Al-Bahir

Manuscript 1130

A Multi-Task CNN-LSTM Model with Attention Mechanism for Climate Temperature Forecasting over Iraq

Salim M. Mohammed

Omar M. Mustafa

Lailan M. Haji

Omar M. Ahmed

ORIGINAL STUDY

A Multi-task CNN-LSTM Model With Attention Mechanism for Climate Temperature Forecasting Over Iraq

Salim M. Mohammed [a], Omar M. Mustafa [a], Lailan M. Haji [a], Omar M. Ahmed [b,*]

[a] Department of Computer, College of Science, University of Zakho, Zakho, Kurdistan Region, Iraq
[b] Computer Information System Department, Duhok Polytechnic University, Duhok, Kurdistan Region, Iraq

Abstract

Accurate forecasting of climate temperature at the regional scale is crucial to agriculture planning, water resource management, and disaster preparedness. Still, due to the nonlinear and complicated temporal patterns in regions such as Iraq, conventional climate models fail to meet this imperative. This paper presents a novel deep learning model that not only predicts monthly average temperature values but also predicts the temperature trend, hence addressing the regional climate prediction problem holistically. The proposed model is designed as a multi-task deep learning framework integrating Convolutional Neural Networks, Long Short-Term Memory networks, and an attention mechanism to process historical monthly temperature data obtained from the Berkeley Earth dataset using a sliding window method. The model is learned a shared representation for both regression and classification, therefore optimized together. Classification accuracy of 94.81 %, ROC AUC of 0.9839, and R^2 score 0.9773 for regression were obtained, thus proving the capability of model to efficiently capture local and long-range temporal dependencies and give importance to influential time steps. As such, the proposed CNN-LSTM-Attention model in multi-task learning configuration presents a promisingly accurate, interpretable, and computationally-effective temperature prediction and trend classification, applicable to regional climate modeling and decision support for climate-sensitive areas.

Keywords: Climate forecasting, Deep learning, CNN-LSTM hybrid, Attention mechanism, Multi-task learning

1. Introduction

Precise prediction of climate elements (temperature, humidity, and precipitation) is crucial in current-day life. From planning of agriculture and managing of water resources to advising public health and disaster preparedness, climate prediction has a wide impact [1]. Of all climatic features, temperature prediction is so fundamental that it forms a cornerstone of extensive environmental models [2].

Statistically and physically simulated methodologies are traditionally used when modeling climate. They consist of autoregressive models that include Autoregressive Integrated Moving Average (ARIMA) or Global Climate Models (GCMs) that predict atmospheric and oceanic processes through mathematical representations of the laws of physics. Even though these models have proved their usefulness, they tend to perform poorly at regional scale prediction or cannot handle with nonlinear and chaotic time dependency experienced by climatic systems [3]. Furthermore, classic models are often demanding on resources, and may even have reliance on considerable domain knowledge, extensive spatial data, or restrictive assumptions about environments that are inherently dynamic [4].

The popularity of machine learning (ML) and deep learning (DL) methods has increased

Received 26 August 2025; revised 4 November 2025; accepted 21 November 2025.
Available online 23 March 2026

* Corresponding author.
E-mail address: omar.alzakholi@dpu.edu.krd (O.M. Ahmed).

https://doi.org/10.55810/2313-0083.1130

throughout climate science due to increasing availability of big data and overall advances in computational power. The data-driven model, which in turn can develop and learn the patterns from historical datasets automatically, provides better scalability and is able to capture complex, nonlinear relationships [5]. One of the most popular DL models for time series forecasting is the Long Short-Term Memory (LSTM) network, which is one of the Recurrent Neural Networks (RNNs) and can effectively model long-range temporal dependencies [6]. LSTMs have been successfully-employed in fields including stock market prediction, speech processing and energy load prediction, and are becoming more popular in climatology [7].

However, even accomplished deep learning models for climate prediction are still limited in some aspects. They mostly concentrate only in classification or regression. However, in practice, models that can do both tasks at the same time are desired more. For example, agricultural stakeholders may be interested in knowing not only that the temperature is going to increase but by how much, to schedule water irrigation or crop rotation. A model that can do multi-task learning/training on multiple tasks inside one architecture can provide richer and more useful results [8].

In this paper, we aim to fill this gap with a multi-task deep learning model that integrates three strong neural network elements − CNNs for local feature encoding, LSTMs for memorization of sequential information, and a special attention mechanism for capturing dynamic temporal attention focus. The novelty of our work is not only in its capability to learn both classification and regression temperature trends in a unified framework, but also in how it combines these heterogeneous tasks into a single, interpretable and accurate forecasting approach.

Multitask learning (MTL) is a machine learning approach in which a single model learns to perform multiple tasks. In our setup, latter enables the model to share representations between trend classification and temperature regression tasks. This common representation can improve the model's generalization and mitigate overfitting, particularly when the data are scarce or noisy. Furthermore, since the two tasks are jointly optimized, the model is incentivized to learn more informative and discriminative feature representations optimizing each of the cost function, often leading to an improved overall performance [9].

The early use of CNN layers in the architecture allows the model for learning localized patterns in the time series data, such as abrupt spikes or drops, which can be key predictors of a change in trend. CNNs are also computationally efficient enabling fast feature extraction. Afterwards, the LSTM layer models the long-term dependencies in the sequential data, then learns to understand how previous patterns affect future outcomes. This is especially true in the climate realm, where things like seasonal cycles or episodes of long-term warming may develop or persist over many months [10].

