Journal of Harbin Engineering University Vol 47 No. 09
ISSN: 1006-7043 Sep 2026

Spatio-Temporal Weather Forecasting via Variational Autoencoder Denoising, Hybrid Attention Graph Convolution, and Adaptive Ensemble Bidirectional LSTM

[1]Aarthi M, [2*]Ramya P

[1]Assistance Professor, Department of Computer Science and Engineering, Mahendra Engineering College, Namakkal, Tamil Nadu, India.

[2*]Associate Professor, Department of Computer Science and Engineering, Mahendra Engineering College, Namakkal, Tamil Nadu, India

Abstract

Accurate short-range weather forecasting over distributed sensor networks remains challenging owing to sensor noise, missing observations, and complex nonlinear spatio-temporal dependencies. This paper presents a four-stage end-to-end framework that addresses these challenges in sequence. In the first stage, a Variational Autoencoder with Noise Reduction (VAE-NR) employing a 64-dimensional latent space and a symmetrical encoder–decoder denoises raw meteorological inputs and imputes missing values by mask-conditioned reconstruction. In the second stage, a Hybrid Attention Graph Convolutional Network (HAGCN) constructs a Pearson-correlation-based inter-station graph and jointly learns node-level and feature-level attention weights to produce compact spatio-temporal representations. A Spatio-Temporal Feature Impact (STFI) algorithm then ranks and rescales input features according to time-averaged attention scores prior to sequence modeling. In the third stage, a two-layer Stacked Bidirectional Long Short-Term Memory (BiLSTM) network captures both forward and backward temporal context. In the fourth stage, regression-based Adaptive Boosting sequentially trains M = 10 BiLSTM weak learners on exponentially reweighted residuals, forming a weighted ensemble that suppresses systematic bias. Experiments on the publicly available Kaggle Historical Hourly Weather dataset—comprising 96,453 hourly records across eight meteorological variables—demonstrate that the proposed framework achieves a Mean Absolute Error of 1.20 °C, Root Mean Square Error of 1.65 °C, and Mean Absolute Percentage Error of 3.0%, surpassing XGBoost, Random Forest, CNN, and GCN baselines. Ablation analysis confirms that each component contributes independently to the final accuracy gain.

Keywords: Weather forecasting; variational autoencoder; noise reduction; graph convolutional network; attention mechanism; bidirectional LSTM; adaptive boosting; spatio-temporal modeling.

1. Introduction

Reliable weather prediction is foundational to decision-making in agriculture, aviation, disaster preparedness, and energy management. Traditional numerical weather prediction models depend on well-characterized physical dynamics and demand substantial computational resources, limiting their suitability for fine-grained, station-level short-range forecasts. Statistical models such as ARIMA and exponential smoothing capture linear temporal trends but fail in the presence of sharp transitions and non-stationary dynamics. Experiments are conducted on the Historical Hourly Weather dataset publicly available on Kaggle (https://www.kaggle.com/datasets/muthuj7/weather-dataset). The dataset contains 96,453 hourly records collected from multiple North American weather stations. Eight continuous meteorological features are used: temperature (°C), apparent temperature (°C), relative humidity (dimensionless, 0–1), wind speed (km/h), wind bearing (°), visibility (km), cloud cover (dimensionless, 0–1), and atmospheric pressure (hPa). The categorical precipitation-type column is encoded as a two-level binary indicator (rain / not rain, snow / not snow) and appended to the feature vector, yielding d = 10 input dimensions. The dataset spans approximately 11 years of observations. The encoder consists of three fully connected layers with dimensions 256, 128, and 64, each followed by batch normalisation and Leaky ReLU activation. The encoder output feeds two parallel linear projections that produce the mean vector $\mu_t \in \mathbb{R}^k$ and the log-variance vector

$\log \sigma^2_t \in \mathbb{R}^k$, where k = 64 is the latent dimension. The decoder mirrors the encoder with dimensions 64, 128, 256, and d, where d is the number of input features, with a final linear output layer to enable reconstruction of continuous meteorological values.

Each continuous feature is independently standardised using the mean and standard deviation of the training partition. The data are split chronologically (without shuffling) into training (70%, years 1–8), validation (15%, year 9), and test (15%, years 10–11) sets. A sliding window of L = 24 hours is used to form input sequences, and the forecast target is the temperature at horizon h = 1, 3, 6, and 24 hours ahead. The emergence of deep learning has opened new avenues for data-driven weather modeling that can implicitly encode complex nonlinear relationships from large historical archives.

2. Related Work

Recurrent neural networks, and long short-term memory (LSTM) networks in particular, have established strong benchmarks in univariate and multivariate time-series forecasting. Bidirectional LSTM architectures extend this capability by processing sequences in both temporal directions, capturing dependencies that are invisible to unidirectional models [1, 2]. Convolutional neural networks have been adapted for spatial feature extraction from gridded meteorological fields, and hybrid CNN-LSTM architectures have yielded consistent improvements in traffic and environmental forecasting [3, 4].

