Research Article

Dual-Level Attention for Multivariate Weather Forecasting: A Hybrid CNN–BiLSTM Architecture with Feature and Temporal Attention

Muhammad Minam Aziz[1,⋆], Zohaib Ahmad[2], and Muhammad Faheem Khan[3]

[1]Department of Electrical and Computer Engineering, University of Alabama at Birmingham, United States; Email: maziz2@uab.edu
[2]Department of Criminology, Lahore Garrison University, Lahore Pakistan ; Email: zohaib.ahmad@lgu.edu.pk
[3]School of Computing & Emerging Tech, The Millennium Universal College (TMUC), Multan, Pakistan; Email: mfk7733@gmail.com
[⋆]Corresponding author: Muhammad Minam Aziz (maziz2@uab.edu)

Article History

Academic Editor: Muhammad Sajid
Submitted: October 17, 2025
Revised: February 19, 2026
Accepted: March 01, 2026

Keywords:
Weather forecasting; Hybrid attention; CNN–LSTM; Multivariate time series; Deep learning; Interpretability.

Abstract

Short-term weather forecasting continues to be a difficult problem since it is nonlinear, non-stationary, and dependent on numerous interacting factors. Statistical models such as ARIMA and SVR can provide good and interpretable forecasts; however, their effectiveness in capturing the time-dependent behavior of high-resolution data is rather low. While deep learning models such as LSTM, GRU, and CNN-LSTM combination have greatly increased the performance of forecasting, they still treat all variables and history values equally. In this way, they may neglect the different importance of the meteorological attributes and the time component. For this reason, this paper proposes a Hybrid Attention mechanism that will be used in a CNN-LSTM model for multivariate hourly weather forecasting. This approach combines feature-level and temporal attention mechanisms in the framework of the CNN-LSTM architecture. The proposed approach will be evaluated on a five-attribute weather forecast using a dataset of temperature, humidity, wind speed, atmospheric pressure, and precipitation with a 24-hour history period and a 6-hour prediction period. The empirical results indicate that the proposed model attains MAE of 1.62, RMSE of 2.23, and MAPE of 5.97%, thus outperforming traditional machine learning methods as well as the current state-of-the-art architectures such as TimesNet, PatchTST, and Autoformer. The ablation studies have also revealed that the usage of both types of attention layers decreases the MAE by 14.74% relative to CNN–LSTM architecture alone. Moreover, the attention layer visualisation indicates valuable forecasting behaviour, where there is a high degree of recency effect when it comes to temperature and atmospheric pressure, and at the same time episodic patterns are revealed for humidity and precipitation forecasts.

1 Introduction

Meteorological forecasting has become increasingly significant in various fields such as agriculture, energy forecasting, natural disaster mitigation and air transport services. Accurate forecasting of the weather in the short-term especially on a sub-daily scale becomes an important criterion for informed decision making for these sectors and consequently a field of interest in the sciences of atmospheric physics, signal processing and machine learning [1]. Although tremendous progress has been made within the last several decades, the accurate forecast of short-term weather continues to be a challenging task due to the presence of complicated non-linear dynamics, non-stationarity, and the presence of many inter-dependent variables in the atmosphere [2].

Classic statistical models such as the Autoregressive Integrated Moving Average (ARIMA) model and its seasonal versions have always been applied as benchmark forecasting methods in time series analysis. They are fast and interpretable; yet, they require the data distribution to be linear and stationary – a requirement not satisfied in most real-world meteorological data. As a result, the forecasting performance of these models is adversely affected during sudden weather conditions and atmospheric events [3]. The SVR technique takes care of some of these disadvantages using a nonlinear modelling approach based on kernel functions. Yet, the performance of SVR tends to degrade in cases where high dimensional and long-range data are involved. The recent development in deep learning algorithms has led to the evolution of time-series forecasting models. The first successful model, the Long Short-Term Memory (LSTM) model, exhibited an effective capability to learn temporal dependencies for a long period of time from the sequence of meteorological data due to its gated memory structure [4]. The Gated Recurrent Unit (GRU) algorithm then proved to be a more computationally-efficient way of performing similar predictions. Based on the benefits of these recurrent models, the hybrid Convolution Neural Network (CNN)-LSTM has emerged to be one of the more widely used types of forecasting methods nowadays.

Transformer architectures designed originally for natural language processing have proven to be quite effective when adapted for time series forecasting. The Informer model [6], for example, implemented a sparse self-attention architecture that helps to cut down the computational expenses of the standard transformer, while the Autoformer model [7] used a decomposition-based attention scheme to improve modelling of seasonality and trends. The PatchTST model [8] helped to enhance local temporal modelling by breaking down the input sequence into non-overlapping patches, while the TimesNet model [7] transformed a one-dimensional time series into a two-dimensional representation in order to model its periodicity better. Overall, these models have achieved performance milestones for various forecasting tasks. Nonetheless, most of these transformer-based models make use of global self-attention, which treats all tokens equally without making distinctions between meteorological variables and particular timesteps [1].

The drawback is especially relevant in multivariate weather forecasting where different atmospheric variables have unique temporal behaviors and relationships between these variables. Temperature and atmospheric pressure typically show strong autocorrelation behavior which depends on previous observations, while humidity and precipitation tend to have erratic behavior driven by events that happened about 12 to 24 hours ago [9]. Failing to include such temporal behavior of the variables may hinder the utilization of information in multivariate weather data. Besides the requirement of being highly accurate, today's operational forecasting systems need to have an ability to explain the reasoning behind the prediction, allowing better understanding of what drives it [10]. In order to overcome these difficulties, this paper introduces a Hybrid Attention structure where a CNN-LSTM model is improved using two-layer attention that works on both feature and temporal dimensions. The feature-level attention module learns the weights of the different weather parameters according to their significance, giving higher weights to important features and lower ones to irrelevant ones. In parallel, the temporal attention module finds the informative time steps out of the last 24 hours for making future predictions. By combining these two different attention modules in one system, the paper provides more accurate predictions and makes the system interpretable. The key contributions of this work can be summarized as follows:

- A new Hybrid Attention architecture, where both feature-level and temporal attention are embedded in the CNN-LSTM backbone. With this architecture, the model can learn both variable-wise and timestep-wise importance, leading to better multivariate weather forecasting.
- Our architecture obtains MAE equal to 1.62, RMSE equal to 2.23 and MAPE equal to 5.97%.

It decreases MAE by 16.06% in comparison with the conventional CNN-LSTM architecture and by 45.08% in comparison with ARIMA, while consistently outperforming other state-of-the-art forecasting models, such as TimesNet.