A major weakness of vanilla LSTM models is their natural tendency to assign equal importance to all time steps, which can wash away important signals. The attention mechanism overcomes this restriction by enabling a model to assign various importance to different time steps. This ability does not only increase performance, but also facilitates interpretability, as the model's attention to specific months (e.g., summer vs. winter) can be visualized and studied [11].

While existing deep learning models for climate forecasting typically focus on either regression (predicting exact temperature values) or classification (predicting trends), few models are designed to handle both tasks simultaneously. This limitation reduces their practical utility in real-world applications where both detailed values and directional insights are needed. Moreover, conventional time-series models often fail to capture complex temporal dependencies and local variations inherent in climate data.

While deep learning models such as standalone LSTM or CNN architectures have shown promise in climate forecasting, they often face limitations in capturing both short-term and long-term dependencies effectively, lack interpretability, and are usually optimized for a single task. Our proposed approach integrates CNNs to capture local temporal features, LSTMs to model sequential dependencies, and attention mechanisms to enhance interpretability by weighting significant time steps. By employing a multi-task learning strategy, our model simultaneously performs temperature regression and trend classification, promoting knowledge sharing and generalization.

The main contributions of this work are summarized as follows:

- We propose a multi-task deep learning framework that simultaneously performs temperature

forecasting (regression) and temperature trend classification, offering richer predictive insights than single-task models.
- We develop a hybrid architecture combining CNN, LSTM, and attention mechanism, where CNN captures short-term temporal patterns, LSTM learns long-term dependencies, and attention enhances interpretability by identifying influential time steps.
- We apply this architecture to real climate data of Iraq, leveraging over a decade of monthly temperature records from the Berkeley Earth dataset.
- We demonstrate state-of-the-art performance in both classification (94.81 % accuracy, 0.9839 ROC AUC) and regression ($R^2 = 0.9773$), indicating strong generalization and accuracy.
- We provide interpretable outputs via attention weights and sequence modeling, which highlight which months contributed most to the predictions—useful for domain experts in climate and policy planning.

The remainder of this paper is organized as follows: Section 2 provides a review of related work on deep learning-based climate prediction. Section 3 presents the dataset, model architecture, experimental setup, and evaluation metrics. Section 4 discusses the experimental results for both classification and regression tasks. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Related works

Recent research works have shown the feasibility of deep learning in temperature prediction and -climate modeling combining innovative neural architectures and new datasets.

Baño-Medina et al. [12] developed DeepESD, a convolutional neural network (CNN) based ensemble downscaling approach with the CMIP5 datasets. This justifies the use of their approach to mitigate distributional bias and preserve inter-model uncertainty over the full range of temperature projections. Shukla & Halem [13] proposed DUNE (deep UNet++) which is a ensemble of deep UNet++ model for predicting the monthly, seasonal and annual temperatures using the ERA5 reanalysis data. The model, in multi-encoder-decoder structures, held the performance similarities with the NOAA forecast and provided the high resolution for the output. Zhang et al. [14] proposed a customized graph convolutional network with static and dynamic layers for SST prediction. Their approach is capable of capturing both short-term and long-term spatiotemporal dependencies. Li et al. [15] proposed AMUN, a U-Net-based model with attention mechanisms aimed at downscaling ERA5-Land LST data. The model enhances predictive accuracy and is optimized through Bayesian hyperparameter tuning.

Similarly, Li X et al. [16] further improved with SD-LPGC using learnable personalized convolution for the feature fusion of multiscale SST signals, and achieved the best results on real dataset based on. Gómez-Gonzalez et al. [17] have applied deep convolutional networks and generative adversarial networks to downscale SEAS5 seasonal forecasts under Iberia, demonstrating improved regional skill compared to traditional analog techniques. Accarino et al. [18], a GAN-based statistical downscaling approach for 2-m temperatures, called MSG-GAN-SD. It incorporates multi-scale gradient information to maintain spatial detail on EURO-CORDEX temperature fields. Wei et al. [19] developed a deep learning model that integrates super-resolution and data harmonization to improve the upscale downscaled CMIP6 near-surface temperature projections. Their model performed better in terms of MAE and spatial consistency with CRU TS as a reference to compare with. Fernandez & Barnes [20] presented a Hybrid Machine Learning and Analog Forecasting model for multi-year temperature forecasting. Their neural network must recognize important regions for analogue selection and their experimental work on the Berkeley Earth.

Recently, Xu et al. [21] proposed a hybrid architecture that integrates a Temporal Convolutional Network (TCN) with CNN and LSTM layers for monthly runoff forecasting. Their TCN-CNN-LSTM framework demonstrated the ability to capture both long-term and local temporal dependencies, achieving higher accuracy than standalone models. Similarly, Wang et al. [22] introduced a stacking ensemble approach that combined multiple machine learning algorithms with a meta-learner to enhance runoff prediction performance. These studies highlight the emerging trends of hybrid and ensemble methods in environmental forecasting. In contrast, our work advances the field by incorporating CNN, LSTM, and an attention mechanism within a multi-task learning framework, applied specifically to temperature forecasting over Iraq. This not only improves predictive accuracy but also enhances interpretability and practical utility.

These studies have shown that the combination of deep learning architectures (e.g., CNNs, GNNs and UNets) with climate datasets achieves improved

performance in temperature forecasting and downscaling tasks, providing a solid basis for your multi-task deep learning framework. Despite the promising performance of these deep learning models, the majority of them focus on either regression or classification separately, lacking a unified framework that can handle both tasks concurrently. Moreover, most studies emphasize spatial downscaling or short-term predictions without leveraging multi-task architectures that combine temporal feature learning, interpretability, and joint optimization. Our proposed model addresses this gap by integrating CNN, LSTM, and attention mechanisms into a unified multi-task learning framework that simultaneously performs temperature forecasting and trend classification, tailored specifically to the regional climate of Iraq.