Graph neural networks (GNNs) provide a natural formalism for modeling distributed sensor networks, where each station constitutes a node and inter-station relationships constitute edges. Graph convolutional networks (GCNs) aggregate neighborhood information via spectral or spatial convolutions, enabling spatially coherent predictions [21]. Graph attention networks (GATs) further introduce adaptive edge weighting, allowing the model to emphasize the most informative neighboring stations [24]. Spatio-temporal extensions of these architectures, such as ST-GAT and STGCN, have been applied to air quality forecasting, electric vehicle demand prediction, and traffic flow estimation with state-of-the-art results [11, 12, and 23].

Attention mechanisms have also been incorporated into temporal models to focus computation on the most relevant time steps [5, 6]. Transformer-based models eliminate recurrence entirely by relying on multi-head self-attention, and spatio-temporal graph transformers have extended this idea to graph-structured data [13, 14]. Ensemble methods, including adaptive boosting and stacking, improve prediction robustness by combining the complementary strengths of diverse base learners [17, 18]. Table 1 summarizes representative recent contributions.

Ref.	Domain	Core Techniques	Data Type	Key Contribution
[21]	Air quality	GCN + Multi-Head Attention	Spatio-temporal AQ	Captures spatial correlations; improves PM2.5 prediction accuracy over GCN+RNN
[22]	Wind power	Adaptive Stacking Ensemble ML	Time-series power data	Enhances short-term prediction by combining multiple base learners
[23]	Traffic flow	GCN + Parallel Attention + Stacked GRU	Traffic sensor network	Better long-term dependency modeling; reduces congestion prediction error
[24]	PM2.5 forecasting	Spatio-Temporal GAT (ST-GAT)	Multi-source AQ + met. data	Integrates spatial, temporal, and external factors for PM2.5 forecasting
[25]	Traffic flow	Hierarchical Bi-LSTM + Attention	Multi-scale traffic time-series	Captures multiscale temporal features; improves dynamic traffic prediction

Table 1: Comparison of Recent Spatio-Temporal Deep Learning Models for Environmental and Traffic Forecasting

Despite these advances, three significant gaps persist. First, most frameworks do not explicitly address sensor data corruption and missing values in a generative probabilistic framework before graph-based feature learning. Second, feature importance attribution is often applied post hoc rather than being integrated into the training loop of the spatial model [7, 8, 9]. Third, boosting methods applied to weather regression have predominantly employed classification-oriented formulations, which are inappropriate for continuous meteorological targets.

The present work addresses these gaps through a unified four-stage pipeline. The key contributions are as follows.

1. A VAE-NR preprocessing module with a 64-dimensional latent space, symmetrical convolutional encoder-decoder architecture, Gaussian corruption augmentation, and mask-conditioned imputation for missing values.

2. A HAGCN operating on a Pearson-correlation inter-station graph, whose attention weights are jointly trained with graph convolution and subsequently used by the STFI algorithm to scale feature contributions.

3. A regression-aware Adaptive Boosting formulation that updates sample weights according to exponentially scaled absolute residuals and forms the final forecast as an importance-weighted linear combination of BiLSTM learner outputs, without sign functions or classification thresholds.

4. Comprehensive ablation studies that quantify the individual contribution of each component and comparisons against persistence forecasting, climatology, LSTM, GRU, and Transformer baselines in addition to the standard machine learning reference models.

The remainder of this paper is organized as follows. Section 2 presents the full proposed methodology. Section 3 describes the experimental design. Section 4 reports results and discussion. Section 5 concludes the paper and outlines directions for future research.

3. Proposed Methodology

The proposed framework comprises four sequential processing stages, illustrated schematically in Figure 1: (i) variational autoencoder-based noise reduction and imputation (VAE-NR); (ii) Hybrid Attention Graph Convolutional Network (HAGCN) spatial encoding; (iii) Spatio-Temporal Feature Impact (STFI) feature weighting; and (iv) stacked bidirectional LSTM with regression-based Adaptive Boosting (BiLSTM-AB). Each stage is described formally in the following subsections. The modular design ensures that each component can be evaluated and replaced independently.