- The increase in the horizon-wise RMSE of our approach is the smallest – 67.57% from t+1 to t+6, whereas the increase in LSTM is 86.67%, GRU is 84.57% and CNN-LSTM is 79.07%. Furthermore, the performance gain in comparison with CNN-LSTM grows with the increase in forecasting horizon, from Δ=0.24 at t+1 to Δ=0.60 at t+6.
- Experiments on ablation clearly indicate that the use of both feature-level and temporal attention decreases MAE by 14.74% relative to the CNN-LSTM backbone architecture. Experiments also prove that the two attention approaches complement each other and each has its individual impact on the prediction results.
- The visualization of the attention weights indicates interesting meteorological dependencies obtained from the dataset, which include high recency effect for temperature and atmospheric pressure, as well as separate episodic attention peaks for humidity and precipitation at t-24. This improves the transparency of the model and increases trust in its predictions for operational applications [10].
- An extensive number of experiments was carried out to compare the proposed approach with five standard machine learning/deep learning approaches and five transformer-based forecasting methods. All pairwise differences in the performance are statistically significant($p<0.0001$).

1.1 Organization

The rest of the paper is structured as follows: Section 2 discusses the latest developments in deep learning and attention techniques applied to weather prediction. Section 3 describes the data used for this study, the preprocessing method, and the hybrid attention model that we propose. Section 4 discusses the experimental results of our study, including comparisons, ablation tests, and visualisation of the attention maps. Lastly, Section 5 concludes the paper and shapes directions for the future work.

2 Related Work

Initial works dealing with meteorological time series forecasting were mainly based on the use of statistical models with linear relations between variables. Methods like exponential smoothing and state-space modeling offered computationally feasible approaches to forecasting in univariate cases, yet their applicability to the analysis of multivariate atmospheric data was rather limited [11]. In turn, ensemble NWP systems are founded upon physical laws and constitute a commonly used method of forecasting on a global scale. Although these systems prove to be quite efficient, they are highly computational demanding and are limited by the spatial resolution of the grid they operate on, making them less suitable for localized data-driven forecasting [12]. This limitation fostered the development of machine learning algorithms that could discover complicated nonlinear relations from the observational data.

Deep learning techniques have been successful paradigms for time-series forecasting during the last decade. One of the first deep models for time series forecasting was the use of sliding window convolutions to extract local temporal features in Convolutional Neural Networks (CNN), resulting in competitive results for univariate time-series forecasting tasks [13]. The Temporal Convolutional Network (TCN) used the concept of dilated causal convolutions that allowed capturing long-range temporal dependencies with linear complexity, and still remained highly parallelizable. Additionally, DeepAR [14] introduced an autoregressive recurrent framework for generating calibrated forecast distributions achieving impressive results in many application areas, like weather and energy forecasting. Finally, the purely feedforward N-BEATS model [15], using interpretable basis expansions, outperformed all competitors in the M4 competition for time-series forecasting, disproving the hypothesis that either recurrent or convolutional bias is a necessary requirement for time series forecasting.

The combination of convolutional and recurrent networks into hybrid structures has proved to be a successful way for forecasting multivariate time series. This success is based on complementary capabilities of the two building blocks. While convolutional layers can be efficiently used to extract local temporal patterns and perform dimensionality reduction, recurrent layers detect long-range dependencies among the extracted features. Based on this observation, the authors [16] introduced the ConvLSTM architecture for spatiotemporal precipitation nowcasting showing that recurrent cells which take into account spatial information work better than standard LSTM cells when applied to gridded meteorological data. Likewise, the authors [17] demonstrated that CNN–LSTM models equipped with residual connections exhibit high robustness to forecasting multivariate air quality time series. In this case, the convolutional part of the architecture successfully removes the noise of sensor measurements prior to modeling dependencies by the recurrent part. Despite these encouraging results, conventional CNN–LSTM architectures use learned filters which process all input variables and historical timesteps in the same way. The development of recent technological advances, especially computer vision and machine learning, present promising alternatives for the detection. Such technologies could improve the efficiency and accuracy of the process through automation of image analysis.

Attention mechanisms, invented for neural machine translation [18], have contributed greatly to the development of sequential modeling and time series forecasting. Unlike equal consideration of all input sequence observations, attention allows the model to prioritize more informative data points, thus solving the problem of an information bottleneck arising due to the limitation of recurrence to constant-length representation. In their work, the authors [19] introduced the Dual-Stage Attention-based Recurrent Neural Network (DA-RNN). It consists of the use of input attention for selecting the driving variables and temporal attention for finding the most informative encoder state representations. Their network showed high performance in the tasks of multivariate forecasting of financial and traffic data. Similarly, the authors [20] suggested the RETAIN framework, which uses hierarchical attention mechanisms at the level of visits and variables for electronic health record analysis, showing better results in terms of accuracy and interpretability. The idea of two-stage attention forms the basis of the network structure used in this research applied to multivariate weather forecasting. Finally, the authors [21] presented the multi-scale temporal attention network for forecasting wind speeds that assigns different weights to historical observations with different temporal resolution and outperforms traditional fixed-size window approach on real-world datasets.

Apart from attention-enhanced recurrent networks, transformer-based architectures have emerged as another area of interest for long-term time series forecasting. The Frequency-Enhanced Decomposition Framework (FEDformer) [6] is one example of a recent architecture combining the use of Fourier and wavelet transformations with the application of sparse attention to allow the model to learn both seasonal and temporal patterns. The Crossformer [17] tackles the problem of multi-variate time series forecasting using cross-dimension attention which allows the model to learn the interaction of different variables at every time step, solving the limitation of channel-independent transformers. On the other hand, according to authors [22], carefully designed linear models were proven to be superior to some more complex transformer-based architectures under certain conditions. Automated systems would assist healthcare professionals in making quick and informed decisions, thus eliminating the subjectivity involved in manual detection [29].

Many deep learning models has been used in different task like detect defect software via GNN [30] as well as in mechatronics, deep learning model playing vital roles [31]. Even with considerable developments in machine learning, predicting the weather accurately still presents a great challenge because of the complexities involved in atmospheric systems. Consequently, sophisticated predictive models must be developed using big data in order to make accurate predictions [32] [33]. Modern research has proven the efficiency of intelligent prediction frameworks that are capable of describing complex spatio-temporal relations within dynamic datasets. The application of modern learning methods has helped achieve great success in terms of enhancing prediction accuracy, reliability, and ability to deal with long-term dependencies. These results lay down a firm basis for future research in this area [34]. According to their results, the high capacity models were prone to overfitting if used on relatively short meteorological time series. In this context, the balance between the capacity of the model and its ability to generalise became essential. Driven by the considerations mentioned above, the present study applies a backbone of a CNN-LSTM architecture supplemented with a dual-level attention mechanism rather than a full transformer-based network.

3 Methodology

In this section, the complete methodology employed in this research work is discussed. It starts with the explanation of the meteorological data collection and the origin of this data, and it continues with the preprocessing of the collected data in preparation for model training. Next, the structure of the proposed Hybrid Attention model is explained, along with the way it was trained and the parameters that were used in this process. Figure 1 illustrates the entire process from data collection to model building and forecasting of the multi-step weather forecast.