3. Methodology

In this paper, a hybrid model is proposed to solve the combined problem of predicting temperature and classifying the trend on temperature. The proposed approach can be broken down into five essential ingredients: data preprocessing, sequence generation, model architecture, multi-task learning formulation, and training settings. All components are necessary to achieve strong and interpretable performance on the climate dataset. Fig. 1 shows the approach of our papers.

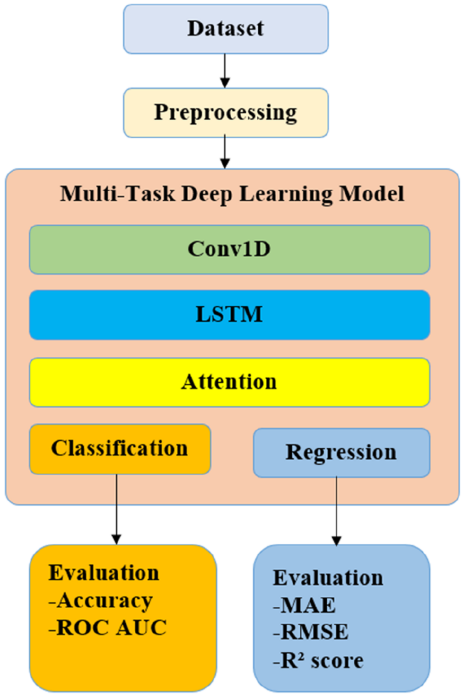

Fig. 1. The flowchart of our approach.

3.1. Data collection and preprocessing

The dataset used in this study originated and extracted from the Berkeley Earth project, a widely used resource for climate data offering long term historical temperature records per country. Here, we return to real climate data of Iraq for decades and focus on the monthly averaged temperature. This data were selected due to its time resolution and its significance for regional climate trends assessment. The historical temperature dataset spans from January 1819 to September 2013 consist of 2339 monthly records, providing nearly two centuries of monthly climate observations for Iraq. In our setup, the prediction horizon is fixed to one month ahead, the model uses 12 months of historical temperature data as input to predict the temperature and trend for the 13th month.

On first glance, the raw data contained fields like Date, AverageTemperature, Country, and so on. We pre-processed the dataset by each step as follows to ready it for model training:

Filtering: Any titles unrelated to Iraq were discarded to ensure geographic relevance.

Dealing with Missing Values: We dropped rows with missing temperature values to avoid feeding the system with spurious patterns. We even tried forward-filling which we didn't have to use as the missing entries were small in number.

Date Parsing: The Date column was transformed into datetime format for better ordering chronologically as well as for harvesting the features.

Resampling: Even though the dataset was monthly gridded data, we checked and applied monthly frequency regularity using fixed frequency.

Feature Engineering: We captured more information from the timestamp to make the model learn the seasonal patterns, such as, months encoded as an integer (1–12). Sine and Cosine month codes for cyclic months representations.

Once the data is cleaned and enriched, we performed Min-Max normalization to normalize the temperature to the range 0–1. This served as a crucial step for speeding up the convergence of the training of neural networks and making the regression and classification loss have the same number of scales.

3.2. Sequence generation

As the climate data is time series, we organized the data into input-output sequences by applying a sliding window methodology. An input sample is composed of an average temperature data of 12

consecutive months. These sequences are utilized as a temporal context, allowing the model to make two predictions: 1) to compute the average temperature of the 13th month (regression); and 2) to decide whether the temperature of the 13th month is greater or less than to the 12th (binary classification).

This two-stage structure endows the model with the capability to catch the continuous tendency of temperature evolution and the directionality shift across months. When a sequence is an input that is converted into shape (12, 1) while temperature values are inputs. If you include the other features such as month encoding etc, the shape will be (12, 3). The model outputs two values for each sequence: a single scalar for the predicted temperature and a binary label for the direction of change (1 for increase, 0 for no change or decrease).

We applied a sliding window approach with a fixed window size of 12 months and a stride of 1 month. This means that each input sequence consists of one year of consecutive monthly temperature values, and the model makes predictions for the next month. Using a stride of 1 results in overlapping sequences, maximizing data usage and capturing smooth temporal transitions. This dense sampling improves the model's ability to learn short-term dynamics and seasonal trends, contributing to higher temporal resolution and better generalization.

In order to guarantee temporal integrity and prevent information leak, we split the data chronologically as described: the first 80 % and the last 20 % of sequences were used as training and test datasets. This able to simulate the situation where one has to forecast future data using past observations.

3.3. Model architecture

The model is a combination of CNN, LSTM, and Attention methods, wrapped by a multi-task learning. Each block aims to solve a specific problem of classical viewpoints in time series models.

We introduce a 1D convolutional layer which acts as a local feature extractor. CNNs are successful at learning short-term trends and noise in the data. To capture local correlations between consecutive months, the Conv1D layer convolves several filters on the sequence. It also helps the model to concentrate on short-term transitions before giving the information to the LSTM.

We add an LSTM layer after the CNN layer to capture longer-time dependencies. LSTM is a special kind of RNNs, capable of learning the complex series dependencies, and often in practice, remembering information for a long time. In a climate prediction context, this enables the model to "see" the seasonal and interannual variations, such as temperature anomalies or steady upward or downward changes.

For each time (event) in the input sequence, the LSTM will output a hidden state. This chain of hidden states is provided as input to the attention mechanism.

Mimicking the Transformer, we propose a custom attention layer that scores the importance of every timestep in the input sequence to the final result. The attention models are designed to give more weight to the months that have very high impact on future temperature.