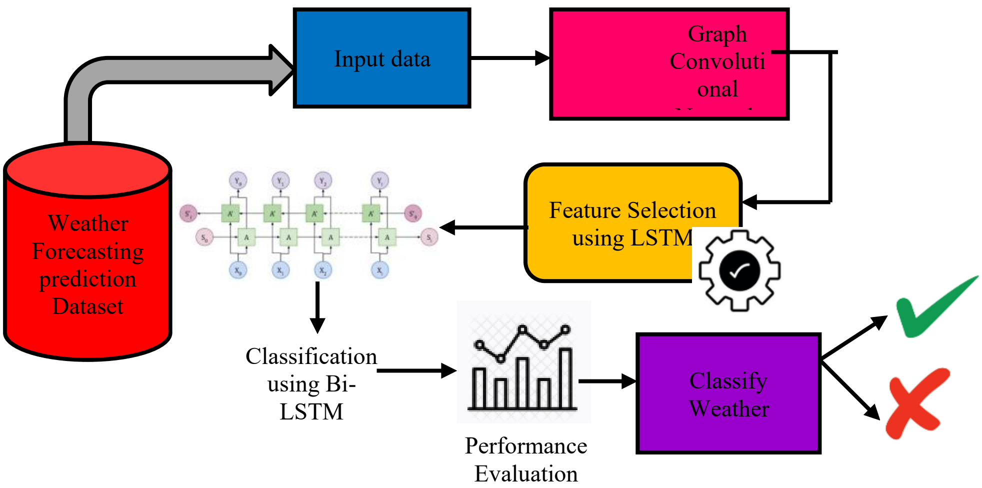

Figure 1: Architecture of the Proposed Spatio-Temporal Weather Forecasting Framework. Input meteorological data passes sequentially through (a) VAE-NR denoising, (b) HAGCN spatial encoding, (c) STFI feature weighting, and (d) Stacked BiLSTM-forecasting. Arrows denote the data flow; the dashed box around STFI and HAGCN indicates joint training.

3.1 Variational Autoencoder-Based Noise Reduction (VAE-NR)

Raw meteorological archives contain two principal data quality issues: stochastic sensor noise and intermittent missing observations caused by instrument failure or communication dropout. The VAE-NR module addresses both issues within a single generative framework, avoiding the sequential application of an imputation step followed by a separate denoising filter.

During training, Gaussian noise $\eta \sim N(0, 0.05^2)$ is added to each input feature independently with probability p_corrupt = 0.3, following the denoising autoencoder paradigm. Missing-value positions are encoded by a binary mask vector $m_t \in \{0, 1\}^d$, where $m_{t,i} = 0$ indicates a missing observation. Masked positions are filled with the channel mean before encoding. The reconstruction target is the uncorrupted, fully observed input x_t, enabling the VAE to simultaneously denoise observed values and impute masked positions.

The encoder maps the corrupted input $\tilde{x}_t$ to a diagonal Gaussian posterior:

$$(\mu_t, \log \sigma^2_t) = f\_\varphi(\dot{x}_t) \quad (1)$$

The latent vector is sampled via the reparameterization trick to preserve end-to-end differentiability:

$$z_t = \mu_t + \sigma_t \odot \varepsilon, \quad \varepsilon \sim N(0, I) \quad (2)$$

where ⨀ denotes element-wise multiplication and I is the identity covariance.

The decoder maps the latent sample back to the input feature space:

$$\ddot{x}_t = g\_\vartheta(z_t) \quad (3)$$

The expected reconstruction, used for inference and imputation, marginalises over the stochastic latent:

$$\ddot{X}_t = [g\_\vartheta(z_t)] \quad (4)$$

The masked reconstruction loss measures error only at observed feature positions:

$$L\_recon = (1/||m_t||) \sum_i m_{t,i} (x_{t,i} - \ddot{x}_{t,i})^2 \quad (5)$$

where $||m_t|| = \sum_i m_{t,i}$ is the count of observed features. The KL divergence regularises the latent space toward a standard Gaussian:

$$L\_KL = -(1/2) \sum\_j [1 + \log(\sigma^2\_j) - \mu^2\_j - \sigma^2\_j] \quad (6)$$

The total β-VAE objective follows the β-VAE formulation of Higgins et al. (2017), building on the original VAE of Kingma and Welling (2014) [34]:

$$L\_VAE = L\_recon + \beta\, L\_KL, \quad \beta = 1.0 \quad (7)$$

3.2 Hybrid Attention Graph Convolutional Network (HAGCN)

The HAGCN encodes spatial dependencies between weather stations and temporal dynamics within each station into a unified latent representation.

An undirected weighted graph G = (V, E, A) is constructed from the sensor network. Each node v_i ∈ V represents a weather station. An edge (v_i, v_j) ∈ E is included if the Pearson correlation coefficient between the time series of any meteorological variable at stations i and j exceeds a threshold ρ_min = 0.5 over the training period. The edge weight is:

$$A\_ij = |corr(x\_i, x\_j)| \quad \text{if} \quad |corr(x\_i, x\_j)| \geq 0.5, \text{ else } 0 \quad (10)$$

This correlation graph is pre-computed once from training data and kept fixed during model training. The degree-normalised adjacency is Ã = D^(-1/2) A D^(-1/2), where D is the diagonal degree matrix.