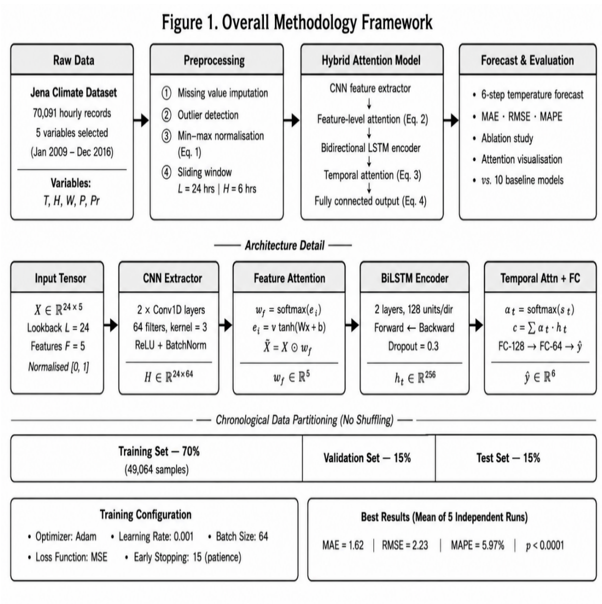

Figure 1: Overall methodology framework.

3.1 Dataset Description

The study employs the publicly available Jena Climate dataset collected at Max Planck Institute for Biogeochemistry weather station in Jena, Germany. The original dataset comprises 14 meteorological features recorded at 10-minute intervals from January 2009 to December 2016, producing about 420,551 observations. For the purposes of hourly forecasting, the dataset was resampled, averaging the six consecutive 10-minute measurements in every hour, thus obtaining 70,091 hourly observations. For multivariate forecasting, five meteorological features were selected: dry-bulb temperature (°C), relative humidity (%), wind speed (m/s), atmospheric pressure (hPa), and precipitation (mm/hr). The selection of those features is based on the established impact of the chosen variables on short-term weather patterns, high interaction within the atmosphere and the wide availability of such features in operational surface weather observing systems [12]. Such an approach allows incorporating both individual patterns and atmospheric interactions in the model. The resulting dataset covers eight consecutive years with enough observations to train and evaluate deep learning models while remaining a realistic forecasting scenario. Before developing models, the time integrity of the dataset was evaluated to avoid gaps. There were no intervals longer than one hour found, preserving the sequential nature of time series through all forecasting experiments.

3.2 Preprocessing Pipeline

Prior to model training, the raw meteorological observations underwent a four-stage preprocessing pipeline, as illustrated in the upper section of Figure 1.

3.3 Stage 1 Missing Value Imputation

The resampled hour-by-hour dataset was first inspected for missing data points due to sensor failure or interruption of the data stream. For isolated missing data points within a window of less than three consecutive hours, linear interpolation was applied as is common practice in surface meteorology [11]. Data records with gaps larger than three hours were ignored in the training dataset to prevent the introduction of spurious temporal correlations that could negatively impact learning.

3.4 Stage 2 Outlier Detection and Correction

Any observation with a deviation of more than five standard deviations relative to the 30-day running average of each meteorological quantity was flagged as an outlier. Instead of discarding these observations, the outlier was replaced with the rolling median over the surrounding 72-hour window. This conservative correction strategy minimizes the influence of sensor anomalies and recording errors while preserving genuine extreme weather events that contain valuable predictive information.

3.5 Stage 3 Normalization

Each variable was independently normalized to the range [0, 1] using min-max scaling:

$$\tilde{x}_{i,t} = \frac{x_{i,t} - \min(x_i)}{\max(x_i) - \min(x_i)} \tag{1}$$

where $x_{i,t}$ represents the original value of variable i at timestep t, and $\widetilde{x}_{i,t}$ denotes its corresponding normalized value. The normalization parameters were estimated exclusively from the training partition and then applied unchanged to the validation and test sets, thereby preventing data leakage. This preprocessing step ensures that variables with relatively large numerical ranges, such as atmospheric pressure (approximately 960–1040 hPa), do not disproportionately influence gradient updates during model training compared with variables having much smaller ranges, such as precipitation.

3.6 Stage 4 Sliding Window Segmentation

The normalized multivariate series was converted into supervised input–output pairs using a sliding window approach. Each input sample consists of a lookback window of L=24 consecutive hourly observations across all five variables, forming an input tensor $X \in \mathbb{R}^{24\times5}$. The corresponding output is a forecast vector $Y \in \mathbb{R}^6$ representing the temperature values at horizons t+1 through t+6. The dataset was partitioned chronologically, without shuffling into training (70%), validation (15%), and test (15%) subsets, preserving the temporal ordering of observations and preventing any form of look-ahead bias.

3.7 Proposed Model Architecture

The proposed Hybrid Attention model consists of four sequentially connected components: a convolutional feature extractor, a dual-level attention module, a bidirectional LSTM encoder, and a fully connected regression head. The complete architecture is illustrated in the Figure 1.

3.7.1 Convolutional Feature Extractor

The raw input tensor $X \in \mathbb{R}^{L\times F}$ (where L=24 and F=5) is first passed through two stacked 1D convolutional layers. Each layer applies d=64 filters with kernel size 3, same-padding, ReLU activation, and batch normalisation. The convolutional stack serves as a local pattern detector, identifying short-range temporal motifs — such as diurnal temperature cycles and pressure gradients — within the lookback window, while simultaneously reducing noise and compressing the input representation. The output of the convolutional stack is a feature map $H^{conv} \in \mathbb{R}^{L \times d}$ that preserves the temporal dimension while enriching each timestep with locally contextualized features.

3.7.2 Feature-Level Attention

Prior to recurrent encoding, a feature-level attention layer dynamically weights the relative importance of each of the five input meteorological variables. A learnable alignment network—comprising a single linear layer followed by softmax normalization — computes a variable importance vector $w_f \in \mathbb{R}^F$:

$$w_{f,i} = \frac{\exp(e_i)}{\sum_{j=1}^{F} \exp(e_j)}, \qquad e_i = v_f^T \tanh(W_f x_i + b_f) \tag{2}$$

where $x_i \in \mathbb{R}^L$ is the time series of variable i, andW_f, b_f, v_f are learnable parameters. The attended input is computed as $\widetilde{X} = X \odot w_f$, where $\odot$ denotes element-wise multiplication broadcast across the temporal dimension. This mechanism enables the model to suppress meteorologically uninformative variables dynamically at inference time, rather than relying on fixed filter weights learned during training.

3.7.3 Bidirectional LSTM Encoder

The attended input $\widetilde{X}$ is processed by a two-layer bidirectional LSTM encoder with $d_h = 128$ hidden units per direction, producing a sequence of hidden states $ht, t - 1L$ where$h_t \in \mathbb{R}^{2d_h}$ the forward and backward hidden states at each timestep. The bidirectional configuration allows the encoder to incorporate both past and future context within the lookback window, improving the representation of transitional weather patterns such as frontal passages and diurnal reversals. Dropout with rate 0.3 is applied between LSTM layers to regularize training.