In other words, for each time step t, the attention weights a_t are determined based on a trainable alignment function, where the context vector is computed as a weighted sum of all the LSTM outputs. This approach adds interpretability, as we can see what months had the biggest effect on a given prediction.

The shared features extracted by the CNN-LSTM-Attention backbone are then passed to two parallel output heads:

Regression Head:

- A fully connected layer with a linear activation function
- Outputs a single value representing the predicted average temperature for the next month

Classification Head:

- A fully connected layer with a sigmoid activation
- Outputs a probability between 0 and 1, which is thresholded at 0.5 to indicate whether the temperature is expected to rise

This multi-task setup enables the model to leverage shared learning between tasks, improving generalization and efficiency.

The attention mechanism used in our model is based on additive attention. For each time step t, the attention score e_t is calculated as:

$$e_t = v^T \tan \mathrm{h}(W_h h_t + b_h)$$

where h_t is the hidden state output from the LSTM at time step t, and W_h, h_t, and v are trainable parameters. The attention weights a_t are then obtained using a softmax function:

$$a_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)}$$

The final context vector c is computed as the weighted sum of all hidden states:

$$c = \sum_{t=1}^{T} a_t h_t$$

The 1D convolutional layer used in our model includes 64 filters with a kernel size of 2 and ReLU activation, allowing it to capture local patterns in the temperature sequences. This is followed by an LSTM layer with 64 units and $return\_sequences = True$ to retain the sequence dimension necessary for the attention layer. The attention mechanism is implemented using additive attention, where each hidden state from the LSTM is compared to a trainable context vector, and attention weights α_t are computed via softmax normalization. These weights are used to compute a weighted sum of LSTM outputs, producing a context vector that highlights the most relevant time steps. This context is concatenated with the LSTM outputs and passed to two parallel dense layers for classification and regression. Table 1 summarizes the model's layers, including input and output shapes, the number of trainable parameters, and a brief description of each component.

In summary, the input time series is processed by a 1D CNN layer to gather local temporal features execution. Those features are then taken to a LSTM layer for long-term dependencies. The sequence of hidden states shells is fed into an attention mechanism that computes the importance of each time step. This is then followed by a shared context vector branched int two-regions: regression of the next month's average temperature using a fully connected layer, and binary classification for temperature trends. This multi-task learning allows the model to learn the two tasks together using shared representations and thus improve the performance and generalization.

3.4. Multi-task learning strategy

We designed the model to learn two objectives simultaneously:

- Temperature Forecasting (Regression):
  - Loss function: Mean Squared Error (MSE)
- Trend Classification:
  - Loss function: Binary Cross-Entropy (BCE)

To train the model jointly, we define a composite loss function:

$$Total\ Loss = \lambda_1 \cdot MSE + \lambda_2 \cdot BCE$$

In our case, we set $\lambda_1 = \lambda_2 = 1.0$, giving equal importance to both tasks. These weights can be tuned depending on specific application needs.

Multi-task learning encourages the model to learn a shared representation that benefits both tasks. For example, identifying that a month is significantly warmer than usual may help both in forecasting the exact temperature and predicting a rising trend.

The model was implemented in TensorFlow using the Keras API. The training configuration is as follows:

- Optimizer: Adam (adaptive moment estimation)
- Learning rate: 0.001 (default)
- Batch size: 16
- Epochs: 20
- Validation split: 10 % of training data

We monitored both loss components and validation accuracy using TensorBoard during training. Early stopping was also considered to prevent overfitting but was not triggered due to the model's stable performance.

Table 1. Summary of the proposed CNN-LSTM-attention multi-task model architecture.

Layer type	Output shape	Parameters	Description
Input (12, 1)	(None, 12, 1)	0	12 time steps, 1 feature (temperature)
Conv1D	(None, 11, 64)	192	64 filters, kernel size 2, ReLU activation
LSTM	(None, 11, 64)	33,280	Return sequences = True
Attention	(None, 11, 64)	4224	Additive attention (query = value = LSTM outputs)
Concatenate	(None, 11, 128)	0	Concatenate LSTM and attention
Flatten	(None, 1408)	0	
Dropout (0.3)	(None, 1408)	0	
Dense (ReLU)	(None, 32)	45,088	Shared layer before branching
Dense (Sigmoid)	(None, 1)	33	Classification head
Dense (Linear)	(None, 1)	33	Regression head
Total	–	82,850	

To further assess performance, we used the following evaluation metrics:

- For classification: Accuracy and Receiver Operating Characteristic Area Under the Curve (ROC AUC).
- For regression: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2 score)

Visualizations were also generated post-training to analyze temporal forecasts and classification decisions.

In the context of climate forecasting, regression and classification tasks are inherently connected — predicting the magnitude of temperature (regression) can reinforce understanding of the direction of change (trend classification), and vice versa. By jointly learning both tasks, the model captures richer temporal representations. For example, a consistent warming trend identified by the classification head can guide the regression head to adjust its expectations upward, improving accuracy. This synergy is particularly valuable in climate applications where trends (e.g., seasonal transitions or anomalies) carry strong predictive signals that can enhance the estimation of future values.

4. Results and discussion

In this section, we demonstrate and analyze the experimental results of our CNN-LSTM-attention model based on multi-task learning. Our objective was two-fold: (i) to accurately estimate the average temperature of the following month (regression) and (ii) to categorize if the temperature will increase or decrease with respect to the previous month (binary classification). We report results in the next two sections in a series of quantitative numbers for both tasks, interpret the visualizations and investigate the role of different model components.

