the next hour's Atmospheric Mercury Index (Hg($t + 1$)). In the prediction process, t_i represents the time the prediction is made, with iii indicating the corresponding data point. For instance, at t_0, the input variables $X_1(t)$ to $X_6(t)$ are used to predict the next hour's atmospheric mercury index (Y_{i+1}), which corresponds to the prediction for the next state, t_1. Once the prediction for (Y_{i+1}) is completed, the process moves forward to predict the next hour's atmospheric mercury index (Y_{i+2}), as shown from left to right in the figure. In this study, the input air variables and current atmospheric mercury levels were combined to predict the next atmospheric mercury value, thereby analyzing the gas distribution of mercury over time.

The window size is carefully selected to balance between capturing a comprehensive history of the time series data and keeping the dataset manageable for efficient processing in this study. A smaller window size might be more responsive but could miss longer-term patterns, while a larger window size captures more historical context at the cost of increased computational load and potential overfitting. In this study, the window size was optimized based on validation tests to maximize predictive accuracy without compromising the model's responsiveness to recent data changes.

At the beginning of the prediction series, where there are insufficient data to fill the window, specific strategies were implemented to avoid inaccuracies. Our approach was padding, where initial values are filled with zeros or the first available data point, ensuring consistent window size without introducing abrupt changes. Alternatively, predictions may start only after the window is fully populated, which provides stable and reliable input data for the model without needing artificial adjustments.

5. Experiment and Discussion

5.1. Experimental Process

Figure 7 illustrates the experimental workflow of this study. At the beginning of the research, atmospheric mercury air data underwent feature engineering, which included feature adjustment, feature normalization, and feature importance assessment for data preprocessing and feature comparison. (1). Feature Adjustment: Outlier detection and data cleaning is the first step. Outlier detection aims to modify the data by reducing skewness and effectively correcting skewed distributions, making the data more symmetric. Data

cleaning removes erroneous data and missing values to enhance the effectiveness of data training. (2). Feature Normalization: This step scales all features to a uniform numerical range to improve the model's training performance. The Min-max method is applied to optimize the model's training efficiency. (3). Feature Importance Assessment: The final step involves evaluating the importance of features by exploring their relationships using a Pearson correlation matrix. SHAP values are then used to quantify each feature's contribution to the prediction, highlighting the most significant features for the model's performance. This comprehensive process ensures that the data are well-prepared for the predictive modeling phase and that the most important features are identified to enhance model accuracy.

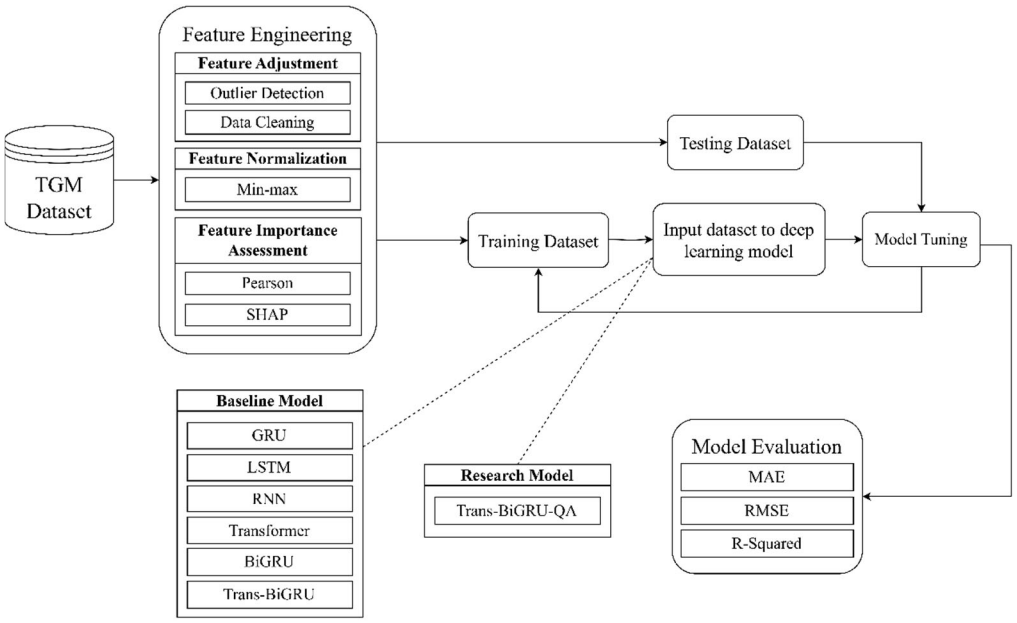


Figure 7. Flowchart of the experimental process.

Next, the data were split into training and testing sets, and the datasets were fed into baseline models for training. The baseline models included Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Transformer, Bidirectional Gated Recurrent Unit (BiGRU), and the hybrid Transformer and Bidirectional Gated Recurrent Unit model (Trans-BiGRU). These models were compared with the proposed Trans-BiGRU-QA hybrid model through benchmarking tests. Model hyperparameter tuning and data validation were conducted during the training process. The model performance was evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination R^2 (R-Squared). Each baseline and hybrid model was run ten times, and the average MAE and RMSE were used as the final evaluation metrics. The standard deviation (STD) was calculated to assess the stability of the models. This study employed grid search to optimize the proposed model's hyperparameter settings to identify the best hyperparameters and achieve the most accurate predictions. This process ensures that the model is finely tuned for optimal performance in predicting atmospheric mercury concentrations.

The dataset used in this study came from Vietnam and is related to air pollution. The original dataset consists of 1272 records, covering from 13 September 2022 to 21 June 2023. Table 4 summarizes the descriptive statistics for all variables. The table includes the count of data entries for each variable (Count), minimum value (Min), median value (Median), mean value (Mean), maximum value (Max), and standard deviation (STD). The detailed statistical results are presented in Table 4. This descriptive analysis not only provides an overview of the distribution and variability of the variables but also offers valuable insights into the data, enhancing their value for model training and evaluation.

Table 4. Descriptive statistics of all variables.

Variables	Count	Min	Median	Mean	Max	STD
TGM	1272	1.297	2.031	2.134	4.66	0.558
Temp	1272	23.262	28.445	29.104	38.735	3.04
RH	1272	34.354	77.025	74.306	94.254	12.718
$PM_{2.5}$	1272	1.003	32.938	35.057	264.416	19.347
CO_2	1272	390.312	421.638	426.938	529.913	22.038
$Hg(t)$	1272	1.297	2.031	2.135	4.66	0.558

To enhance the accuracy of data predictions, it is necessary to split the entire dataset into subsets and establish separate models based on different conditions during the data processing phase. This approach relies on large datasets to provide sufficient data for training and testing each model [46]. In this study, a 70:30 ratio divided the dataset for training and testing. After feature adjustment in the feature engineering process, the total number of available data entries was 1245. The training dataset covered from 13 September 2022 to 10 January 2023, representing 70% of the total data. The testing datasets spanned from 11 January 2023 to 21 June 2023, representing 30% of the total data.

5.2. Experimental Parameter Settings

5.2.1. Environment Setup

This study used PyCharm as the development environment and Python version 3.11 as the programming language. The libraries used for machine learning model analysis included Pandas, Numpy, Matplotlib, and Scikit-learn. The hardware configuration for the experiments is outlined in Table 5. This setup ensured compatibility with the required libraries and provided the necessary computational resources for executing the machine-learning models efficiently.

Table 5. Specifications of the experimental equipment.

Name	Model/Version
Operating System	Windows 10 22H2
Processor	Intel Core i7-13700
Memory	ADATA DDR4-3200 64G
SSD	Kingston KC3000 1TB
Graphics Card	Nvidia GeForce RTX 3060

5.2.2. Hyperparameter Experiment Setup

To perform model tuning, this study employed grid search to find the optimal parameters during the training of the learning model. The experimental steps were as follows: (1). Extract the raw data. (2). Partition of the raw data into training and testing sets. (3). Adjust or create machine learning model objects. (4). In Scikit-learn, set up and instantiate the grid search method. (5). Fit the grid search model to the data, identifying the best parameters and model for prediction [31]. Various combinations were tested during the parameter selection process, and the best-performing parameters were chosen as the final result. This approach ensured that the most influential parameter configuration was selected for accurate and efficient model predictions.

The learning rate is a crucial parameter that affects the convergence speed of the model. If the learning rate is too large, it may cause the model to fail to converge; if the learning rate is too low, while it may more stably find the global optimum, the convergence speed will be slower, and it might not reach the optimal value within an acceptable number of training epochs [31]. Epochs refer to the number of times the entire dataset is passed through the model during training, while batch size determines the number of samples fed into the model before each weight update. Balancing these parameters is critical to achieving efficient and accurate model training.