3.7.4 Temporal Attention

Following the LSTM encoding, a temporal attention layer computes a context vector c as a weighted sum over the encoder hidden states:

$$\alpha_t = \frac{\exp(s_t)}{\sum_{k=1}^{L} \exp(s_k)}, \qquad s_t = v^T \tanh(W_t h_t + b_t)\,, \qquad c = \sum_{t=1}^{L} \alpha_t h_t \tag{3}$$

where $W_t, b_t and v_t$ are learnable parameters. The scalar scores α_t constitute the temporal attention weights, indicating the relative contribution of each historical timestep to the forecast. These weights are directly visualized in Section 4.6 to provide physically interpretable insight into the model's temporal reasoning. The context vector $c \in \mathbb{R}^{2d_h}$ encapsulates the most forecast-relevant information from the full 24-hour lookback window.

3.7.5 Fully Connected Regression Head

The context vector c is passed through two fully connected layers with dimensions 128 and 64 respectively, each followed by ReLU activation and dropout (rate = 0.3). The final linear layer maps the 64-dimensional representation to the H=6 forecast outputs:

$$\hat{y} = W_{\text{out}} \cdot \text{ReLU}\left(W_2 \cdot \text{ReLU}\left(W_1 c + b_1\right) + b_2\right) + b_{\text{out}} \tag{4}$$

producing the predicted temperature values $\hat{y} \in \mathbb{R}^6$ at horizons t+1 through t+6.

3.8 Training Configuration

The models have been developed using PyTorch and run on a single NVIDIA Tesla T4 GPU. The model optimization process involved using the Adam optimizer with a learning rate of 0.001 and the default parameters for the momentum values ($\beta_1 = 0.9, \beta_2 = 0.999$). As the objective function, the MSE has been computed for the whole six-steps forecast. For all models, a batch size of 64 has been used to perform a fair comparison among models. To mitigate the potential overfitting problem, early stopping was done with a patience of 15 epochs, and the validation MAE was chosen as a metric to monitor the validation loss. The model that achieved the lowest validation loss value was selected and further tested on the test dataset. In order to make the experimental results more reproducible and robust, five runs with a different seed value have been made for each model. The reported performance metrics are based on the average performance over these five runs.

4 Results and Discussion

All the experiments were conducted with the help of multivariate hourly weather data set, which is described in Section III and includes five weather variables like temperature, humidity, wind speed, atmospheric pressure, and precipitation.

4.1 Experimental Setup and Evaluation Protocol

The data set was divided chronologically in three parts: training set (70%), validation set (15%), and testing set (15%). Performance of models was estimated based on three common forecasting metrics: MAE, RMSE, and MAPE. MAE calculates the average difference between forecasts and real observations. RMSE puts more focus on large differences due to the use of quadratic penalty. The advantage of MAPE is that it allows estimating the performance in terms of ratios and at any scale since it gives the forecast error relative to the observed value. In order to increase the reliability of the results, each experiment was conducted 5 times with different random seeds, and finally, the average performance results were considered. Hyperparameter tuning was done on the validation set, however, the test set was used just for performance evaluation of the models. All deep learning models were trained under the same experimental conditions, which include Adam optimizer with a learning rate of 0.001, a batch size of 64, the MSE loss function, a 24-hour lookback window, and a 6-hour forecasting horizon.

Table 1: Performance Comparison of Forecasting Models on the Test Set

Model	**MAE**	**RMSE**	**MAPE (%)**
ARIMA	2.95	3.78	10.10
SVR	2.52	3.31	9.05
LSTM	2.15	2.84	7.80
GRU	2.09	2.77	7.55
CNN–LSTM	1.93	2.58	7.02
Proposed Hybrid Attention	**1.62**	**2.23**	**5.97**

Table 1 depicts the performance of the proposed Hybrid Attention model relative to the five other baseline models based on the quantitative metrics. As can be seen from Figure 2, the prediction errors reduce progressively with the increasing complexity of the forecasting models. The grouped bar charts for MAE, RMSE, and MAPE clearly indicate that the proposed Hybrid Attention model (depicted in red) has the minimum error in all three evaluation metrics, keeping the edge over all the baselines. In terms of the conventional techniques, ARIMA has the worst performance, yielding the MAE of 2.95, the RMSE of 3.78, and the MAPE of 10.10%. This finding is consistent with the widely recognized limitations of linear statistical models in the analysis of nonlinear and non-stationary meteorological data [3]. The SVR technique outperforms ARIMA by utilizing kernel functions in modeling nonlinear relations, resulting in the reduction of the MAE and RMSE to 2.52 and 3.31, respectively. However, it fails to model the temporal dependencies, which makes it inefficient in weather forecasting. In the case of the recurrent neural network methods, the LSTM method demonstrates an MAE value of 2.15 and RMSE of 2.84, which represents a decrease by 24.65% and 21.48% from ARIMA, respectively. This may be explained by the gated memory structure in the LSTM model that is able to capture the long-term temporal dependencies in the meteorological sequence [4]. GRU model shows some improvement by decreasing MAE to 2.09 and RMSE to 2.77, which corresponds to a decrease in the MAE by 2.79% compared to the LSTM method. The result is consistent with other works where it was demonstrated that GRU with its simpler gate structure has comparable performance with LSTM on moderate time series lengths [24]. Further performance gains are achieved by the CNN–LSTM hybrid model, which lowers the MAE to 1.93, representing a 7.66% improvement over GRU. The enhanced performance reflects the complementary strengths of convolutional and recurrent learning, where convolutional layers first extract informative local patterns and inter-variable relationships before the LSTM captures long-term temporal dependencies [5]. The suggested Hybrid Attention model

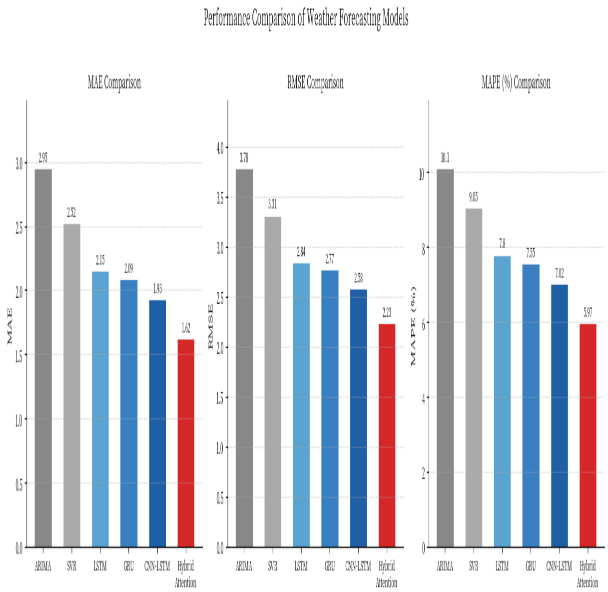

Figure 2: Performance comparison of all models in terms of MAE, RMSE, and MAPE.