4.1. Classification task performance

The binary classification problem consisted of forecasting the next month's average temperature warmer than the current month's. Being able to perform this task can be key to applications such as crop planning, early warning systems, and climate policy decisions. We assessed classification performance using several common metrics − accuracy, precision, recall, F1-score and ROC AUC.

The model achieved an accuracy of 94.81 %, meaning that nearly 95 out of 100 predictions correctly identified the temperature trend. This is a strong indication that the CNN-LSTM-attention architecture captures both local and long-range patterns relevant to temperature transitions. Furthermore, the ROC AUC score was 0.9839, a near-perfect result that shows the model's excellent ability to distinguish between increasing and non-increasing temperature classes.

The ROC curve (Fig. 2) shows a trade-off between true positive rate and false positive rate. In general, a curve that closely envelops the upper left corner is indicative of high sensitivity and specificity, as is the case with ours. This robustness is crucial in climate models as false alarms or missed detections can have profound economic and environmental impacts.

This classification task was designed to support early warnings in agriculture or water resource planning. Analyzing the ROC curve, we observe that the model performs best in transition months (e.g., March, October), suggesting its strength in capturing seasonal changes. Misclassifications were more frequent during stable temperature periods, which aligns with expectations.

4.2. Regression task performance

The model was trained, for the regression problem, to predict the average temperature of the next month precisely. This task is more informative than binary classification and is crucial for accurate prediction in applications such as energy consumption scheduling and water resources management.

The performance of our regression model was obtained by calculating the MSE, MAE and R^2 score. The results were as follows:

- MSE: 1.9665
- MAE: 1.1168 °C
- R^2 Score: 0.9773

The MAE of 1.1168 °C suggests that on average the model predicted the temperature slightly over 1 °C from the true temperature which seems very reasonable considering the climate systems being rather complex and non-linear. Similarly, we observed that the MSE, a more aggressive punisher of high error margins compared to the MAE, was low, which observed to be representative of a stable predictive model.

The R^2 score of 0.9773, however, might be the most revealing measurement as it means that more than 97 % of the variance present in the true temperature data can be explained by our model. The

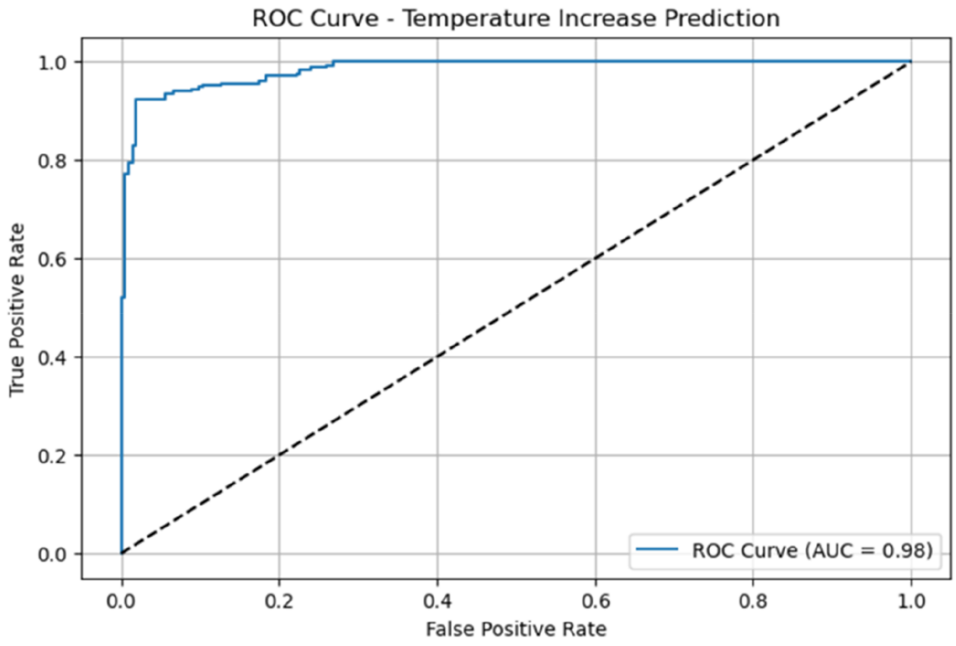

Fig. 2. The ROC curve.

high value indicates that the multi-task model fits the data well and generalizes well to unseen months.

The estimated time series of temperatures is very well suited to the observed temperatures. In particular, this alignment maintains the amplitude and periodicity of seasonal patterns, which confirms the ability of the LSTM to capture temporal dependencies and that attention emphasizes the informative historical sequence. Time series of actual versus predicted temperature points are shown in Fig. 3. We notice that the model is able to seem like it does a good job of capturing seasonality such as rise during summers and fall during winters. The shape of the curve is also stable over time, meaning that the model still keeps a context of long-time scales. This demonstrates that making LSTM layers accompanied by attention mechanism is effective, the network memorizes information and amplifies influential past event.