Additionally, the Num_head parameter in the Transformer model determines the number of attention heads in the Multi-Head Attention mechanism. The Multi-Head mechanism enables the model to focus on different parts of the input, and increasing the number of heads can improve the model's ability to capture diverse spatiotemporal features [47]. The Key_dim defines the dimension of each attention head responsible for processing the internal information within the model, and it influences both the complexity and the processing capability of the model [47].

Table 6 summarizes the reference literature for the GRU, LSTM, RNN, Transformer, BiGRU, and Trans-BiGRU models, providing a detailed comparison of their hyperparameter settings, including Batch_size, Epoch, Learning rate, Num_head, and Key_dim.

Table 6. Summary of hyperparameter tuning.

Model	Batch_Size	Epoch	Learning Rate	Num_Head	Key_Dim
GRU	128	30	0.001	-	-
LSTM	[1, 100]	-	[0, 0.01]	-	-
RNN	100	[100, 500]	0.01	-	
Transformer	32	8	1×10^{-4}	4	8
BiGRU	128	100	0.005	-	-
Trans-BiGRU	150	150	0.005	4	8

5.2.3. Parameter Settings of the Proposed Trans-BiGRU-QA Hybrid Model

This study's proposed Trans-BiGRU-QA hybrid model integrates the characteristics of the Transformer model, BiGRU model, and Quick Attention mechanism. This study's batch size was set to 32, the epochs to 100, and the learning rate to 0.0001.

In the hybrid Trans-BiGRU-QA model, the fully connected layer (Dense layer) was set to 64 neurons. Additionally, since this study introduced the Quick Attention mechanism to enhance the extraction of data features, the fully connected layer in the Quick Attention model (Dense_layer—QA) was set to 32 neurons. The BiGRU layer, responsible for capturing both forward and backward deep learning features, was set to 64 neurons in the BiGRU model. In the Transformer model, the default values for the Multi-Head mechanism (Num_head) and the size of each head (Key_dim) were set to 4 and 64, respectively. In the initial design phase, default hyperparameters were used for experimentation, and future experiments will employ grid search to optimize and progressively determine the best model parameter settings. Table 7 outlines the detailed parameter settings for the Trans-BiGRU-QA hybrid model.

Table 7. Parameter settings of the Trans-BiGRU-QA hybrid model.

Model	Batch_Size	Epoch	Learning Rate	Dense_Layer
Trans-BiGRU-QA	32	100	0.0001	64
	Dense_layer(QA)	BiGRU layer	Num_head	Key_dim
	32	64	4	64

In the Trans-BiGRU-QA model, the results obtained from grid search hyperparameter tuning include several fields, each representing different hyperparameter configurations within the model. Batch size indicates the number of samples used in each parameter update, while epoch refers to the total number of complete learning cycles over the entire training dataset. Learning rate controls the step size during each parameter update; a smaller learning rate, such as 0.0001, enhances training stability, reducing the risk of overshooting the optimal solution, which is particularly suitable for complex deep learning models. The dense layer denotes the number of units in the fully connected layer. Num_head represents the number of heads in the Multi-Head Attention mechanism within Quick Attention. At the same time, Key_dim specifies the dimensionality of the Key vector

for each head in the attention mechanism, set to 64 in this paper, meaning that each head's Key vector contained 64 dimensions.

5.3. Experimental Results

5.3.1. Performance Comparison of Each Model

To evaluate the accuracy of the predictive models, this study employed several performance metrics to compare the precision of each model. Table 8 presents a performance comparison using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2) as critical evaluation criteria for the predictive models. Standard deviation (STD) is an essential indicator for assessing model performance, particularly in deep-learning model evaluations. Standard deviation is often used to measure the variability of results across multiple experimental runs, allowing researchers to assess the stability of model performance. A more minor standard deviation indicates more stable model performance. In this study, the standard deviation was used to evaluate the stability of each model. The MAE, RMSE, and R-squared values in Table 8 represent the average of ten runs for each model. The table also includes the average MAE and RMSE values and their respective standard deviations for each model.

Table 8. Performance comparison of each model.

	MAE ± SD	**RMSE ± SD**	**R^2**
GRU	$0.0567 \pm 7.63 \times 10^{-4}$	$0.0860 \pm 3.82 \times 10^{-4}$	0.69
LSTM	$0.0939 \pm 4.10 \times 10^{-4}$	$0.1260 \pm 2.72 \times 10^{-4}$	0.34
RNN	$0.0930 \pm 46.44 \times 10^{-4}$	$0.1197 \pm 34.38 \times 10^{-4}$	0.33
Transformer	$0.0644 \pm 19.54 \times 10^{-4}$	$0.0909 \pm 22.14 \times 10^{-4}$	0.66
BiGRU	$0.0574 \pm 39.94 \times 10^{-4}$	$0.0865 \pm 38.49 \times 10^{-4}$	0.69
Trans-BiGRU	$0.0543 \pm 18.29 \times 10^{-4}$	$0.0817 \pm 22.44 \times 10^{-4}$	0.67
Trans-BiGRU-QA	$0.0509 \pm 15.80 \times 10^{-4}$	$0.0787 \pm 19.14 \times 10^{-4}$	0.72

Based on the performance comparison in Table 8, the results for each model are as follows:

- GRU model: MAE = $0.0567 \pm 7.63 \times 10^{-4}$, RMSE = $0.0860 \pm 3.82 \times 10^{-4}$, R-squared = 0.69.
- LSTM model: MAE = $0.0939 \pm 4.10 \times 10^{-4}$, RMSE = $0.1260 \pm 2.72 \times 10^{-4}$, R-squared = 0.34.
- RNN model: MAE = $0.0930 \pm 46.44 \times 10^{-4}$, RMSE = $0.1197 \pm 34.38 \times 10^{-4}$, R-squared = 0.33.
- Transformer model: MAE = $0.0644 \pm 19.54 \times 10^{-4}$, RMSE = $0.0909 \pm 22.14 \times 10^{-4}$, R-squared = 0.66.
- BiGRU model: MAE = $0.0574 \pm 39.94 \times 10^{-4}$, RMSE = $0.0865 \pm 38.49 \times 10^{-4}$, R-squared = 0.69.
- Trans-BiGRU model: MAE = $0.0543 \pm 18.29 \times 10^{-4}$, RMSE = $0.0817 \pm 22.44 \times 10^{-4}$, R-squared = 0.67.
- For the proposed Trans-BiGRU-QA hybrid model, the results are: MAE = $0.0509 \pm 15.80 \times 10^{-4}$, RMSE = $0.0787 \pm 19.14 \times 10^{-4}$, and R-squared = 0.72.

Compared to baseline models such as GRU, LSTM, and RNN, the Trans-BiGRU-QA model demonstrates significant advantages across the evaluation metrics MAE, RMSE, and R-squared. The lowest MAE and RMSE indicate that the Trans-BiGRU-QA model excels in average deviation and performance on high-error samples, capturing the finer fluctuations in atmospheric mercury concentration more accurately. The lower MAE and RMSE suggest that this model exhibits more minor prediction deviations in typical and anomalous cases than the other baseline models. The highest R-squared value signifies that the Trans-BiGRU-QA model can better explain the total variance in concentration data, indicating a more precise capturing of overall trends. This is particularly important for

analyzing long-term dependencies in time series data. However, the computational time for this hybrid model may be longer than that of the other baseline models.

After multiple runs, the evaluation metrics (MAE, RMSE, and R^2) showed minimal variation, indicating that the model is not sensitive to the randomness of the training data and exhibits stable performance. Consistently outperforming baseline models across multiple runs suggests a low variability in model performance. This demonstrates that the Trans-BiGRU-QA model architecture and training process are relatively stable.

These results indicate that the Trans-BiGRU-QA hybrid model outperforms the other models in both performance and stability, as reflected by the lower MAE and RMSE values, the higher R-squared, and the more minor standard deviations. This shows that the proposed model is more accurate and consistent in predicting atmospheric mercury concentrations.

The performance of LSTM, RNN, and GRU in predicting the atmospheric mercury concentration (as measured by MAE, RMSE, and R^2) is inferior to that of the Trans-BiGRU-QA model, likely due to their relatively limited ability to capture long-term dependencies. In contrast, the Self-Attention mechanism in the Transformer component enables Trans-BiGRU-QA to capture distant dependencies within the data, providing the model an advantage in forecasting long-term trends. Additionally, LSTM and RNN are typically unidirectional, meaning they process sequences in a single direction, from past to future. In contrast, BiGRU is bidirectional and can consider the past and future contexts within a sequence.

Moreover, LSTM, RNN, and GRU lack an Attention mechanism, which limits their ability to selectively adjust the influence of different time points or features, leading to a poorer performance when handling data with high diversity and randomness. The Quick Attention (QA) mechanism in Trans-BiGRU-QA plays a crucial role by enabling the model to focus on features closely related to changes in atmospheric mercury concentration, reducing the impact of noise on prediction outcomes and thereby enhancing model accuracy and stability.