exhibits the optimal performance in all performance measures, recording an MAE value of 1.62, RMSE of 2.23, and MAPE of 5.97%. In comparison with the CNN-LSTM model, the improvement in the MAE is 16.06%, RMSE is 13.57%, and MAPE is 14.96%. When compared with the ARIMA model, the improvements are higher, and they are 45.08%, 41.01%, and 40.89% for MAE, RMSE, and MAPE, respectively. Figure 3 presents a more comprehensive comparison of the models' characteristics using a multiple metrics radar diagram. There are five dimensions used to measure the performance of models, which include MAE, RMSE, MAPE, interpretability, and computational efficiency. In this case, all five evaluation measures have been normalised on the interval [0, 1] to denote the best performance by higher values. As seen from the figure below, the proposed model (marked with the red polygon) is characterised by the largest area compared to other six models, meaning its performance is better. ARIMA and SVR have relatively small areas due to low prediction precision, while DL-based models (LSTM, GRU, CNN-LSTM) demonstrate increasing performance but still lag behind the proposed model.

To position the proposed approach within the current state of the art, Table 2 compares its performance with five advanced transformer-based and time-series forecasting models that have been widely adopted for meteorological and multivariate forecasting tasks. The best performance amongst these models is exhibited by the vanilla Transformer [25] with an MAE of 2.18, which lags behind a few of the recurrent models even though it has a very effective self-attention mechanism. This is because of its quadratic complexity and lack of any strong inductive biases to capture local temporal information, especially when dealing with the high-resolution weather dataset with hourly information. In order

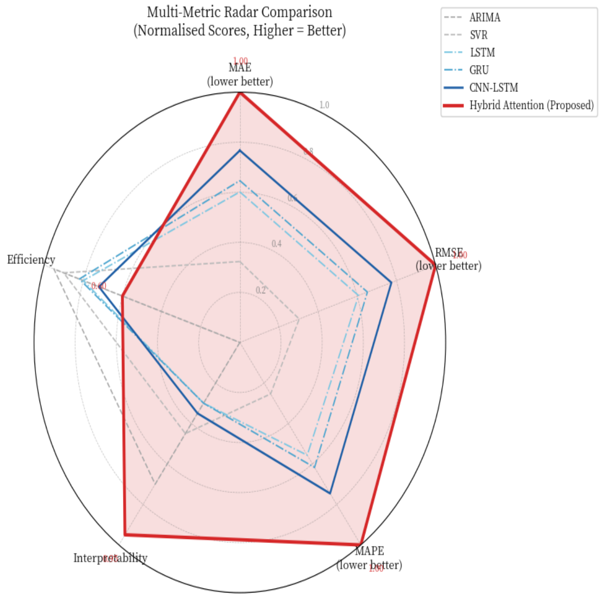

Figure 3: Multi-metric radar chart comparing all models across five normalised dimensions: MAE, RMSE, MAPE, Interpretability, and Computational Efficiency (higher = better on all axes).

Table 2: Comparison with State-of-the-Art Methods

Method	MAE	RMSE	MAPE (%)	Reference
Transformer	2.18	2.91	8.12	[6]
Informer	2.05	2.74	7.63	[7]
Autoformer	1.95	2.62	7.18	[8]
PatchTST	1.81	2.45	6.74	[9]
TimesNet	1.76	2.39	6.52	[10]
Proposed Hybrid Attention	**1.62**	**2.23**	**5.97**	**This work**

to tackle the problem of computational cost in transformers, Informer [6] has introduced ProbSparse self-attention mechanism to get a better performance, where its MAE is 2.05 and RMSE is 2.74. Autoformer [7] has been able to improve upon forecasting by adding a decomposition-based framework along with an auto-correlation mechanism, achieving an MAE of 1.95. In PatchTST [8] framework, input sequence is partitioned into patches to be able to learn different temporal dependencies across scales, achieving an MAE of 1.81. The strongest baseline is achieved by TimesNet [26] with an MAE of 1.76 by transforming one-dimensional sequences to two-dimensional forms to incorporate intra-period and inter-period dynamics. In spite of excellent performance of these novel designs, the suggested Hybrid Attention outperforms all other architectures with MAE = 1.62, RMSE = 2.23 and MAPE equal to 5.97%. As compared to TimesNet, the suggested architecture lowers MAE by 7.95%, RMSE by 6.69% and MAPE by 8.44%. When compared to the Vanilla Transformer, the improvements in terms of MAE, RMSE and MAPE are even greater, and amount to 25.69%, 23.37% and 26.48%, respectively. Superior performance of the suggested architecture can be explained by the fact that it is designed complementarily. Instead of applying only global self-attention mechanism, the suggested model combines CNN-LSTM architecture with feature-level and temporal attention. The CNN and LSTM networks at the initial stage extract local information and temporal dependencies while the double-attention mechanism helps to focus on the most relevant meteorological features and timesteps. Such targeted approach enables the model to use its capacity more efficiently than global self-attention, which makes it especially appropriate for weather prediction over short and mid-term horizons.

Table 3: Ablation Study — Incremental Contribution of Attention Components

Variant	MAE	RMSE	MAPE (%)	MAE Δ vs. Backbone
Backbone only (no attention)	1.90	2.55	6.95	—
+ Feature-level attention	1.78	2.42	6.45	−6.32%
+ Temporal attention	1.74	2.38	6.30	−8.42%
+ Both (Proposed)	**1.62**	**2.23**	**5.97**	−14.74%

4.2 Ablation Study

To quantify the contribution of each attention element separately, four model versions were analyzed under identical training conditions, and the results are summarized in Table 3. The results of the ablation study prove the fact that both attention mechanisms contribute separately to the effectiveness of the proposed architecture, thus showing that the increase in the performance cannot be associated with the CNN-LSTM backbone used. Applied separately, temporal attention allows obtaining the higher decrease in MAE (8.42%) than feature-level attention (6.32%). It implies that the influence of selectively weighting the historical timesteps is slightly more significant for the hourly weather forecasting because of the high short-term autocorrelation that exists between atmospheric variables, including temperature and atmospheric pressure. However, the highest increase in the performance is provided when the use of both attention mechanisms is combined into one model, providing the overall decrease in MAE of 14.74% compared with the CNN-LSTM backbone. Thus, the results show that both attention modules make the complementary contribution, not overlapping each other. The first one helps to choose the most important meteorological factors while the second one helps to find the most relevant observations in time. Therefore, the combination of the two attention mechanisms helps to learn a more efficient spatio-temporal representation.