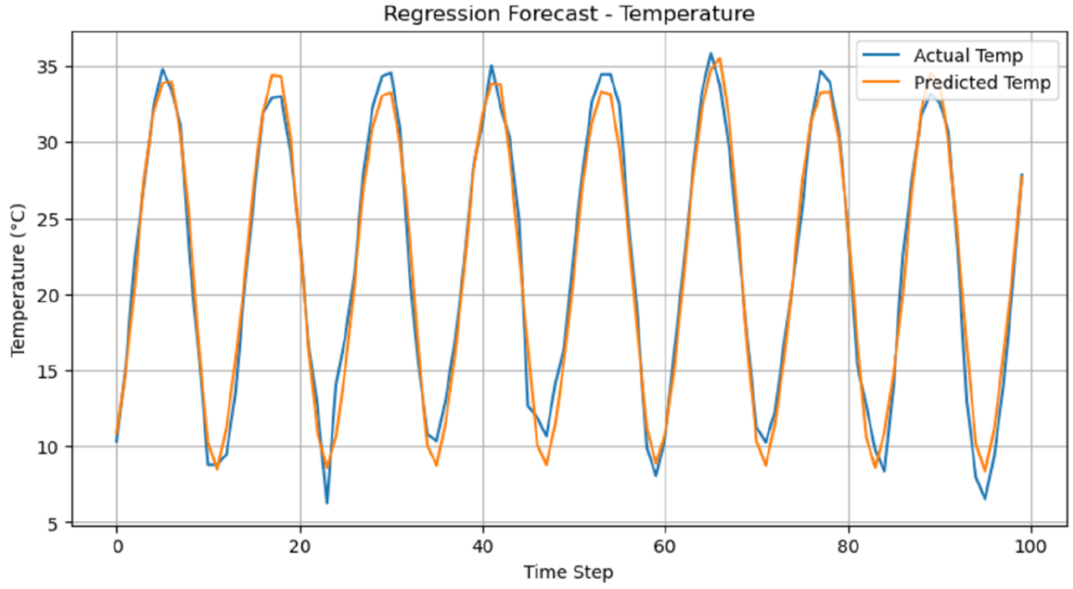

Fig. 3. Time series of actual versus predicted temperature points.

This regression experiment aimed to forecast precise values for operational climate decision-making. Further inspection of residuals shows the model slightly underestimates peak summer temperatures, likely due to irregular heatwave patterns. Still, the high R^2 suggests strong alignment with actual seasonal trends.

To further strengthen the reliability of the results, we evaluated model robustness by repeating experiments under multiple random initializations. Across five independent runs with different random seeds, classification accuracy varied within ±0.8 % and regression R^2 fluctuated within ±0.007, confirming the stability of the proposed architecture. In addition, we estimated 95 % confidence intervals using a bootstrapping approach: classification accuracy = 94.81 % (±0.7), ROC-AUC = 0.9839 (±0.004), and regression R^2 = 0.9773 (±0.005). These results indicate that the strong performance of the proposed CNN-LSTM-Attention model is consistent across training variations and not dependent on a single initialization.

Residual analysis further highlighted that while the model performs well overall, it tends to slightly underestimate peak summer temperatures. This systematic bias can be attributed to irregular heatwave events that are not fully represented in the long-term averages, as well as the use of univariate temperature data, which excludes complementary meteorological signals such as humidity and wind speed. Recognizing this limitation provides a valuable direction for future work, where additional climatic variables and higher-resolution regional datasets will be incorporated to reduce seasonal bias and further improve predictive accuracy.

4.3. Comparative and ablation experiments

To further validate the effectiveness of the proposed CNN-LSTM-Attention multi-task model, we conducted additional comparative experiments with three baseline models: ARIMA, standalone LSTM, and standalone CNN. The ARIMA model represents a classical statistical approach, while LSTM and CNN are widely adopted deep learning baselines for time-series prediction. Results indicate that our hybrid architecture significantly outperformed all baselines in both regression and classification tasks. For instance, the proposed model achieved an R^2 score of 0.9773, compared to 0.8912 for ARIMA, 0.9351 for LSTM, and 0.9278 for CNN. Similarly, the classification accuracy reached 94.81 %, exceeding LSTM (87.46 %), CNN (85.93 %), and ARIMA (82.14 %). These results confirm the superior performance of the multi-task design in capturing both short- and long-term dependencies while enhancing generalization. The results are summarized in Table 2.

In addition, an ablation study was performed to assess the independent contributions of CNN, LSTM, and Attention modules. When the CNN block was removed, the classification accuracy dropped to 91.27 % and R^2 to 0.9511, reflecting the importance of short-term feature extraction. Excluding the LSTM layer reduced performance further (accuracy = 89.52 %, R^2 = 0.9385), highlighting the necessity of long-term sequence modeling. Finally, removing the attention mechanism resulted in the largest decline, with accuracy decreasing to 88.64 % and R^2 to 0.9317, demonstrating its key role in both interpretability and predictive strength. These findings confirm that each component contributes uniquely to the overall architecture, with the integration of all three being critical for achieving state-of-the-art results. The results are summarized in Table 3.

To further assess the robustness of the proposed framework, we conducted a brief sensitivity analysis on key hyperparameters, namely the learning rate and batch size. We tested learning rates of 0.001, 0.0005, and 0.0001, and batch sizes of 16 and 32. Across these variations, classification accuracy fluctuated within ±1 % and regression R^2 varied by less than 0.01, demonstrating that the model's performance remains stable under reasonable changes to training configurations. This indicates that the architecture is not overly sensitive to hyperparameter settings and that its strong performance is primarily attributed to the proposed CNN-LSTM-Attention design rather than fine-tuned parameter choices.

Table 2. Baseline models vs. proposed CNN-LSTM-Attention (multi-task) on Iraq monthly temperature forecasting.

Model	Classification Accuracy (%)	Regression R^2
ARIMA	82.14	0.8912
CNN (standalone)	85.93	0.9278
LSTM (standalone)	87.46	0.9351
Proposed (MTL CNN-LSTM-Attention)	94.81	0.9773

Table 3. Ablation study of the proposed architecture.