The Mean Absolute Error (MAE) measures the absolute difference between the actual and predicted values. At the same time, the Root Mean Square Error (RMSE) reflects the deviation between the actual results and predictions in real time [31]. According to the equations for MAE and RMSE, the smaller the values, the better the model's performance. The proposed Trans-BiGRU-QA hybrid model, which incorporates the Quick Attention mechanism, had experimental results where the MAE and RMSE were smaller than those of other single models and more accurate than those of the Trans-BiGRU hybrid model.

Including the Quick Attention mechanism enhances the model's prediction accuracy, as evidenced by the R-squared value, which reaches 0.72. This proves that the Trans-BiGRU-QA hybrid model is the best-performing model in this study, providing both a superior accuracy and stability compared to the other models.

LSTM and GRU have their strengths in time series modeling, particularly excelling at capturing short-term contextual dependencies. However, when long-term dependencies exist within a sequence, these models may need help to retain the influence of earlier information, presenting limitations in capturing distant dependencies. While LSTM and GRU can model contextual dependencies within sequences, more than a unidirectional structure is needed to consider past and future information as effectively as BiGRU. By utilizing bidirectional memory, BiGRU can more comprehensively capture both long-term and short-term dependencies within a sequence, which enhances coherence in mercury concentration prediction tasks.

In the Trans-BiGRU-QA model, the Quick Attention mechanism further enhances feature selection capability. It quickly focuses on information highly relevant to mercury concentration changes, suppressing noise and minimizing the influence of non-essential data on prediction outcomes. Quick Attention intensifies focus on specific features, increasing the model's attention to critical data points, leading to a superior performance across evaluation metrics such as MAE, RMSE, and R^2. This design improves prediction accuracy

and strengthens the model's ability to interpret long-term dependency data, demonstrating robust performance in atmospheric mercury concentration forecasting.

Figure 8 and Table 8 present the bar chart of MAE and RMSE for each model. To compare the predictive performance and accuracy of single models versus the hybrid model, this study plotted a bar chart displaying the MAE and RMSE values for a clear and intuitive comparison of each model's performance. The bar chart, from left to right, shows the MAE and RMSE data for the GRU, LSTM, RNN, Transformer, BiGRU, Trans-BiGRU, and Trans-BiGRU-QA models.

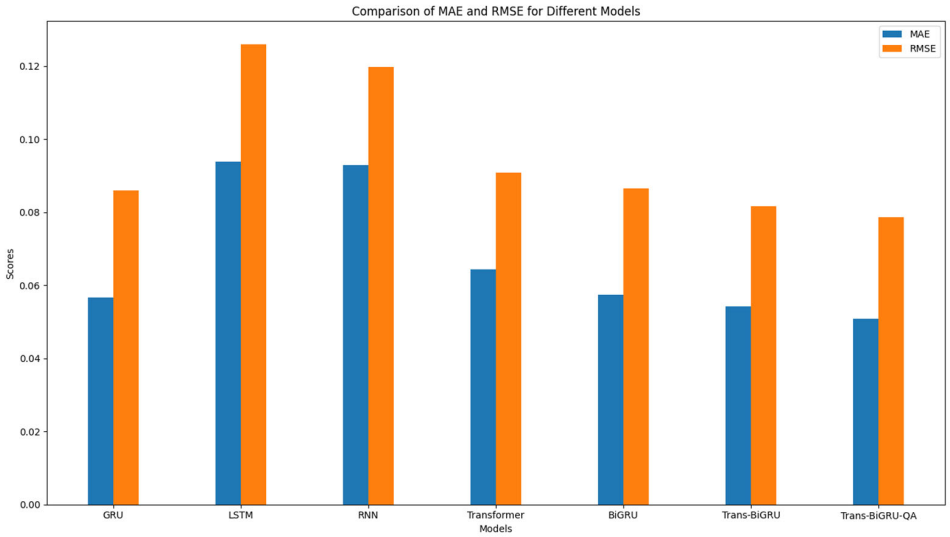
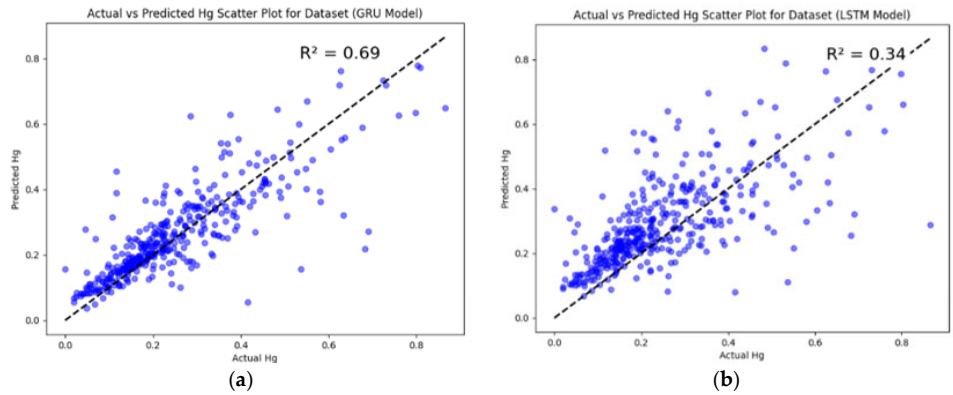


Figure 8. Bar chart of MAE and RMSE for each model.

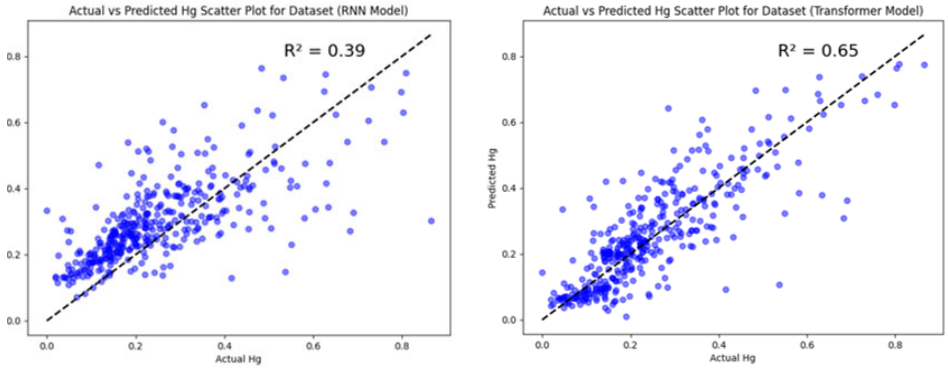
As Figure 7 depicts, the proposed Trans-BiGRU-QA hybrid model exhibits significantly lower MAE and RMSE values than the other single and hybrid models. This result further confirms the superior predictive accuracy of the Trans-BiGRU-QA hybrid model, establishing it as the best-performing model in the study.

Figures 9–12 show the scatter plots of the predicted and actual values for the GRU, LSTM, RNN, Transformer, BiGRU, Trans-BiGRU, and Trans-BiGRU-QA models, respectively. These scatter plots compare the predictive performance of each model and identify outliers, providing a visual representation of the relationship between predicted values and actual values. In the plots, the X-axis represents the actual values, and the Y-axis represents the predicted values. Each point represents a sample, comparing the actual value with the model's prediction. The closer the points are to the diagonal line, the more accurate the model's predictions are. Points that deviate significantly from the diagonal indicate less accurate predictions or potential outliers.



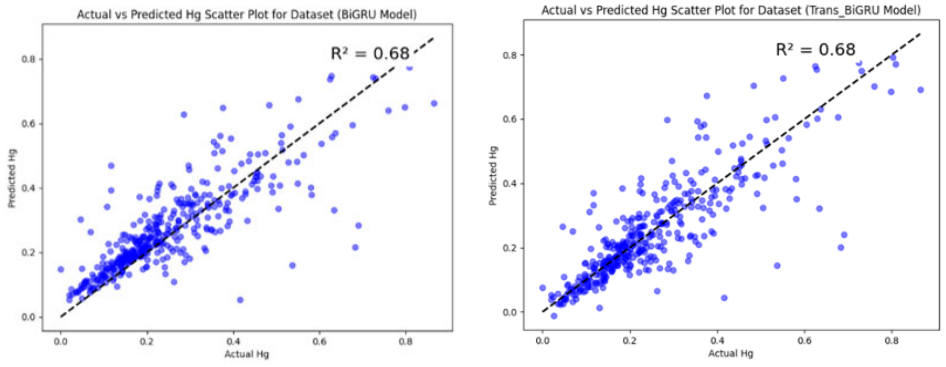
(**a**) (**b**)

Figure 9. GRU (**a**) and LSTM (**b**) model predictions, respectively.