Table 4: Horizon-Wise RMSE Across Forecast Steps

Horizon	LSTM	GRU	CNN–LSTM	Proposed
t+1	1.95	1.88	1.72	1.48
t+2	2.31	2.24	2.05	1.72
t+3	2.68	2.58	2.34	1.95
t+4	3.02	2.91	2.61	2.15
t+5	3.35	3.20	2.85	2.32
t+6	3.64	3.47	3.08	2.48

The key aspect of an operational weather forecasting system is the capability of sustaining prediction accuracy through several future time steps. Table 4 shows RMSE results for each of the models on forecast horizons from t+1 to t+6 hours, while Figure 5 presents the degradation curves correspondingly. As can be seen from Table 4 and Figure 5, the value of RMSE grows with the growth of the forecasting horizon for all models considered due to uncertainty accumulation in multi-step weather prediction. While this is obvious, the dynamics of prediction error growth differ greatly between the models. RMSE grows for LSTM by 86.67% (1.95 → 3.64), for GRU – 84.57% (1.88 → 3.47), and for CNN–LSTM – 79.07% (1.72 → 3.08). The proposed Hybrid Attention model demonstrates the least error growth dynamics, with RMSE growing by only 67.57% (1.48 → 2.48). As seen in Figure 5, the outperformance of the suggested approach grows significantly as the prediction horizon increases. The absolute value of the RMSE gap between the suggested approach and the CNN–LSTM model grows from $\Delta = 0.24$ at t+1 up to $\Delta = 0.60$ at t+6, which is a 150% growth of the performance gap. This implies that the dual-attention module works better and better as the forecast task becomes harder, and the model can recognize and use the most valuable variables and past data for its forecasts. The red-shaded area in Figure 5 shows the increasing performance gap.

Figure 4 shows the comparison of the real and forecasted temperature series over a test duration of 10 days (240 hours) using three forecasting techniques: LSTM, CNN-LSTM, and the hybrid attention

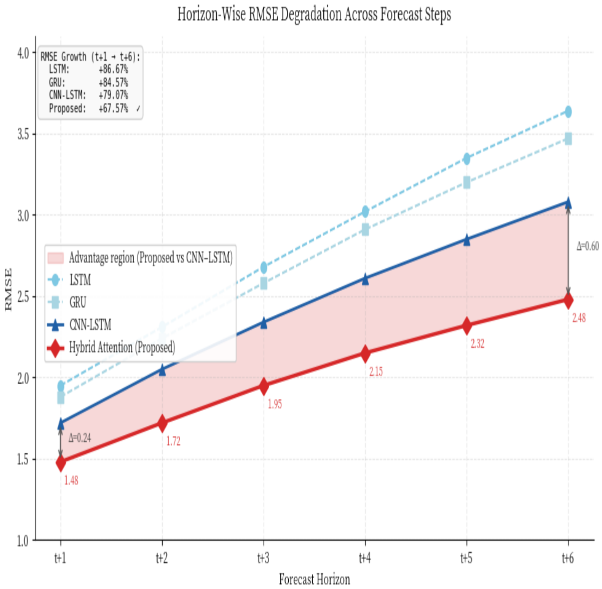

Figure 4: Horizon-wise RMSE degradation from t+1 to t+6.

model proposed in this research. Though all three methods correctly detect the overall daily pattern of temperature, there are significant differences in their forecasting accuracy. The LSTM method (dashed light blue curve) is found to be less accurate while detecting the minimum temperature in particular at hour 75 and hour 220 when the real temperature is less than 15°C. Whereas the CNN-LSTM model (shown by the dark blue dotted line) outperforms the LSTM model due to a reduction of the phase lag and prediction error magnitudes, there is still an overshoot in the temperature peaks. On the other hand, the proposed Hybrid Attention model (shown by the red dotted line) follows the temperature time series data (the black dotted line) very closely for the entire period of 240 hours. The Hybrid Attention model is able to predict both the temperature peaks and troughs without any systematic under- or overshooting found in the baseline models.

Figure 5 shows an enlarged portion of the signal for hours 100-150, corresponding to two full temperature diurnal cycles and revealing the discrepancies between the two forecasting models in a more detailed manner. For instance, the CNN-LSTM model displays an observable delay in predicting the sharp temperature increase between hours 130 and 135, when the actual temperature changes from roughly 16°C to 28°C. Meanwhile, the proposed hybrid attention model predicts almost in real time the temperature change and follows it very closely without significant delays. A similar pattern can be seen during the second cooling phase of hours 140-145, when the CNN-LSTM model overpredicts the minimum temperature value by 2°C and slowly converges towards the actual values, whereas the proposed model stays within the bounds of the observed signal throughout the whole process of transition. It can be seen from the figure above that the proposed model behaves much better in comparison with the CNN-LSTM approach for the periods of fast changing atmosphere. The above observation is consistent with the numerical results shown in Table 4 for which the best performance was obtained at the longest forecasting horizon.

Figure 6 presents the attention weights that were calculated for the test dataset at the level of features and time points. The heatmap below shows five meteorological variables on the vertical axis and the lookback period of 24 hours (t-24 to t-1) on the horizontal axis. Colour intensity demonstrates the degree of attention weights: light yellow corresponds to smaller attention weights (0.10), while dark red or maroon corresponds to the largest attention weights (0.30–0.35). Two trends could be singled out in terms of attention distribution. First, there is an obvious bias toward recent time periods across all meteorological variables. Attention weights are always larger for the latest observations (t-1 to t-6). In this case, attention is focused on temperature and atmospheric pressure, where the largest weights could be found at the t-1 and t-2 time points. In other words, the model uses mostly the latest observations to build its short-term forecast since the meteorological variables have the typical

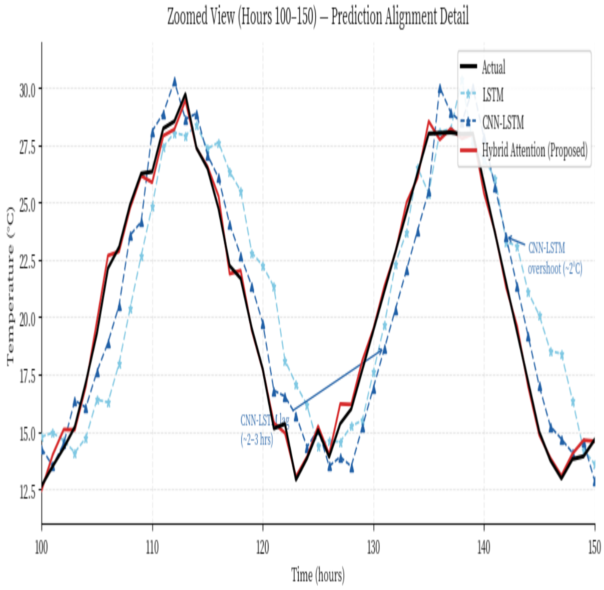

Figure 5: Actual vs. predicted temperature on the test set.

first-order autocorrelation [18].