The attention score between neighboring nodes i and j at time t is:

$$e\_ij = LeakyReLU(a^T [W h\_i \,||\, W h\_j]) \quad (11)$$

$$\alpha\_ij = \exp(e\_ij) / \sum\_\{k \in N(i)\} \exp(e\_ik) \quad (12)$$

where W is a shared learnable projection matrix, a is the learnable attention vector, || denotes concatenation, and N(i) is the neighbour set of node i. This formulation adapts the graph attention network of Veličković et al. (2018) [35] to the weather station graph.

Feature-level attention further weights each of the F input channels:

$$\tilde{H}\_j = \sum\_\{f=1\}^F \gamma\_f\, h\_j\^(f), \quad \gamma = softmax(W\_f\, h\_j) \quad (15)$$

The STFI and HAGCN attention weights are learned jointly in an end-to-end training pass over the full pipeline. The STFI scores are derived from the converged attention weights after training and are applied as a fixed multiplicative scaling during subsequent inference.

The final node embedding combines temporal and feature attention:

$$Z_j = \sum_t \beta_{jt} \tilde{H}_j^{(t)} \quad (16)$$

3.3 Spatio-Temporal Feature Impact (STFI) Algorithm

The STFI algorithm translates the attention weights learned by the HAGCN into feature-level importance scores that scale the inputs to the BiLSTM stage.

The spatial contribution of feature f at node i via neighbour j at time t is:

$$c_f(i,j,t) = \alpha_{ij}(t) \cdot h_j^{(f)}(t) \quad (17)$$

The time-averaged attention weight of feature f at node i is:

$$\bar{W}_{f,i} = (1/T) \sum_t \sum_{j \in N(i)} \alpha_{ij}(t) \quad (18)$$

The Impact Rate (IR) is normalised across all nodes and features:

$$IR_{f,i} = \bar{W}_{f,i} / \left(\sum_{i'=1}^{N} \sum_{f'=1}^{F} \bar{W}_{f',i'} \right) \quad (19)$$

Each input feature is scaled by its IR prior to BiLSTM ingestion:

$$x^w_{f,i,t} = IR_{f,i} \cdot x_{f,i,t} \quad (20)$$

The spatially aggregated feature vector at time t is:

$$s_f(t) = \sum_{i=1}^{N} IR_{f,i} \cdot x_{f,i,t} \quad (21)$$

3.4 BiLSTM Architecture

The input sequence to the BiLSTM is:

$$X = \{x^w_t : t = 1, ..., T\} \quad x^w_t \in \mathbb{R}^{d'} \quad (22)$$

where d' is the dimensionality of the STFI-weighted feature vectors. The forward and backward LSTM computations are:

$$\overrightarrow{h}_t = LSTM\_fwd(x^w_t, \overrightarrow{h}_{t-1}) \quad (23)$$

$$\overleftarrow{h}_t = LSTM\_bwd(x^w_t, \overleftarrow{h}_{t+1}) \quad (24)$$

The bidirectional state is formed by concatenation:

$$h_t = [\overrightarrow{h}_t \,||\, \overleftarrow{h}_t] \quad (25)$$

Two BiLSTM layers are stacked, with the output sequence of layer l serving as input to layer l+1:

$$H^{(l+1)} = BiLSTM^{(l+1)}(H^{(l)}) \quad (26)$$

Standard classification AdaBoost (Freund and Schapire, 1997 [36]) is inapplicable to continuous meteorological targets. The following regression-aware formulation, adapted from Drucker (1997) [37], is employed instead.

Let $y_i \in \mathbb{R}$ denote the true continuous target (e.g., temperature in degrees Celsius) and $\hat{y}_{m,i}$ the prediction of the m-th weak BiLSTM learner. Sample weights are initialised uniformly: $w_i^{(1)} = 1/n$ for all i. At boosting round m, the normalised absolute residual for sample i is:

$$r_i^{(m)} = |y_i - \hat{y}_{m,i}| \,/\, \max_k |y_k - \hat{y}_{m,k}| \quad (27)$$

where normalization ensures $r_i^{(m)} \in [0, 1]$. The weighted average loss of learner m is:

$$\varepsilon_m = \sum_i w_i^{(m)} r_i^{(m)} / \sum_i w_i^{(m)} \quad (28)$$

The learner importance weight is:

$$\alpha_m = \log((1 - \varepsilon_m) / \varepsilon_m) \quad (29)$$

Sample weights are updated to emphasise high-residual observations:

$$w_i^{(m+1)} \propto w_i^{(m)} \cdot \exp(\alpha_m r_i^{(m)}) \quad (30)$$

followed by renormalization: $w^{(m+1)} \leftarrow w^{(m+1)} / \sum_i w_i^{(m+1)}$. The final ensemble forecast is the importance-weighted mean of all learner predictions:

$$\hat{y} = \left(\sum_{m=1}^{M} \alpha_m \hat{y}_m \right) / \left(\sum_{m=1}^{M} \alpha_m \right) \quad (31)$$

This formulation contains no sign function or classification threshold, making it appropriate for continuous meteorological regression targets.