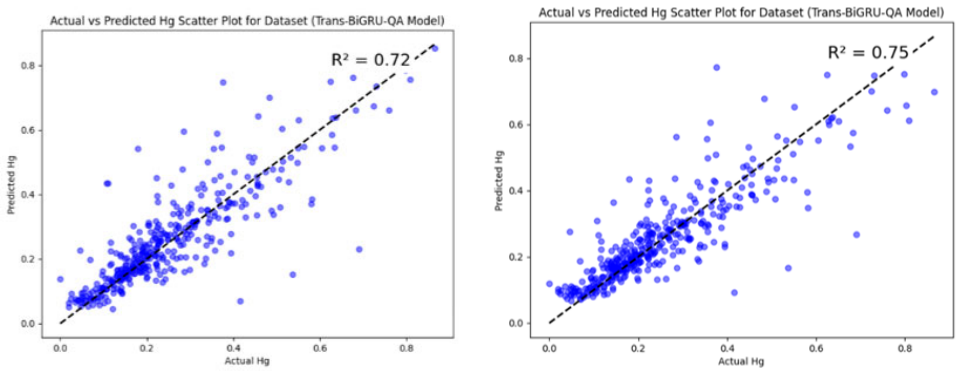
(**a**) (**b**)

Figure 10. RNN (**a**) and Transformer (**b**) model predictions, respectively.



(**a**) (**b**)

Figure 11. BiGRU (**a**) and Tran-BiGRU (**b**) model predictions, respectively.



(**a**) (**b**)

Figure 12. Trans-BiGRU-QA model predictions (**a**,**b**).

Figures 9–12 show that the GRU model demonstrates a relatively dense distribution of the actual and predicted values, indicating a good predictive accuracy. In contrast, the LSTM and RNN models show more dispersed distributions, reflecting a poorer performance in matching the actual values with the predictions.

The scatter points for the Transformer, BiGRU, and Trans-BiGRU models are more closely grouped and show a more stable predictive performance compared to the LSTM and RNN models. These models exhibit a more consistent alignment between the predicted and actual values, though there is still room for improvement in their predictive precision.

As shown in Figure 12a, the proposed Trans-BiGRU-QA hybrid model demonstrates a points distribution that aligns closely with the diagonal line, indicating a higher predictive accuracy. The model achieved an R-squared value of 0.72, confirming its superior performance compared to the other models.

This study employed grid search to further optimize the hyperparameters of the Trans-BiGRU-QA hybrid model during the training process. Figure 12b displays the scatter plot of data predictions after applying the optimized hyperparameters using grid search. Figure 12b shows that, after hyperparameter optimization, the model's predictive accuracy improved further, with the R-squared value increasing to 0.75. This demonstrates an even stronger model performance, highlighting the effectiveness of grid search in fine-tuning the Trans-BiGRU-QA model for optimal results.

Figure 13 shows the line chart of the actual vs. predicted values for the Trans-BiGRU-QA model.

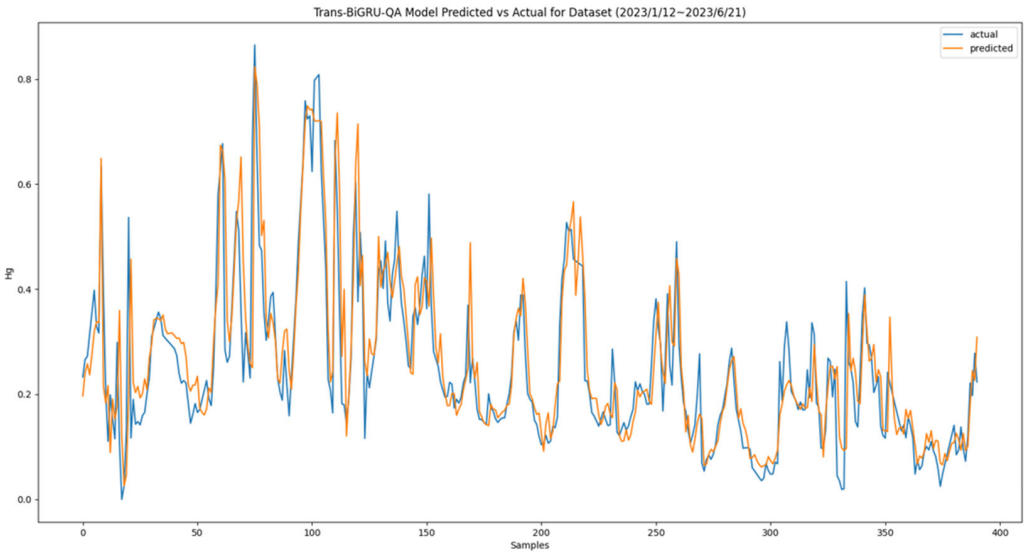


Figure 13. Trans-BiGRU-QA model prediction and actual data line chart.

5.3.2. Statistical Analysis of the Performance of the Proposed Model

Table 9 presents the Analysis of Variance (ANOVA) results for the models. According to the table: The sum_sq under Model represents the total variance the Model explains, i.e., the Model's sum of squares. The sum_sq under Residual represents the residual sum of squares, indicating the variance not explained by the Model. Df refers to the degrees of freedom, where the Model's df is 6, indicating the presence of 7 model groups, while the Residual's df is 63.

Table 9. Analysis of variance (ANOVA) for each model.

	Sum-sq	**DF**	**F**	**PR (>F)**
Model	0.020227	6.0	442.999226	1.590326×10^{-49}
Residual	0.000479	63.0	NaN	NaN

The F-value is a statistic that tests whether the variance between groups is significantly greater than within groups. The F-value for this Model is 442.999226, indicating that the variance between groups is significantly more significant than within groups.

The *p*-value (PR (>F)) represents the probability of observing an extreme F-value if the null hypothesis is true. In this case, the *p*-value is 1.590326×10^{-49}, much smaller than 0.05, suggesting that the null hypothesis can be rejected. This means there is a statistically significant difference between the model groups.

The ANOVA results demonstrate that the differences between the models are statistically significant, given the very low *p*-value. Consequently, it is appropriate to proceed with Tukey's HSD Test to compare the specific differences between the models.

Table 10 presents the results of Tukey's HSD Test for the models. The table summarizes the Mean Absolute Error (MAE) after running each model ten times. The models compared

included GRU, LSTM, RNN, Transformer, BiGRU, Trans-BiGRU, and the proposed Trans-BiGRU-QA model. The comparisons between two machine learning models are shown in the Group1 and Group2 columns.

Table 10. Summary of Tukey's HSD test results for each model.

Group 1	Group 2	Mean Diff.	P-Adj	Lower	Upper	Reject
BiGRU	GRU	−0.00165	0.9	−0.00518	0.00188	False
BiGRU	LSTM	0.03761	0.001	0.03407	0.04114	True
BiGRU	RNN	0.03616	0.001	0.03263	0.03969	True
BiGRU	Trans-BiGRU	0.00223	0.9	−0.00131	0.00576	False
BiGRU	Trans-BiGRU-QA	−0.00587	0.001	−0.00941	−0.00234	True
BiGRU	Transformer	−0.00473	0.002	−0.00826	−0.00119	True
GRU	LSTM	0.03926	0.001	0.03572	0.04279	True
GRU	RNN	0.03781	0.001	0.03427	0.04134	True
GRU	Trans-BiGRU	0.00388	0.622	−0.00017	0.00794	False
GRU	Trans-BiGRU-QA	−0.00421	0.013	−0.00825	−0.00016	True
GRU	Transformer	−0.00308	0.239	−0.00713	0.00098	False
LSTM	RNN	−0.00145	0.9	−0.00548	0.00258	False
LSTM	Trans-BiGRU	−0.03538	0.001	−0.03942	−0.03135	True
LSTM	Trans-BiGRU-QA	−0.04346	0.001	−0.04749	−0.03942	True
LSTM	Transformer	−0.04233	0.001	−0.04637	−0.03830	True
RNN	Trans-BiGRU	−0.03393	0.001	−0.03796	−0.02989	True
RNN	Trans-BiGRU-QA	−0.04201	0.001	−0.04605	−0.03798	True
RNN	Transformer	−0.04088	0.001	−0.04491	−0.03684	True
Trans-BiGRU	Trans-BiGRU-QA	−0.00808	0.001	−0.1211	−0.00404	True
Trans-BiGRU	Transformer	−0.00695	0.001	−0.01098	−0.0029	True
Trans-BiGRU-QA	Transformer	0.00113	0.9	−0.00290	0.00517	False

Mean diff. represents the mean difference between the two models being compared. P-adj is the adjusted *p*-value used to test whether the mean difference is statistically significant. Lower and Upper represent the lower and upper limits of the 95% confidence interval. Reject indicates whether the null hypothesis (that the means are equal) is rejected. If the Reject column shows "True", the mean difference between the two models is statistically significant.

According to the results of Tukey's Test, when compared to the other baseline models, the Trans-BiGRU-QA model consistently demonstrates a significant performance advantage. The model has a significantly lower MAE in all comparisons, confirming its predictive performance is noticeably better than the other models. This further validates the superiority of the Trans-BiGRU-QA hybrid model in this study.