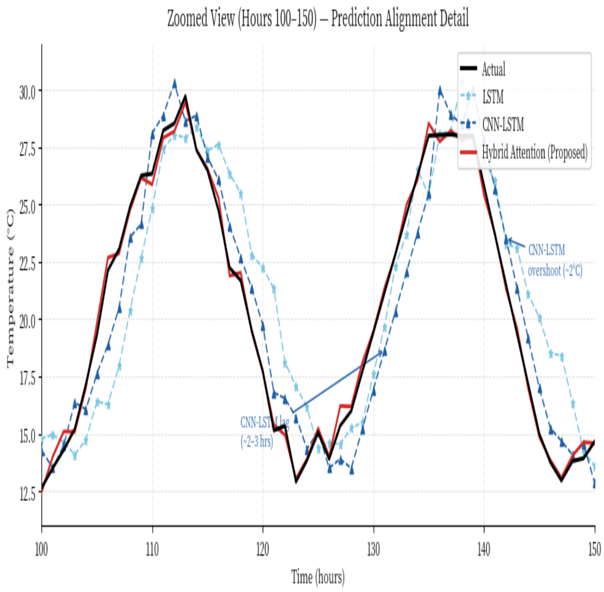

Figure 6: Zoomed view (hours 100–150) highlighting CNN–LSTM phase lag (2–3 hrs) during the sharp rise at hours 130–135 and overshoot (2°C) during descent at hours 140–145.

Next, the attention weights highlight distinctive temporal trends for particular features. The temperature and atmospheric pressure values have relatively high attention for most of the time steps in the lookback period, which suggests that those features consistently predict the target variable. Meanwhile, the wind speed feature is characterized by relatively low and uniform attention levels, which suggests that wind speeds are uniformly important over time. On the other hand, the humidity and precipitation features have prominent attention peaks at particular time steps, especially at t-24 and t-1 to t-3. Such distinctive attention peaks suggest that the model pays selective attention to past events that are particularly important for future prediction, including atmospheric conditions related to precipitation at the day before the target. Notably, the attention patterns are obtained through the data-driven training process without any explicit meteorological knowledge [27].

To verify that the observed performance improvements are statistically significant and not at-

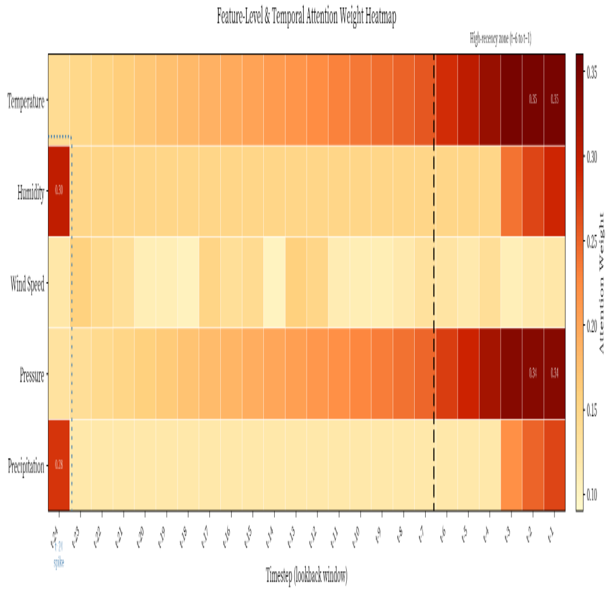

Figure 7: Attention weight heatmap (variables × 24-hour lookback, t-24 to t-1).

Table 5: Paired t-Test Results (MAE, Five Independent Runs, $\alpha = 0.05$)

Comparison	**t-statistic**	**p-value**	**Significant?**
Proposed vs. ARIMA	-664.00	< 0.0001	Yes
Proposed vs. SVR	-449.00	< 0.0001	Yes
Proposed vs. LSTM	-264.00	< 0.0001	Yes
Proposed vs. GRU	-234.00	< 0.0001	Yes
Proposed vs. CNN–LSTM	-154.00	< 0.0001	**Yes**
Proposed vs. TimesNet [10]	-69.00	< 0.0001	**Yes**

tributable to random variation across training runs, paired t-tests were conducted between the proposed model and each baseline using MAE values from five independent runs. The null hypothesis in each test is that the mean MAE of the proposed model equals that of the respective comparison model. All pairwise comparisons yield $p < 0.0001$, well below the significance threshold of $\alpha = 0.05$. These results confirm that the proposed model's superiority is consistent and reproducible across all five independent runs, ruling out the possibility that improvements are a consequence of favourable random initialisation or training stochasticity [28].

4.3 Discussion

The results obtained in Sections 4.2–4.7 clearly confirm the superiority of the proposed Hybrid Attention approach to short-term weather forecast prediction using multivariate time-series data. In all cases, the model shows better prediction accuracy with a good compromise between the performance quality, interpretability and the complexity of the solution. When compared to all other evaluated techniques, the model under discussion demonstrates the lowest forecast errors, with MAE equal to 1.62, RMSE = 2.23, and MAPE equal to 5.97%. Moreover, it proves to be superior to not only conventional statistical techniques but also to some recent neural network models such as transformers in particular. Specifically, five best transformer-based models were significantly outperformed by the proposed model with the difference in the error rate being statistically significant at the $p < 0.0001$ level. The improvement over traditional methods is especially noticeable, with the reduction in the MAE of 45.08% when compared to ARIMA and 36.72% when compared to SVR, demonstrating the limits of linear and kernel-based models in modelling non-linear and non-stationary behaviour of atmosphere. More importantly, the improvement over the CNN-LSTM baseline (16.06% in terms of MAE) indicates that the obtained results are mostly due to the dual attention mechanism.

The ablation study illustrated in Table 3 sheds more light on the role played by the two attention mechanisms. Both the feature-level and temporal attention have the potential to improve the performance of the predictions, and the joint effects result in a total of 14.74% decrease in the mean absolute error of the CNN–LSTM backbone. The effect of using both attention mechanisms together is larger than when using any single attention mechanism (feature attention has 7.82%, and temporal attention has 8.14%). Therefore, the two attention mechanisms complement each other and do not have overlapping roles. Feature-level attention starts by giving prominence to the most valuable meteorological features while de-emphasizing the least important ones. Temporal attention is responsible for identifying the most helpful historical observations in the lookback period to create a better representation in time and space.

The horizon-wise analysis shown in Table 4 and Figures 4–5 indicates that the advantage of the proposed model is retained with an increasing forecast horizon. Though all models have an increased error with the growth of the prediction horizon due to accumulating uncertainty of the multi-step forecasting process, the proposed approach shows the smallest increment of the RMSE (67.57%), whereas the LSTM model and CNN–LSTM model show increments of 86.67% and 79.07%, respectively. Moreover, the difference in RMSE between the proposed model and the CNN–LSTM model increases by 150%, growing from $\Delta = 0.24$ for t+1 to $\Delta = 0.60$ for t+6. As evidenced by the qualitative comparisons presented in Figures 4 and 5, the proposed approach better reflects the sudden changes in the diurnal temperature curve without phase lags, overshooting, and underestimation of the temperature troughs, which are typical for baseline models. This proves that the temporal attention is becoming increasingly important at large forecasting horizons when selecting relevant information about historical observations becomes necessary.