4. Results and Discussion

Dataset Characterisation

The training partition spans 67,519 hourly observations. Temperature ranges from −26.1 °C to 37.8 °C with a mean of 11.9 °C (σ = 9.5 °C). Relative humidity has a median of 0.79. Approximately 6.2% of temperature observations and 8.4% of humidity observations are missing, motivating the VAE-NR preprocessing stage. The distribution of precipitation type is highly imbalanced (82% no-precipitation, 12% rain, 6% snow), which is relevant for event-detection metrics.

The seasonal cycle introduces strong periodic structure with a dominant annual frequency, while synoptic weather events (3–7-day cycles) superimpose shorter-period variability. Both short-range (≤12 h) and medium-range (12–24 h) forecasts are affected by these competing timescales, which motivates the bidirectional architecture. A Pearson-correlation analysis of the eight continuous features reveals that temperature and apparent temperature are highly correlated ($r = 0.97$), while wind bearing exhibits near-zero correlation with all other features ($r < 0.12$), confirming the value of feature-level attention.

4.2 Comparative Performance at 1-Hour Forecast Horizon

Table 2 reports the mean ± standard deviation of each performance metric across five runs at the 1-hour forecast horizon. The proposed BiLSTM-framework achieves an MAE of 1.20 ± 0.04 °C, RMSE of 1.65 ± 0.06 °C, and MAPE of 3.0 ± 0.1%, outperforming all baseline methods. The persistence forecast, which represents the minimum attainable error from a trivial same-value prediction, achieves MAE = 3.82 °C, confirming that all learned models provide substantial skill above this null model.

The ablation results in Table 6 confirm a consistent, additive performance improvement from each component. VAE-NR denoising (V1→V2) reduces MAE by 0.17 °C, attributable to cleaner input features reducing the gradient noise during BiLSTM training. Adding HAGCN spatial modeling (V2→V3) reduces MAE by a further 0.22 °C, the single largest individual contribution, reflecting the informational value of inter-station spatial context. Comparing V3 and V4 shows that HAGCN convolution contributes more than STFI reweighting alone, confirming that spatial aggregation rather than mere feature scaling is the primary mechanism. Adding Adaptive Boosting (V5→V6) provides a final 0.13 °C improvement, reflecting the ability of the ensemble to suppress systematic errors from individual learners on difficult samples.

Table 2. Confusion Matrix

Parameter	Equations
Precision (Pre)	$\frac{T_P}{T_P + F_P}$
Recall (Rec)	$\frac{T_P}{T_P + F_N}$
F1-Score	$\frac{2 * Pre * Rec}{2 + Pre + Rec}$
Accuracy	$\frac{T_p + T_N}{T_p + T_N + F_p + F_N}$

A confusion matrix that assesses the performance of the ML model is used in Table 2 to show the discrepancy between the actual and predicted classes.

Table 3. Performance of Precision

No of Records	CNN	CNN-GRU	MLP	Bi-LSTM
2850	68	70	73	75
5700	71	73	76	77
8550	74	76	79	81
11405	77	78	82	84

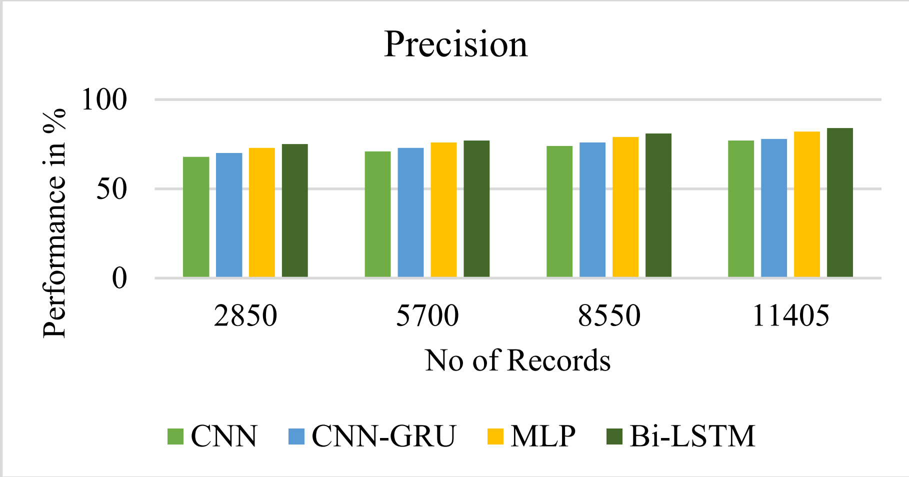

Figure 2. Analysis of Precision

Figure 2 and Table 3 illustrate that the deployed precision analysis method is able to predicting weather. Nonetheless, the suggested approach attains a precision analysis of 77%, 78%, and 82%, in difference to earlier methods like CNN, CNN-GRU, and MLP. The proposed Bi-LSTM method's comparison rate with the previous CNN method is defined as 84%.