5.3.3. Robustness Analysis

Model robustness refers to a model's ability to maintain reliable performance and consistency across different datasets, environments, or training conditions [48]. A robust model can withstand small data fluctuations or variations in testing sets while maintaining high predictive accuracy. This study conducted a model robustness analysis to ensure that machine learning models perform stably in practical applications. By studying and improving model robustness, the credibility and reliability of machine learning systems can be enhanced, promoting more widespread research into trustworthy machine learning. A robust model can effectively handle data variability, reduce the risk of overfitting, and increase model trustworthiness [48]. Robust models possess a strong generalization ability, allowing them to adapt to data changes while maintaining a stable performance and increasing confidence in the model's results.

In this study, the standard deviation (STD) was used to assess the robustness of each model. In deep learning models, performance stability is typically evaluated by calculating multiple experimental results' mean and standard deviation. A more minor standard deviation indicates excellent stability in the model's performance. Therefore, the standard deviation is a crucial measure to evaluate the robustness of the models.

Table 11 compares the robustness of each machine learning model, showing the overall average Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) along

with their corresponding standard deviations. Lower MAE and RMSE values indicate smaller prediction errors, while the standard deviation reflects the variability in performance across multiple experiments. A smaller standard deviation means the model's performance remains stable across different training and testing conditions, indicating a stronger robustness.

Table 11. Comparison of robustness across models.

	MAE	RMSE
GRU	$0.0567 \pm 7.63 \times 10^{-4}$	$0.0860 \pm 3.82 \times 10^{-4}$
LSTM	$0.0939 \pm 4.10 \times 10^{-4}$	$0.1260 \pm 2.72 \times 10^{-4}$
RNN	$0.0930 \pm 46.44 \times 10^{-4}$	$0.1197 \pm 34.38 \times 10^{-4}$
Transformer	$0.0644 \pm 19.54 \times 10^{-4}$	$0.0909 \pm 22.14 \times 10^{-4}$
BiGRU	$0.0574 \pm 39.94 \times 10^{-4}$	$0.0865 \pm 38.49 \times 10^{-4}$
Trans-BiGRU	$0.0543 \pm 18.29 \times 10^{-4}$	$0.0817 \pm 22.44 \times 10^{-4}$
Trans-BiGRU-QA	$0.0509 \pm 15.80 \times 10^{-4}$	$0.0787 \pm 19.14 \times 10^{-4}$

As shown in Table 11, the proposed Trans-BiGRU-QA model achieves the best results in both MAE and RMSE, followed by the Trans-BiGRU and GRU models, demonstrating a good robustness and low error rates. On the other hand, the LSTM and RNN models have higher standard deviations and more significant errors, indicating poor robustness. The Transformer model shows moderate error levels, but its more significant standard deviation suggests some volatility in prediction performance. The BiGRU model achieves low error rates, but its higher standard deviation implies that its robustness may fluctuate in some instances.

From this robustness analysis, it is clear that the Trans-BiGRU-QA model demonstrates the best robustness and accuracy, with the GRU model being a close second. The Trans-BiGRU-QA and Trans-BiGRU models are recommended for those prioritizing low error and high stability.

5.4. Ablation Experiment

The purpose of ablation experiments is to investigate the contribution of various structural configurations to the performance of a new model. Different variant model structures are typically constructed and compared using the same training data when evaluating different combinations in deep learning models and their impact on overall performance [49]. Ablation experiments allow for assessing how specific features or components of a model contribute to the overall performance. This process typically involves modifying the model or progressively removing specific components, followed by observing the resulting changes in performance. This helps to understand better how each model component influences the outcome.

Table 12 presents the comparison results from the ablation experiments for the hybrid model in this study. According to Table 12, this study follows Sheikholeslami et al.'s [49] ablation experiment methodology, using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared to evaluate the performance of the Trans-BiGRU-QA hybrid model and the contribution of each of its components.

Table 12. Comparison of hybrid model experiments.

	MAE	RMSE	R^2	Equation
Trans-BiGRU-QA	0.0509	0.0787	0.72	$C_m(M)$
Trans-BiGRU	0.0543	0.0817	0.67	$C_m(M, \{C\})$
BiGRU-QA	0.0580	0.0871	0.68	$C_m(M, \{A\})$
Trans-QA	0.0578	0.0840	0.71	$C_m(M, \{B\})$
BiGRU	0.0574	0.0865	0.69	$C_m(M, \{A, C\})$

The equations for the ablation experiments, Equations (13) and (14), are provided as follows:

$$C_m(M) = C_m(A,\ B,\ C) \tag{13}$$

$$C_m(M,\ \{B\}) \tag{14}$$

As shown in Equation (13), the total sum of the Trans-BiGRU-QA model is defined as C_m, with models A, B, and C defined as three different model combinations for cross-testing. Model A refers to the Transformer model, model B refers to the BiGRU model, and model C refers to the Quick Attention model. Model M is the Trans-BiGRU-QA model [49]. Since this study used a model that integrates all three models, the equation for all combined models was defined as $C_m(M) = C_m(A,\ B,\ C)$, as shown in Equation (14).

Since this study removed model B (BiGRU) for an ablation test, the equation during the testing phase was $C_m(M,\ \{B\})$ [46]. Each testing phase's MAE, RMSE, and R-squared were executed ten times, and the average was taken as the final test value.

This study's ablation research wasdivided into five testing phases to assess whether the novel Trans-BiGRU-QA fusion model effectively improves training performance. The first phase tested the predictive structure of the Trans-BiGRU-QA fusion model, referred to as $C_m(M)$. The second phase removed the Quick Attention model and generated the second phase $C_m(M, \{C\})$ (Trans-BiGRU). The third phase removed the Transformer model to generate $C_m(M, \{A\})$ (BiGRU-QA). The fourth phase removed the BiGRU model to generate the fourth phase $C_m(M, \{B\})$ (Trans-QA). Finally, both the Transformer and Quick Attention models were removed for an individual test of the BiGRU model, generating the fifth phase $C_m(M, \{A, C\})$ (BiGRU).The results indicate that the proposed Trans-BiGRU-QA fusion model is the best, with an MAE of 0.0509, an RMSE of 0.0787, and an R-squared of 0.72.

5.5. SHAP

SHAP (SHapley Additive exPlanations) is a tool (Python library) used to explain the predictions of machine learning models. Its primary goal is to help this study understand the factors influencing the prediction process and the decision-making of the model and to explain the outcomes of machine learning model predictions [50]. SHAP's application mainly covers four areas: explaining model predictions, evaluating feature importance, improving the model, and enhancing model trust and transparency.

Explaining model predictions: SHAP quantifies the contribution of each feature to the model's prediction, helping to interpret the logic behind the model's predictions. Feature importance evaluation: SHAP assesses the average impact of each feature on the dataset's predictions, allowing this study to identify the most critical features for overall model performance. Model improvement: Adjusting the model structure or features based on SHAP values can enhance model stability or prediction accuracy. Increasing trust and transparency: SHAP boosts trust in the model's decision-making process by making the typically black-box nature of machine learning models more understandable. SHAP makes the model's prediction process more interpretable and transparent, clarifying the internal operations that are otherwise difficult to comprehend [50].

As an essential tool for analyzing machine learning model predictions, SHAP can quantify the contribution of each feature to the prediction results. In a SHAP plot, the X-axis represents the magnitude of SHAP values, which indicates the impact of each feature on the model output—the larger the value, the more significant the impact. The Y-axis lists all the features: TGM, CO_2, Temp, $PM_{2.5}$, and RH. Each point's color represents the magnitude of the feature value, ranging from blue (low feature value) to red (high feature value).

Figure 14 displays the SHAP plot for the input variables. According to Figure 12b:

- TGM has the greatest influence on the model output, followed by CO_2, Temperature (temperature), $PM_{2.5}$, and RH (Relative Humidity).
- For TGM, CO_2, and Temp, higher feature values have a positive impact on the model output, driving the predicted results in an upward direction.

- In contrast, $PM_{2.5}$ and RH have relatively minor influences, and the shading indicates a low impact on the prediction output, suggesting that variations in these features contribute less to the model output than TGM and other primary features.

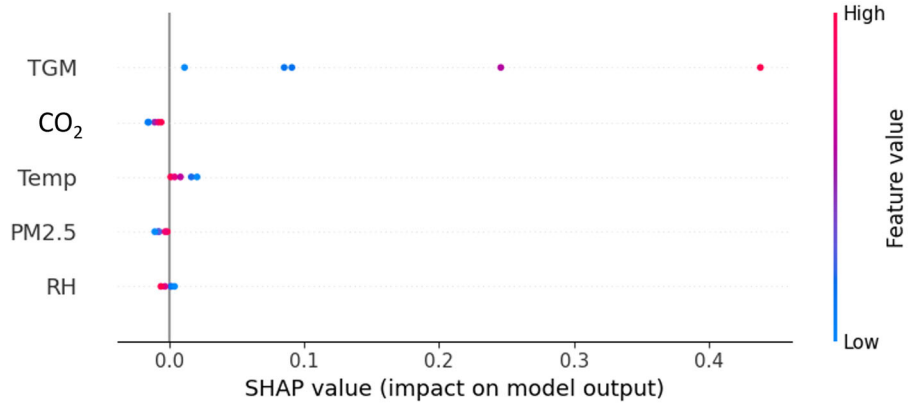


Figure 14. SHAP plot of input variables.