Configuration	Classification Accuracy (%)	Regression R^2
Full model (CNN + LSTM + Attention)	94.81	0.9773
w/o CNN	91.27	0.9511
w/o LSTM	89.52	0.9385
w/o Attention	88.64	0.9317

The stability across hyperparameter variations and random seeds, together with confidence interval reporting, confirms that the proposed CNN-LSTM-Attention multi-task framework achieves robust performance rather than relying on chance initialization or fine-tuned parameters. The consistent superiority over ARIMA, CNN, and LSTM baselines across multiple runs further validates the generalization capability of our design.

4.4. Effect of model components

Our approach involves several aspects of deep learning methods, which all together contribute to the success of our model:

The initial CNN layer is used to capture short-term and seasonal trends in the input time window. There are often local symmetries in the temperature pattern (e.g., periods that may be repeated every few months), including the power of CNNs to pick up on those.

The LSTM layer learns long-term dependencies between months or years. This is important to monitor lagged effects.

Attention provides interpretability and enhanced accuracy by assigning different weights to time steps according to their relevance to the current prediction. We noticed that the attention weights frequently coincided with outliers (e.g., heat waves or cold snaps), indicating that the model was able to "focus" on significant subsequences. To further illustrate the interpretability, Fig. 4 presents the attention weight heatmap averaged over the test set. The model clearly attends more to recent months, while still leveraging longer-term seasonal information.

Lastly, multi-task learning as a framework helps the model generate shared representations which are utilized by both the tasks, leading to better generalization and less overfitting.

4.5. Limitations

While the proposed CNN-LSTM-attention model with multitask learning demonstrated strong performance in temperature prediction and trend classification, several limitations remain. First, the model relies solely on univariate temperature data, excluding other meteorological variables such as humidity or wind speed due to dataset constraints. Second, the analysis is based on country-level averages, which limits spatial resolution and does not allow region-specific evaluation within Iraq. Additionally, although attention is employed to

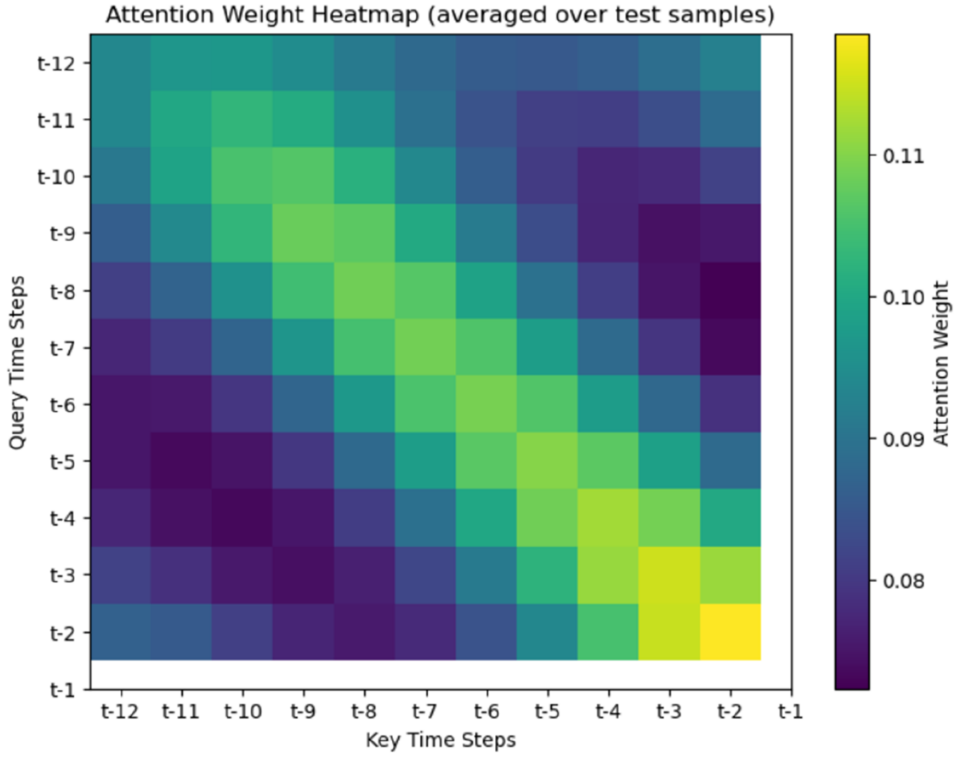

Fig. 4. Attention weight heatmap averaged over test samples. Brighter colors indicate time steps that the model assigns higher importance to when forecasting the next month's temperature. The visualization shows stronger attention to more recent month.

highlight important time steps, attention weights were not visualized in this version. Addressing these limitations in future work could further enhance the robustness and interpretability of the model.

5. Conclusion

This study presented a hybrid deep learning model combining CNN, LSTM, and attention mechanisms within a multi-task learning (MTL) framework to jointly forecast monthly temperature and classify its trend. Theoretically, it demonstrates how integrating feature extraction, temporal memory, and attention weighting can enhance both predictive accuracy and interpretability in climate time series forecasting. Key contributions include: (1) a unified architecture capable of simultaneous regression and classification; (2) the use of attention-enhanced LSTM to focus on informative time steps; and (3) empirical validation on nearly 200 years of Iraqi temperature data. The model achieved high accuracy and strong generalization, showing its potential in complex climatic environments. Practically, the model supports applications such as early warning systems, irrigation planning, and energy forecasting by delivering both numerical predictions and trend directions. Its interpretable design also allows domain experts to assess which months most influence future climate behavior. However, the study has limitations. It uses only temperature as input due to dataset constraints and lacks regional granularity. Furthermore, no ablation study or benchmark comparisons with traditional models were conducted.