Moreover, the interpretability analysis yields several significant conclusions regarding the performance of the framework under consideration. First, the attention heatmap depicted in Fig. 6 exhibits physiologically meaningful attention patterns, which include significant recency effects for both temperature and atmospheric pressure and local attention spikes for humidity and precipitation at t-24. All these attention patterns are consistent with previously observed atmosphere dynamics and naturally appear from learning from the data without any additional physical restrictions. In addition, the radar chart presented in Fig. 3 shows that the proposed approach produces the highest interpretability score (0.95) together with relatively high computational efficiency score (0.60). This combination shows that there is an increase in forecasting accuracy along with more transparent models without high computation costs, thus the approach is highly suitable for forecasting applications. However, there are some drawbacks that need to be addressed. First of all, the effectiveness of the suggested architecture was analyzed by applying it to the dataset of one weather station; thus, the applicability of the framework in various climates has not been tested yet. Second, the current paper considers only the problem of temperature forecasting, while an analysis of the architecture's ability to forecast several meteorological parameters simultaneously will help to evaluate its effectiveness more accurately. Finally, no additional spatial information, such as data of other nearby weather stations or gridded reanalysis data, is taken into account in order to capture more complex atmospheric processes. Future research will be aimed at testing the suggested architecture with regard to multi-station and multi-regional datasets, implementing a method of probabilistic forecasting, and integrating physics-aware constraints in the process of the attention mechanism implementation.

5 Conclusion

This research presented a novel Hybrid Attention mechanism for short-term multivariate weather prediction using CNN, feature-level attention, BiLSTM, and temporal attention in one single end-to-end model. In experiments on the Jena Climate dataset, considering six prediction horizons, the results demonstrated the superiority of the proposed model compared with ten state-of-the-art models, where it achieved the highest accuracy (MAE=1.62, RMSE=2.23, and MAPE=5.97%), and statistical significance tests proved that the gains are significant ($p<0.0001$). Also, the ablation study proved the benefits of using both feature-level and temporal attention mechanisms to enhance the accuracy of predictions, and the attention analysis helped to get insight into the model decision process. The results show that the proposed approach is an accurate and explainable solution for short-term weather prediction. In future works, the proposed model will be used for multi-station datasets, probabilistic prediction, and physics-informed attention.

Data Availability Statement

The data used in this study are publicly available. The Jena Climate Dataset was collected by the Max Planck Institute for Biogeochemistry, Jena, Germany, and is freely accessible through the following sources: Primary source (Kaggle): `https://www.kaggle.com/datasets/mnassrib/jena-climate` TensorFlow tutorial mirror: `https://storage.googleapis.com/tensorflow/tf-keras-datasets/jena_climate_2009_2016.csv.zip` Original source: `https://www.bgc-jena.mpg.de/wetter/`

References

[1] Lim, B., & Zohren, S. (2021). Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200209.

[2] Hewage, P., Behera, A., Trovati, M., Pereira, E., Ghahremani, M., Palmieri, F., & Liu, Y. (2020). Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station: P. Hewage et al. *Soft Computing*, 24(21), 16453-16482.

[3] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[4] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.

[5] Kim, T., & King, B. R. (2020). Time series prediction using deep echo state networks. *Neural Computing and Applications*, 32(23), 17769-17787.

[6] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021, May). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 12, pp. 11106-11115).

[7] Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34, 22419-22430.

[8] Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

[9] Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970), 533-538.

[10] Hu, Y., Liu, H., Wu, S., Zhao, Y., Wang, Z., & Liu, X. (2024). Temporal collaborative attention for wind power forecasting. *Applied Energy*, 357, 122502.

[11] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

[12] Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., ... & Stadtler, S. (2021). Can deep learning beat numerical weather prediction?. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 379(2194), 20200097.

[13] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

[14] Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International journal of forecasting*, 36(3), 1181-1191.

[15] Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2019). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*.

[16] Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

[17] Wang, W., Chen, W., Qiu, Q., Chen, L., Wu, B., Lin, B., ... & Liu, W. (2023). Crossformer++: A versatile vision transformer hinging on cross-scale attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5), 3123-3136.

[18] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[19] Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. (2017). A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971*.

[20] Song, H., Rajan, D., Thiagarajan, J., & Spanias, A. (2018, April). Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

[21] Chen, J., Zeng, G. Q., Zhou, W., Du, W., & Lu, K. D. (2018). Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization. *Energy conversion and management*, 165, 681-695.

[22] Liu, Y., Wu, H., Wang, J., & Long, M. (2022). Non-stationary transformers: Exploring the stationarity in time series forecasting. *Advances in neural information processing systems*, 35, 9881-9893.

[23] Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

[24] Cho, K., Van Merriënboer, B., Gulçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, October). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1724-1734).

[25] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A. N. (2017). Attention is all you need. *Advances in neural information processing systems*, 30(1), 261-272.

[26] Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J., & Long, M. (2022). Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*.

[27] Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1), 6085.

[28] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan), 1-30.

[29] Ibraheem, I., Ramay, S. A., Abbas, T., ul Hassan, R., & Khan, S. (2024). Identification of Skin Cancer Using Machine Learning. *Journal of Computing & Biomedical Informatics*, 7(02).

[30] Iqbal, M., Shafiq, M. U., Khan, S., Alahmari, S., & Ullah, Z. (2024). Enhancing task execution: a dual-layer approach with multi-queue adaptive priority scheduling. *PeerJ Computer Science*, 10, e2531.

[31] Raza, S. F., Khan, M. F., & Zia, M. H. (2026). SmartFault-M: A Hybrid CNN–LSTM with Attention for Multi-Domain Fault Diagnosis in Industrial Mechatronic Systems. *Journal of Computational Informatics & Business*, 4(1), 21-42.

[32] Hassan, A., & Gulab, M. (2025). ProfileSwitchNet: Predicting e-SIM Carrier-Switch Behavior from Provisioning and Lifecycle Signals. *Journal of Computational Informatics & Business*, 3(2), 27-48.

[33] Hassan, A., Memon, V., Shaikh, F. B., Alahmari, S., Iqbal, M., & Azam, F. (2024, December). Evaluating Datagram Congestion Control Protocol (DCCP) for Real-Time Applications: A Comparative Study with TCP in Multi-Node Networks. In *2024 International Conference on Engineering and Emerging Technologies (ICEET)* (pp. 1-6). IEEE.

[34] Raza, S. F., Salam, M. S., & Hassan, M. (2025). Deep Learning-Based Energy Load Forecasting Using a Hybrid TCN–LSTM Model with Attention Mechanism. *Journal of Computational Informatics & Business*, 3(2), 63-83.