5.6. Discussion

5.6.1. Discussion of Ablation Experiments

Based on the results of the ablation experiments, the proposed Trans-BiGRU-QA hybrid model demonstrated the best performance, with an MAE of 0.0509, RMSE of 0.0787, and an R-squared of 0.72. The ablation experiments confirmed that integrating Transformer, BiGRU, and Quick Attention within the proposed architecture significantly enhances the model's predictive performance.

During the ablation experiments, different model components were gradually removed, allowing the observation of changes in model performance. This process provided insights into the contribution of each component to the overall model performance. The results indicate that integrating these structural elements effectively improves the model's predictive capabilities. Furthermore, removing any single part resulted in a decline in model performance, which further validated each component's critical role in enhancing the model's overall effectiveness. This confirms that the combination of Transformer, BiGRU, and Quick Attention is essential for achieving optimal predictive accuracy in the hybrid model.

When the Quick Attention mechanism was removed, the model reverted to Trans-BiGRU, resulting in an MAE of 0.0543, RMSE of 0.0817, and decreased the R-squared value to 0.67. This decline in performance indicates that the Quick Attention mechanism plays a significant role in enhancing the model's capabilities. The Quick Attention mechanism helps the model extract key features more effectively, thereby improving the prediction accuracy. Therefore, the contribution of this mechanism within the overall hybrid model is substantial and cannot be overlooked.

When the Transformer model was removed, the model became BiGRU-QA, resulting in an MAE of 0.0580, RMSE of 0.0871, and a decrease in the R-squared value to 0.68. This further decline in performance demonstrates that the Transformer model plays a crucial role in improving overall predictive performance. The Transformer model effectively captures long-distance dependencies and key features in the data, which has a significant impact on the model's prediction accuracy. Thus, its importance within the model structure is substantial and cannot be overlooked.

When the BiGRU component was removed, the model became Trans-QA, resulting in an MAE of 0.0578, RMSE of 0.0840, and an R-squared value of 0.71. This indicates a decline in performance, though the impact is less severe compared to the removal of the Transformer model or the Quick Attention mechanism. This suggests that BiGRU also contributes to improving the model's performance. The BiGRU model can extract forward and backward information from sequential data, thus enhancing prediction accuracy.

However, its contribution is slightly less significant than that of the Transformer model and the Quick Attention mechanism in the overall model structure.

When testing with a single BiGRU model, the results show an MAE of 0.0574, an RMSE of 0.0865, and an R-squared value of 0.69. The model's performance significantly declined, indicating that a single model cannot achieve the predictive accuracy of the hybrid model. The hybrid model effectively combines the strengths of different models, significantly improving prediction accuracy and stability. In contrast, a single model's performance is more limited, as it cannot fully exploit the complex features in the data.

Through ablation experiments, this study demonstrated the impact of different combinations on the overall performance of the hybrid model. The results indicate that integrating the three models—Transformer, BiGRU, and Quick Attention—within the Trans-BiGRU-QA hybrid model significantly enhances the model's predictive capabilities. The experimental findings further confirm that the proposed hybrid model offers a notable advantage in data prediction performance, effectively improving prediction accuracy and stability.

5.6.2. Robustness

This study evaluated each model's stability by calculating the mean and standard deviation across multiple experiments. The standard deviation reflects the variability in model performance; a smaller standard deviation indicates that the model performs more consistently across different training and testing scenarios. The experimental results show that the LSTM model demonstrated the best stability, followed by the GRU model.

The proposed Trans-BiGRU-QA model exhibited a significant advantage in predictive performance, with an MAE of 0.0509 (standard deviation: $\pm 15.80 \times 10^{-4}$) and an RMSE of 0.0787 (standard deviation: $\pm 19.14 \times 10^{-4}$). This model achieved the lowest prediction error in the experiments, highlighting its superior accuracy.

In contrast, the LSTM and RNN models showed weaker predictive performance. Although their standard deviations were relatively small, indicating a degree of stability, their overall effectiveness was not ideal. For instance, the LSTM model had an MAE of 0.0939 (standard deviation: $\pm 4.10 \times 10^{-4}$) and an RMSE of 0.1260 (standard deviation: $\pm 2.72 \times 10^{-4}$). This suggests that, while the LSTM and RNN models were stable across multiple experiments, they suffered from more significant prediction errors and lacked the predictive accuracy required for more precise results.

The Transformer model exhibited moderate error levels, but its more significant standard deviation indicates that its predictions may experience more significant fluctuations under certain conditions. While the BiGRU model had a relatively low error, it showed a higher standard deviation, suggesting that its stability might vary under some conditions.

The robustness experiments concluded that the proposed Trans-BiGRU-QA model, as well as the Trans-BiGRU model, both performed exceptionally well in terms of stability and low error rates. The analysis of the different models' robustness demonstrated that integrating multiple models can lead to a better performance and stability compared to single models. The Trans-BiGRU-QA hybrid model developed in this study is a significant contribution, effectively managing data variability, reducing overfitting, and enhancing the reliability of the prediction results.

5.6.3. Limitations

This study's total number of data points was 1245, obtained after removing outliers and handling missing values. During the data collection process for Vietnam's air quality data, factors such as equipment malfunctions, power outages, or human errors led to missing or abnormal values. The final dataset used for analysis underwent a thorough cleaning and processing.

However, models are prone to overfitting when the training dataset is limited. This means that, while a model may perform well on the training data, it may still perform poorly on the test set. The primary reason is that the model may need to learn the underlying

patterns in the data effectively instead of memorizing specific features present in the training set.

The limited data quantity can restrict the prediction accuracy of the models, potentially causing the performance of various models to be quite similar. This challenge underscores the urgent need for larger and more representative datasets for training. With such datasets, models can better generalize to unseen data and achieve improved predictive accuracy, addressing this issue and improving the overall performance of our models. Atmospheric mercury monitoring stations are relatively scarce, and currently, most can only collect relevant concentration data within limited areas. This results in spatial and temporal data sparsity, making it challenging to comprehensively reflect mercury concentration variations across more significant regions. Additionally, environmental conditions, equipment aging, and maintenance quality can affect atmospheric mercury measurement equipment, impacting the accuracy and consistency of the data. If the data contains a high noise level or outliers, the accuracy of model predictions may be compromised.

5.6.4. Integration and Application

The outstanding performance of the Trans-BiGRU-QA model in predicting atmospheric mercury concentrations demonstrates its feasibility for future implementation in early warning systems or environmental monitoring networks. The model must handle real-time data streams from various monitoring points in an early warning system. Trans-BiGRU-QA can be deployed on edge servers, directly receiving updated data from the source and performing real-time predictions. Based on the model's predictions, specific mercury concentration thresholds can be set to trigger alerts. When the model's predicted values exceed these thresholds, the system automatically initiates warning mechanisms and activates response protocols.

Trans-BiGRU-QA can be configured as a distributed model deployed across different monitoring stations for single-point predictions and integrated analysis. As predictions from multiple monitoring points are aggregated, the system can identify spatial trends in mercury concentration, aiding in tracking the sources and spread pathways of mercury pollution.

In the future, the Trans-BiGRU-QA model could incorporate a broader range of meteorological variables, such as atmospheric pressure, rainfall, and hours of sunlight. These factors directly influence the mercury's transport, transformation, and deposition, enabling the model to capture the impact of environmental conditions on concentration levels accurately. Additionally, an ensemble model could be considered to strengthen the prediction framework. By combining models, the bias of a single model can be reduced, leading to more robust predictions. The final prediction could be determined through methods like weighted averaging or voting.

5.6.5. Application to an Additional Dataset

To verify the applicability of the proposed Trans-BiGRU-QA model used in this study, this study further used another EU27&UK gas dataset for verification. The verification dataset was taken from the Zenodo open dataset. These data come from the natural gas data of the entire 27 EU countries and the United Kingdom. The characteristics of natural gas data include the country's total natural gas emissions and some factors that affect the natural gas supply, transmission, consumption structure and consumption change drivers, etc. The data source is (https://zenodo.org/records/11175364, accessed on 31 October 2024), the number of data items is 67,986, and the data period is from 1 January 2016 to 30 April 2024. This study used its input variables to forecast the volume of natural gas heating for homes for the next day. Table 13 is a summary table of each data attribute.