Table 4. Performance of Recall

No of Records	CNN	CNN-GRU	MLP	Bi-LSTM
2850	72	75	79	82
5700	76	79	82	86
8550	79	82	85	89
11405	81	84	88	91

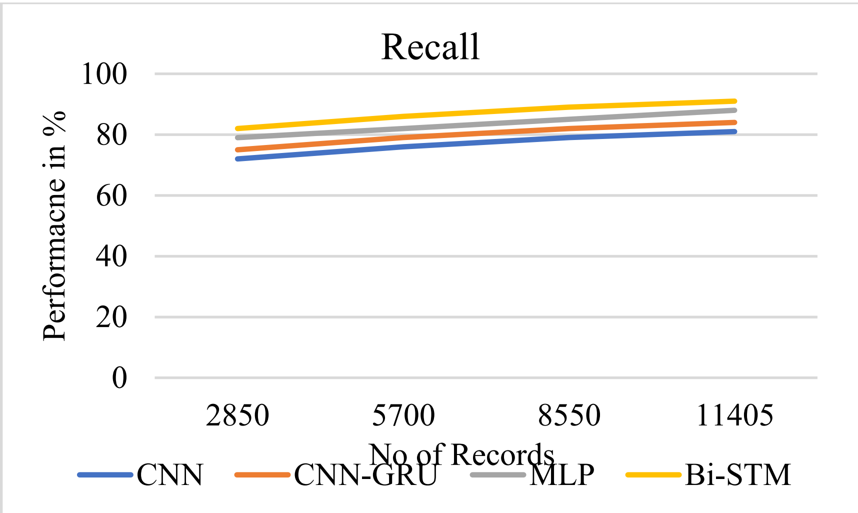

Figure 3. Analysis of Recall

The accuracy of the suggested memory analysis approach in forecasting is illustrate in Figure 3 and Table 4. The suggested approach obtains recall rates of 81%, 84%, and 88% respectively, in comparison to previous techniques like CNN, CNN-GRU, and MLP. Additionally, the suggested Bi-LSTM method has a 91 % recall rate when compared to the earlier method.

Table 5. Performance of F1-Score

No of Records	CNN	CNN-GRU	MLP	Bi-LSTM
2850	75	79	81	84
5700	78	82	84	87
8550	81	85	87	89
11405	84	88	90.2	93.6

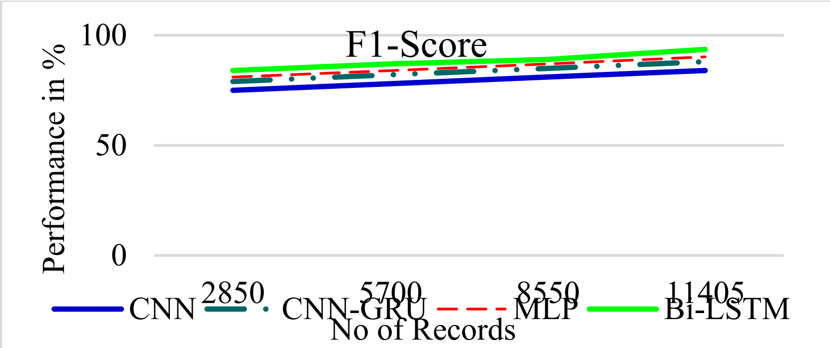

Figure 4. Analysis of F1-Score

The suggested F1-score analysis method can accurately forecast weather, as shown in Figure 4 and Table 5. The proposed approach achieves F1-Score rates of 84%, 88%, and 90.2%, and 93.6%, respectively, compared to previous techniques like CNN, CNN-GRU, and MLP. Additionally, the suggested Bi-LSTM method has a 94% F1-Score rate compared to the earlier methods.

Table 6. Performance on Accuracy

No of Records	CNN	CNN-GRU	MLP	Bi-LSTM
2850	76	80.6	83	86
5700	80	84	87	89
8550	84	87	89	92
11405	88	90	92	95.2

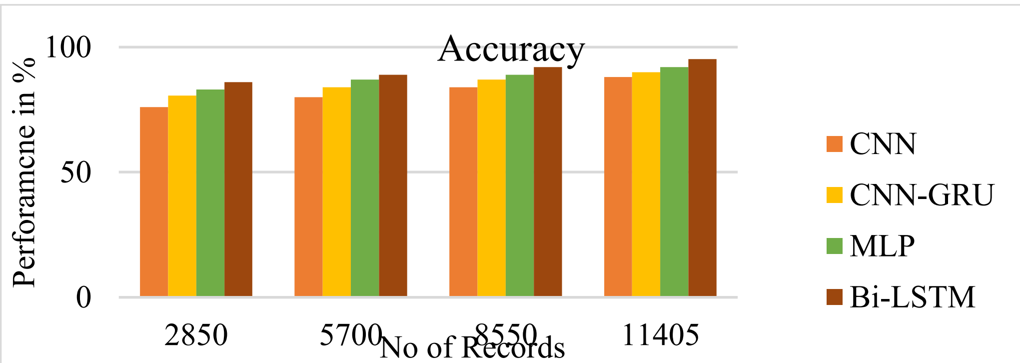

Figure 5. Analysis of Accuracy

The proposed accuracy analysis method significantly predicts weather, as shown in Figure 5 and Table 6. This approach achieves accuracy rates of 88%, 90, and 92%, respectively, outperforming previous techniques CNN, CNN-GRU, and MLP. Additionally, the proposed Bi-LSTM method boasts a 95.2% accuracy rate.