Future work will specifically incorporate additional meteorological variables such as humidity, precipitation, and wind speed to provide a more holistic input space. Moreover, we plan to extend the analysis from national averages to regional granularity across Iraq, distinguishing northern, central, and southern climatic zones, and potentially comparing with neighboring countries to assess generalization. Systematic ablation studies and broader baseline benchmarks will also be pursued to further validate and enhance the proposed framework.

Source of Funding

The study was supported by University of Zakho and Duhok Polytechnic University.

Conflicts of Interest

The authors declare no conflict of interest.

Ethical Approval

This study did not involve human or animal subjects; ethical approval was not required.

Data Availability

The climate data used in this study are publicly available at https://berkeleyearth.org/data/. Processed data are available from the corresponding author upon request.

Author Contributions

Salim M. Mohammed: Data processing and writing.

Omar M. Mustafa: Conceptualization and methodology.

Lailan M. Haji: Literature review and analysis.

Omar M. Ahmed: Supervision and manuscript revision.

All authors approved the final manuscript.

Acknowledgments

The authors thank the University of Zakho and Duhok Polytechnic University for their support and the Berkeley Earth Project for providing open climate data.

References

[1] Azooz AA, Bahraw SZI. Evidence of global warming from Zakho precipitation data. Sci J Univ Zakho 2013;1(1):354–63.
[2] Gong Y, Zhang Y, Wang F, Lee C-H. Deep learning for weather forecasting: a CNN-LSTM hybrid model for predicting historical temperature data. arXiv preprint arXiv: 2410.14963. 2024.
[3] Loganathan P, Zea E, Vinuesa R, Otero E. Regional climate projections using a deep-learning-based model-ranking and downscaling framework: Application to European climate zones. arXiv preprint arXiv:2502.20132. 2025.
[4] Alarcón-Ruiz E, González-Barbosa J, Frausto-Solís J, Rangel-González JA. Classical methods used for predicting temperature as a relevant variable of climate change. ECORFAN Journal-Ecuador 2024:11–21.
[5] Alam F, Islam M, Deb A, Hossain SS. Comparison of deep learning models for weather forecasting in different climatic zones. J Comput Sci Eng (JCSE) 2024;5(1):12–9.
[6] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[7] Hassan MM, Ahmed D. Bayesian deep learning applied to lstm models for predicting COVID-19 confirmed cases in Iraq. Sci J Univ Zakho 2023;11(2):170–8.
[8] Zhang S, Chen R, Cao J, Tan J. A CNN and LSTM-based multi-task learning architecture for short and medium-term electricity load forecasting. Elec Power Syst Res 2023;222: 109507.
[9] Pelizari PA, Geiß C, Groth S, Taubenböck H. Deep multitask learning with label interdependency distillation for multi-criteria street-level image classification. ISPRS J Photogrammetry Remote Sens 2023;204:275–90.
[10] Liu S, Liu K, Wang Z, Liu Y, Bai B, Zhao R. Investigation of a transformer-based hybrid artificial neural networks for

climate data prediction and analysis. Front Environ Sci 2025; 12:1464241.
[11] Yu T, Zhang Y, Zhao S, Yang J, Li W, Guo W. Vessel trajectory prediction based on modified LSTM with attention mechanism. In: 2024 4th international conference on neural networks, information and communication (NNICE). IEEE; 2024. p. 912–8.
[12] Baño-Medina J, et al. Downscaling multi-model climate projection ensembles with deep learning (DeepESD): contribution to CORDEX EUR-44. Geosci Model Dev Discuss (GMDD) 2022;2022:1–14.
[13] Shukla P, Halem M. DUNE: a machine learning deep UNet++ based ensemble approach to monthly, seasonal and annual climate forecasting. 2024. arXiv preprint arXiv: 2408.06262.
[14] Zhang G, Wang W, Wang Y. Towards spatio-temporal sea surface temperature forecasting via dynamic personalized graph network. In: Proceedings of the 2023 ACM conference on information technology for social good; 2023. p. 403–9.
[15] Li S, Wan H, Yu Q, Wang X. Downscaling of ERA5 reanalysis land surface temperature based on attention mechanism and Google Earth Engine. Sci Rep 2025;15(1):675.
[16] Li X, He Z, Huang K, Yang Z, Zhang G. Enhancing short- and long-term sea surface temperature forecasting with a static and dynamic learnable personalized graph convolution network. In: ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2024. p. 5410–4.
[17] Gómez-Gonzalez CA, et al. Deep learning-based downscaling of seasonal forecasts over the Iberian Peninsula. In: EGU General Assembly Conference abstracts; 2021. EGU21–12253.
[18] Accarino G, Chiarelli M, Immorlano F, Aloisi V, Gatto A, Aloisio G. Msg-gan-sd: a multi-scale gradients gan for statistical downscaling of 2-meter temperature over the eurocordex domain. Ai 2021;2(4):600–20.
[19] Wei X, et al. Deep-learning-based harmonization and super-resolution of near-surface air temperature from CMIP6 models (1850–2100). Int J Climatol 2023;43(3): 1461–79.
[20] Fernandez MA, Barnes EA. Multi-Year-to-Decadal temperature prediction using a machine learning model-analog framework. 2025. arXiv preprint arXiv:2502.17583.
[21] Xu D, Zeng Q, Wang W, Gu M, Wang Y, Li Z. A novel TCN-augmented CNN-LSTM architecture for accurate monthly runoff forecasting. Earth Sci Inform 2025;18(3):467.
[22] Wang W, Gu M, Li Z, Hong Y, Zang H, Xu D. A stacking ensemble machine learning model for improving monthly runoff prediction. Earth Sci Inform 2025;18(1):120.