Table 13. Summary of data attributes in the gas dataset.

Variable	Name	Description	Data Type
$XG_0(t)$	TOTAL	All natural gas supplies	integer
$XG_1(t)$	RU	Russian natural gas supply	float
$XG_2(t)$	LNG	LNG supply	float
$XG_3(t)$	PRO	Amount of natural gas produced	float
$XG_4(t)$	AZ	Azerbaijan's natural gas supplies	float
$XG_5(t)$	DZ	Algeria's natural gas supplies	integer
$XG_6(t)$	NO	Norwegian natural gas supplies	float
$XG_7(t)$	LY	Libyan natural gas supply	float
$XG_8(t)$	TR	Natural gas supplies in the Netherlands	float
$XG_9(t)$	RU_from_storage	Natural gas supplies from Russian storage	float
$XG_{10}(t)$	LNG_from_storage	Natural gas supply from LNG storage	float
$XG_{11}(t)$	PRO_from_storage	Natural gas supply from production storage	integer
$XG_{12}(t)$	AZ_from_storage	Natural gas supplies from Azerbaijan storage	integer
$XG_{13}(t)$	DZ_from_storage	Natural gas supplies from Algerian storage	integer
$XG_{14}(t)$	NO_from_storage	Natural gas supplies from Norwegian storage	float
$XG_{15}(t)$	RS_from_storage	Natural gas supplies from Russian storage	float
$XG_{16}(t)$	LY_from_storage'	Natural gas supplies from Libyan storage	float
$XG_{17}(t)$	TR_from_storage	Natural gas supplies from Turkish storage	float
$XG_{18}(t)$	house_heating	Natural gas heating for homes	float
$YG_1(t+1)$	house_heating	Natural gas heating for homes (next day)	float

The research results of all models are shown in Table 14. The Trans-BiGRU-QA model proposed in this study shows the best performance in terms of evaluation indices. Although another Trans-BiGRU model also has good performance, its performance is still not as good as the Trans-BiGRU-QA model proposed in this study.

Table 14. Comparison of hybrid models in the gas dataset.

	MAE	RMSE	R^2
Trans-BiGRU-QA	0.0614	0.0888	0.69
Trans-BiGRU	0.0621	0.0904	0.65
BiGRU-QA	0.0682	0.1003	0.66
Trans-QA	0.0627	0.0945	0.68
BiGRU	0.0754	0.1005	0.67

These results show that, compared to other basic or hybrid models, the Trans-BiGRU-QA model proposed in this study still has the best performance when processing the EU27&UK gas dataset, further supporting the applicability of the Trans-BiGRU-QA model method proposed in this study. This also shows the significant role of the Trans-BiGRU-QA model proposed in this study in predicting performance and is worthy of further exploration and application in future research and practice.

6. Conclusions

This study proposed an advanced Trans-BiGRU-QA hybrid model for atmospheric mercury (TGM) forecasting, comparing its performance against other machine learning models, including GRU, LSTM, RNN, Transformer, BiGRU, and Trans-BiGRU. The evaluation metrics included the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared, providing a comprehensive and rigorous assessment of each model's strengths and weaknesses. The study utilized a dataset of air variables collected in Vietnam,

incorporating five key features—atmospheric mercury concentration (TGM), temperature (Temp), relative humidity (RH), $PM_{2.5}$, and carbon dioxide (CO_2)—to predict hourly atmospheric mercury levels.

In air pollution research, machine learning models are commonly used to integrate air pollution data to assess environmental impacts. This study demonstrates that the hybrid model can significantly improve prediction accuracy through the performance evaluation, statistical analysis, and robustness analysis of the models. The Self-Attention mechanism in the Transformer model enables it to process all positions in input sequences simultaneously, effectively capturing long-term dependencies. When combined with BiGRU, which extracts deep features through forward and backward time series processing, and the Quick Attention mechanism, the model focuses more precisely on critical features in the data.

The proposed Trans-BiGRU-QA hybrid model showed exceptional performance in forecasting atmospheric mercury levels. Leveraging feature engineering and sliding window techniques, the model accurately predicts future trends in atmospheric mercury. When compared to baseline models like GRU, LSTM, RNN, Transformer, BiGRU, and Trans-BiGRU, the results indicate that Trans-BiGRU-QA is the best-performing hybrid model, with superior MAE, RMSE, and R-squared values. This hybrid model offers technical support for real-time predictions, advancing the application of deep learning techniques in air pollution forecasting and providing valuable tools for meteorologists and air quality experts.

Author Contributions: Conceptualization, D.-H.S.; formal analysis, B.-H.W.; investigation, D.-H.S., F.-I.C., and T.-W.W.; methodology, D.-H.S. and B.-H.W.; project administration, D.-H.S.; Resources, D.-H.S. and M.-H.S.; software, F.-I.C. and B.-H.W.; supervision, D.-H.S.; validation, T.-W.W. and M.-H.S.; visualization, M.-H.S.; writing—original draft, T.-W.W. and B.-H.W.; writing—review and editing, F.-I.C. and M.-H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are not publicly available due to privacy or ethical restrictions.

Acknowledgments: The authors thank Ly Sy Phu Nguyen, Vietnam National University, Linh Trung Ward, Thu Duc District, Ho Chi Minh City 700000, Vietnam, for the dataset provided. Atmospheric mercury concentration monitoring is primarily conducted in Ho Chi Minh City.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yan, R.; Liao, J.; Yang, J.; Sun, W.; Nong, M.; Li, F. Multi-hour and multi-site air quality index forecasting in Beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. *Expert Syst. Appl.* **2021**, *169*, 114513. [CrossRef]
2. Samad, A.; Garuda, S.; Vogt, U.; Yang, B. Air pollution prediction using machine learning techniques—An approach to replace existing monitoring stations with virtual monitoring stations. *Atmos. Environ.* **2023**, *310*, 119987. [CrossRef]
3. Yuan, C.-S.; Jhang, Y.-M.; Ie, I.-R.; Lee, C.-E.; Fang, G.-C.; Luo, J. Exploratory Investigation on Spatiotemporal Variation and Source Identification of Atmospheric Speciated Mercury Surrounding the Taiwan Strait. *Atmos. Pollut. Res.* **2021**, *12*, 54–64. [CrossRef]
4. Skalny, A.V.; Aschner, M.; Sekacheva, M.I.; Santamaria, A.; Barbosa, F.; Ferrer, B.; Aaseth, J.; Paoliello, M.M.; Rocha, J.B.; Tinkov, A.A. Mercury and cancer: Where are we now after two decades of research? *Food Chem. Toxicol.* **2022**, *164*, 113001. [CrossRef]
5. Wu, Q.; Tang, Y.; Wang, S.; Li, L.; Deng, K.; Tang, G.; Liu, K.; Ding, D.; Zhang, H. Developing a statistical model to explain the observed decline of atmospheric mercury. *Atmos. Environ.* **2020**, *243*, 117868. [CrossRef]
6. Wang, C.; Wang, Z.; Zhang, X. Two years measurement of speciated atmospheric mercury in a typical area of the north coast of China: Sources, temporal variations, and influence of regional and long-range transport. *Atmos. Environ.* **2020**, *228*, 117235. [CrossRef]
7. Nguyen, L.S.P.; Pham, T.D.H.; Truong, M.T.; Tran, A.N. Characteristics of total gaseous mercury at a tropical megacity in Vietnam and influence of tropical cyclones. *Atmos. Pollut. Res.* **2023**, *14*, 101813. [CrossRef]
8. Pang, Q.; Gu, J.; Wang, H.; Zhang, Y. Global health impact of atmospheric mercury emissions from artisanal and small-scale gold mining. *iScience* **2022**, *25*, 104881. [CrossRef] [PubMed]