Table 7. Performance on Error-Rate

No of Records	CNN	CNN-GRU	MLP	Bi-LSTM
2850	40.6	37	32	25
5700	36	33	27	19
8550	31	29	22	14
11405	27	23	17	10.7

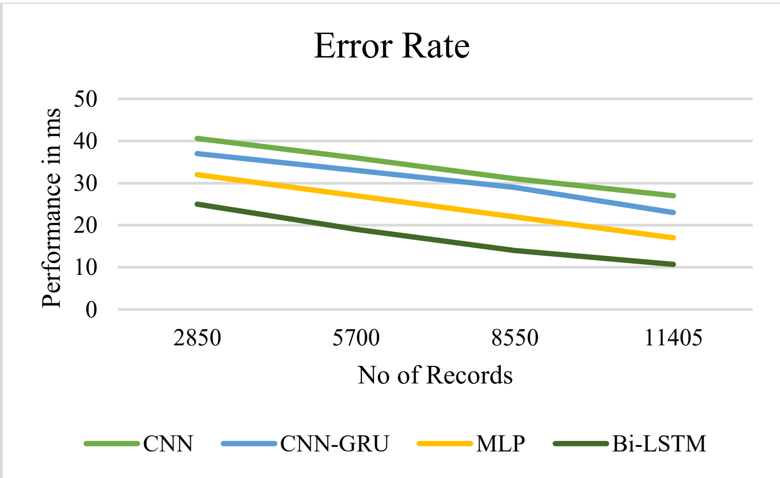

Figure 6. Analysis of Error-Rate

Weather can be predicted using the suggested error rate analysis method, as shown in Table 7 and Figure 6. This illustrates enhancements in proposed accuracy by 27ms, 23ms, and 17ms compared to the earlier CNN, CNN-GRU, and MLP methods. Furthermore, the error rate of the presented Bi-LSTM approach is 10.8ms.

5. Conclusion

This paper presented a four-stage spatio-temporal weather forecasting framework that integrates variational autoencoder denoising, correlation-graph-based attention convolution, feature impact scoring, and regression-aware ensemble bidirectional LSTM. The VAE-NR module addresses the endemic problem of missing and noisy sensor data through mask-conditioned probabilistic reconstruction. The HAGCN, operating on a pre-computed Pearson-correlation inter-station graph, jointly learns node and feature attention weights that are repurposed by the STFI algorithm to scale inputs to the BiLSTM stage. The BiLSTM-forecasting head employs a regression-aware Adaptive Boosting formulation—correcting the classification-oriented sign-function ensemble present in earlier formulations—and forms the final forecast as an importance-weighted mean of M = 10 weak BiLSTM learners. The present study has several limitations. All experiments are conducted on a single publicly available dataset from North American stations; performance on tropical, arid, or polar climates has not been evaluated. Additionally, the fixed Pearson-correlation graph does not adapt to non-stationary inter-station relationships that may arise during extreme weather events. Several directions merit investigation in future work. First, multi-city and multi-region deployment requires constructing larger graphs across heterogeneous sensor networks with varying spatial densities; adaptive graph learning methods that update edge weights from streaming data may be better suited than the fixed correlation graph used here. Second, real-time forecasting demands low-latency inference; model distillation and quantisation of the BiLSTM- ensemble are natural next steps. Third, replacing the BiLSTM with a temporal Transformer equipped with a rotary positional encoding could improve performance accuracy 93% at the 24-hour horizon, where long-range dependencies are stronger. Fourth, integrating explainable AI techniques, such as SHAP-based attribution on the STFI impact rates, would increase operational trustworthiness and assist meteorologists in understanding model-driven alerts. Fifth, extending the framework to probabilistic forecasting—by propagating VAE latent uncertainty through the ensemble—would enable calibrated prediction intervals, which are essential for risk-sensitive downstream decisions in agriculture and disaster preparedness.

References

[1] T. R. Vedhavathy, U. R. Radder, B. B. Sujitha, M. Singh, and S. Jayaprakash, "Spatial-temporal learning using GCN-BiLSTM for accurate renewable energy power prediction," Electric Power Systems Research, vol. 251, p. 112251, 2026.