9. Wang, C.; Wang, Z.; Zhang, Y.; Zhang, X. Sustained high atmospheric Hg level in Beijing during wet seasons suggests that anthropogenic pollution is continuing: Identification of potential sources. *Environ. Res.* **2022**, *214*, 113814. [CrossRef] [PubMed]
10. Xu, N.; Li, L.; Dong, H.; Huang, F. Prediction of Air Quality in the Beijing-Tianjin-Hebei Region Based on LSTM Model. *Acad. J. Comput. Inf. Sci.* **2023**, *6*, 113–118. [CrossRef]
11. Wen, C.; Lin, X.; Ying, Y.; Ma, Y.; Yu, H.; Li, X.; Yan, J. Dioxin emission prediction from a full-scale municipal solid waste incinerator: Deep learning model in time-series input. *Waste Manag.* **2023**, *170*, 93–102. [CrossRef]
12. Sarkar, N.; Gupta, R.; Keserwani, P.K.; Govil, M.C. Air Quality Index prediction using an effective hybrid deep learning model. *Environ. Pollut.* **2022**, *315*, 120404. [CrossRef]
13. Wu, C.-L.; He, H.-D.; Song, R.-F.; Zhu, X.-H.; Peng, Z.-R.; Fu, Q.-Y.; Pan, J. A hybrid deep learning model for regional O_3 and NO_2 concentrations prediction based on spatiotemporal dependencies in air quality monitoring network. *Environ. Pollut.* **2023**, *320*, 121075. [CrossRef]
14. Wu, Z.; Zeng, Y. Air pollution distribution under climate change: Application of geographical artificial intelligence technology. *Civ. Eng. Water Conserv.* **2023**, *50*, 16–23.
15. Wang, S. Air pollution warning for special education. *Taiwan Educ. Rev. Mon.* **2018**, *7*, 121–124. Available online: https://www.airitilibrary.com/Article/Detail?DocID=P20130114001-201807-201807160019-201807160019-121-124 (accessed on 11 November 2024).
16. Fu, X.W.; Zhang, H.; Yu, B.; Wang, X.; Lin, C.-J.; Feng, X.B. Observations of atmospheric mercury in China: A critical review. *Atmos. Meas. Tech.* **2015**, *15*, 9455–9476. [CrossRef]
17. Luo, Q.; Ren, Y.; Sun, Z.; Li, Y.; Li, B.; Yang, S.; Zhang, W.; Wania, F.; Hu, Y.; Cheng, H. Characterization of atmospheric mercury from mer-cury-added product manufacturing using passive air samplers. *Environ. Pollut.* **2023**, *337*, 122519. [CrossRef]
18. Kalyan, K.S.; Rajasekharan, A.; Sangeetha, S. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv* **2021**, arXiv:2108.05542.
19. Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 87–110. [CrossRef]
20. Chen, M.; Tan, X.; Ren, Y.; Xu, J.; Sun, H.; Zhao, S.; Qin, T.; Liu, T.-Y. Multispeech: Multi-speaker text to speech with transformer. *arXiv* **2020**, arXiv:2006.04664.
21. Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; Sun, L. Transformers in time series: A survey. *arXiv* **2022**, arXiv:2202.07125.
22. Gao, Y.; Miyata, S.; Matsunami, Y.; Akashi, Y. Spatio-temporal interpretable neural network for solar irradiation prediction using transformer. *Energy Build.* **2023**, *297*, 113461. [CrossRef]
23. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017.
24. Mao, X.; Ren, N.; Dai, P.; Jin, J.; Wang, B.; Kang, R.; Li, D. A variable weight combination prediction model for climate in a greenhouse based on BiGRU-Attention and LightGBM. *Comput. Electron. Agric.* **2024**, *219*, 108818. [CrossRef]
25. Zhao, L.; Li, Z.; Qu, L.; Zhang, J.; Teng, B. A hybrid VMD-LSTM/GRU model to predict non-stationary and irregular waves on the east coast of China. *Ocean Eng.* **2023**, *276*, 114136. [CrossRef]
26. Busari, G.A.; Lim, D.H. Crude oil price prediction: A comparison between AdaBoost-LSTM and AdaBoost-GRU for improving forecasting performance. *Comput. Chem. Eng.* **2021**, *155*, 107513. [CrossRef]
27. Tang, J.; Hou, H.J.; Chen, H.G.; Wang, S.J.; Sheng, G.H.; Jiang, C.X. Concentration prediction method based on Seq2Seg network improved by BI-GRU for dissolved gas in transformer oil. *Electr. Power Autom. Equip.* **2022**, *42*, 196–202.
28. Ji, Y.; Huang, Y.; Zeng, J.; Ren, L.; Chen, Y. A physical–data-driven combined strategy for load identification of tire type rail transit vehicle. *Reliab. Eng. Syst. Saf.* **2025**, *253*, 110493. [CrossRef]
29. Wazir, S.; Fraz, M.M. HistoSeg: Quick attention with multi-loss function for multi-structure segmentation in digital histology images. In Proceedings of the 12th International Conference on Pattern Recognition Systems (ICPRS), Saint-Etienne, France, 7–10 June 2022; pp. 1–7.
30. Soydaner, D. Attention mechanism in neural networks: Where it comes and where it goes. *Neural Comput. Appl.* **2022**, *34*, 13371–13385. [CrossRef]
31. Huang, Q.; Cui, Z. Study on prediction of ocean effective wave height based on hybrid artificial intelligence model. *Ocean Eng.* **2023**, *289*, 116137. [CrossRef]
32. Sari, Y.; Arifin, Y.F.; Novitasari, N.; Faisal, M.R. Deep learning approach using the GRU-LSTM hybrid model for Air temperature prediction on daily basis. *Int. J. Intell. Syst. Appl. Eng.* **2022**, *10*, 430–436.
33. Ie, I.-R.; Yuan, C.-S.; Lee, C.-E.; Chiang, K.-C.; Chen, T.-W.; Soong, K.-Y. Chemical significance of atmospheric mercury at fishing port compared to urban and suburb in an offshore island. *Atmos. Pollut. Res.* **2022**, *13*, 101538. [CrossRef]
34. Gustin, M.L.; Lindberg, S.E.; Poissant, L. A review of mercury volatilization from soil and sediment. *Environ. Sci. Technol.* **2000**, *34*, 4322–4337.
35. Lindberg, S.E.; Poissant, L.; Gustin, M.L. The influence of temperature on the volatilization of mercury from contaminated soil. *J. Geophys. Res. Atmos.* **1999**, *104*, 21879–21888. [CrossRef]
36. Horowitz, H.M.; Jacob, D.J.; Zhang, Y.; Dibble, T.S.; Slemr, F.; Amos, H.M.; Schmidt, J.A.; Corbitt, E.S.; Marais, E.A.; Sunderland, E.M. A new mechanism for atmospheric mercury redox chemistry: Implications for the global mercury budget. *Atmos. Meas. Tech.* **2017**, *17*, 6353–6371. [CrossRef]

37. Zhang, C.; Zhang, Y.; Wang, D. The influence of relative humidity on the volatilization of mercury from contaminated soil. *Environ. Sci. Technol.* **2012**, *46*, 10342–10347.
38. Li, J.; Liu, X.; Yu, H. The effect of relative humidity on mercury volatilization from water. *Water Res.* **2014**, *58*, 104–111. [CrossRef] [PubMed]
39. Kannan, S.; Vijayan, P. The impact of carbon dioxide on mercury methylation in aquatic systems. *Environ. Sci. Pollut. Res.* **2016**, *23*, 12135–12142.
40. Wu, Y.; Zhang, Y.; Zhang, C. The influence of carbon dioxide on the partitioning and mobility of mercury in environmental systems. *Environ. Sci. Technol.* **2017**, *51*, 7808–7816.
41. Aamir, M.; Zaidi, S.M.A. DDoS attack detection with feature engineering and machine learning: The framework and performance evaluation. *Int. J. Inf. Secur.* **2019**, *18*, 761–785. [CrossRef]
42. Zhu, L.; Husny, Z.J.B.M.; Samsudin, N.A.; Xu, H.; Han, C. Deep learning method for minimizing water pollution and air pollution in urban environment. *Urban Clim.* **2023**, *49*, 101486. [CrossRef]
43. Martínez, J.; Saavedra, Á; García-Nieto, P.; Piñeiro, J.; Iglesias, C.; Taboada, J.; Sancho, J.; Pastor, J. Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain). *Appl. Math. Comput.* **2014**, *241*, 1–10. [CrossRef]
44. Ali, P.J.M.; Faraj, R.H. Data normalization and standardization: A technical report. *Mach. Learn. Tech. Rep.* **2014**, *1*, 1–6.
45. Molnar, C. *Interpretable Machine Learning*; Lulu.com: Morrisville, NC, USA, 2020.
46. Catalano, M.; Galatioto, F.; Bell, M.; Namdeo, A.; Bergantino, A.S. Improving the prediction of air pollution peak episodes generated by urban transport networks. *Environ. Sci. Policy* **2016**, *60*, 69–83. [CrossRef]
47. Wang, S.; Shi, J.; Yang, W.; Yin, Q. High and low frequency wind power prediction based on Transformer and BiGRU-Attention. *Energy* **2024**, *288*, 129753. [CrossRef]
48. Freiesleben, T.; Grote, T. Beyond generalization: A theory of robustness in machine learning. *Synthese* **2023**, *202*, 1–28. [CrossRef]
49. Sheikholeslami, S.; Meister, M.; Wang, T.; Payberah, A.H.; Vlassov, V.; Dowling, J. Autoablation: Automated parallel ablation studies for deep learning. In Proceedings of the 1st Workshop on Machine Learning and Systems, Online, 26 April 2021; pp. 55–61.
50. Xiao, W.; Wang, C.; Liu, J.; Gao, M.; Wu, J. Optimizing Faulting Prediction for Rigid Pavements Using a Hybrid SHAP-TPE-CatBoost Model. *Appl. Sci.* **2023**, *13*, 12862. [CrossRef]