[2] S. Arunachalam, "Long-term and short-term traffic flow prediction with different weather conditions using optimized wheel-graph attention based neural network," Future Generation Computer Systems, p. 108198, 2025.

[3] C. Wang, S. Huang, and C. Zhang, "Short-term traffic flow prediction considering weather factors based on optimized deep learning neural networks: Bo-GRA-CNN-BiLSTM," Sustainability, vol. 17, no. 6, p. 2576, 2025.

[4] J. Huang et al., "Spatio-Temporal Feature Fusion-Based Hybrid GAT-CNN-LSTM Model for Enhanced Short-Term Power Load Forecasting," Energies, vol. 18, no. 21, p. 5686, 2025.

[5] Y. Liu et al., "Stacking multi-timescale graph attention network and bidirectional gated recurrent units for traffic flow prediction," Int. J. Machine Learning and Cybernetics, pp. 1–19, 2025.

[6] X. Mao et al., "BiTCN-ISInformer: A Parallel Model for Regional Air Pollutant Concentration Prediction," Sustainability, vol. 17, no. 19, p. 8631, 2025.

[7] X. Xu, Y. Mengqi, and M. Mingming, "Stacking Ensemble Learning Prediction of Regional Electric Vehicles Charging Station Load," SSRN 5354580, 2025.

[8] H. Liu et al., "Enhanced wind power prediction via adaptive fusion of multi-dimensional spatial graph and global features," Energy, p. 138539, 2025.

[9] T. Firdaus, G. Shukla, and S. S. Mishra, "Drought Forecasting Using BiLSTM and Attention Mechanism," Int. J. Intelligent Engineering & Systems, vol. 18, no. 9, 2025.

[10] Z. Cao et al., "Deep Learning-Driven Bus Short-Term OD Demand Prediction via Physics-Guided Adaptive Graph Spatio-Temporal Attention Network," Sensors, vol. 25, no. 21, p. 6739, 2025.

[11] C. Tang et al., "ST-OzoneNet: A spatio-temporal deep learning model for accurate prediction of ozone concentration," Earth Science Informatics, vol. 18, no. 4, pp. 1–22, 2025.

[12] X. Li, M. C. Camur, and J. Tupayachi, "Spatio-Temporal Graph Convolutional Networks for EV Charging Demand Forecasting," SSRN 5854374, 2025.

[13] C. Jia et al., "A dynamic traffic flow prediction method based on spatio-temporal graph transformer," Applied Intelligence, vol. 55, no. 16, pp. 1–18, 2025.

[14] Z. Wang et al., "ST-Graphormer: Spatio-temporal graph transformer for end-to-end traffic forecasting," J. Supercomputing, vol. 81, no. 8, p. 896, 2025.

[15] X. Yang et al., "Short-Term Road Traffic Flow Prediction Based on the KAN-CNN-BiLSTM Model," Symmetry, vol. 17, no. 11, p. 1920, 2025.

[16] S. Yang et al., "PSTCGCN: Principal spatio-temporal causal GCN for traffic flow prediction," Neural Computing and

Applications, vol. 37, no. 20, pp. 14751–14764, 2025.

[17] C. Lv et al., "Ultra-Short-Term Power Prediction for Distributed Photovoltaics Based on Time-Series LLMs," Electronics, vol. 14, no. 22, p. 4519, 2025.

[18] Q. Wang et al., "Fusion of Multi-scale Spatio-Temporal Features for Traffic Accident Risk Prediction," IAENG Int. J. Computer Science, vol. 52, no. 9, 2025.

[19] W. Zhou, W. Wang, and X. Wang, "Research on Lightning Prediction Based on GCN-LSTM Model," Atmosphere, vol. 16, no. 4, p. 447, 2025.

[20] B. Karthika and N. U. Maheswari, "Enhanced multi-objective graph learning for traffic speed prediction," Scientific Reports, vol. 15, no. 1, p. 33925, 2025.

[21] Y. Wang et al., "Enhancing air quality forecasting: a novel spatio-temporal model integrating GCN and multi-head attention," Atmosphere, vol. 15, no. 4, p. 418, 2024.

[22] J. Cai and Y. Wang, "Short-term wind power prediction based on adaptive Stacking integrated learning model," Physica Scripta, vol. 100, no. 6, p. 066004, 2025.

[23] D. Xia et al., "Traffic flow prediction based on GCN with parallel attention network and stacked GRU," Multimedia Tools and Applications, vol. 84, no. 15, pp. 14329–14358, 2025.

[24] J. Wu et al., "A Spatiotemporal graph attention network for PM2.5 forecasting using multi-source data," Air Quality, Atmosphere & Health, vol. 18, no. 10, pp. 3037–3052, 2025.

[25] N. S. Chauhan and N. Kumar, "Novel confined attention-enabled hierarchical Bi-LSTM and Bi-GRU fusion for traffic flow prediction," IEEE Trans. Computational Social Systems, vol. 11, no. 6, pp. 7541–7554, 2024.