IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15690

BREATHE: AIR QUALITY PREDICTION USING EMBEDDED MACHINE LEARNING AND DEEP LEARNING MODELS WITH QUANTIZATION TECHNIQUES

Anandhu Suresh[1], Lekshmi V[2]

Student, Christ Nagar College, Kerala, India[1]

Assistant Professor, Christ Nagar College, Kerala, India[2]

Abstract: Air quality degradation poses a significant global health challenge, necessitating accurate and intelligent systems capable of real time pollutant forecasting and actionable wellness guidance for everyday users. The complexity of spatial and temporal pollutant dynamics across urban environments demands advanced deep learning architectures capable of multi city, multi pollutant prediction while remaining deployable on resource constrained mobile devices. This project presents Breathe, a cross platform air quality intelligence system employing a hybrid GCN-TransGRU architecture combining Graph Convolutional Networks for spatial inter city relationship modeling with Transformer encoders and Gated Recurrent Units for capturing temporal dependencies across PM2.5, PM10, and NO2.

To enable efficient mobile deployment, a Knowledge Distillation strategy compresses a high capacity teacher model of approximately 2.52M parameters into a lightweight student model of approximately 466k parameters into a lightweight student model of approximately 43k parameters, exported as an ONNX float32 model (~1.51 MB) and deployed via ONNX Runtime for on-device inference without significant accuracy loss. The system is implemented across three integrated components including a deep learning forecasting model, a modular NestJS backend API managing authentication, real time air quality data, and trip planning, and a Flutter based mobile application featuring a dynamic AQI dashboard, personalized health profiles, and resilience on network loss (last loaded data retained in memory). By combining spatial temporal deep learning with scalable cloud infrastructure, Breathe contributes to improved public awareness, reduced health risks, and the advancement of technology driven air quality management.

Keywords: Air Quality Index (AQI), Deep Learning Forecasting, Multi-Pollutant Prediction, Spatial-Temporal Modeling, Graph Convolutional Networks (GCN), Transformer-GRU Hybrid Architecture, Knowledge Distillation, Model Compression, On-Device Inference, ONNX Runtime, Cross-Platform Mobile Applications, Flutter, NestJS Backend API, Modular Architecture, Software Engineering.

I. INTRODUCTION

Air quality has become a pressing global concern, directly impacting human health and quality of life. Pollutants such as PM2.5, PM10, nitrogen dioxide, and ozone contribute to millions of premature deaths annually through respiratory and cardiovascular diseases. Traditional monitoring relies on fixed ground stations and chemical transport models, which are expensive, geographically limited, and unable to capture complex nonlinear pollutant dynamics. With advancements in deep learning, particularly hybrid spatial temporal architectures, accurate multi city pollutant forecasting has become achievable, enabling models to learn patterns from large scale environmental data that conventional approaches fail to capture.

The proposed system, Breathe: Air Quality Prediction Using Embedded Machine Learning and Deep Learning Models with Quantization Techniques, provides an intelligent solution for real time air quality monitoring and user wellness. It processes environmental data through a hybrid GCN-TransGRU architecture combining Graph Convolutional Networks for spatial inter city modeling with Transformer encoders and Gated Recurrent Units for temporal dependency capture across PM2.5, PM10, and NO2. Knowledge Distillation compresses the model into a student of ~466k parameters, exported as ONNX float32 (~1.51 MB) for efficient on-device mobile deployment. Implemented as a Flutter application supported by a modular NestJS backend, Breathe delivers health advisories, personalized alerts, and trip planning with air quality forecasts, aiming to improve public health awareness and advance technology driven environmental management.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15690

II. LITERATURE SURVEY

The domain of air quality forecasting has evolved substantially through the application of advanced machine learning and deep learning methodologies. Research primarily focuses on capturing complex linear, non-linear, temporal, and spatial features across various environmental variables. Traditional statistical approaches and baseline regressions, such as ARIMA and Support Vector Regression (SVR) hybridized with Empirical Mode Decomposition [3], have historically been deployed to model urban pollutant dynamics. While effective for short-term predictions under three hours, these structural approaches suffer severe performance degradation over longer horizons because they lack iterative historical pattern recognition and modern attention mechanisms.

To capture extended sequential trends, recent paradigms have heavily embraced Recurrent Neural Networks (RNNs) and their variants. Architectures incorporating Long Short-Term Memory (LSTM) units and Gated Recurrent Units (GRUs)—such as hyper-heuristic multi-chain models [2] or shallow neural structures for localized atmospheric tasks [16]—demonstrate strong capabilities in handling long-range temporal dependencies. Innovations have also extended into the quantum realm, with frameworks like the CNN Quantum-LSTM integrating Variational Mode Decomposition [4] to denoise inputs. However, despite extreme precision, these frameworks suffer from low practical accessibility due to dependencies on specialized quantum hardware. Furthermore, many high-fidelity configurations, such as Transformer-BiGRU setups optimized via Bayesian optimization [5], carry immense computational burdens that preclude edge execution.

Recent research addresses spatial correlation alongside temporal dynamics, as atmospheric pollution naturally propagates across geographical boundings. Researchers have integrated spatial modeling frameworks using Graph Convolutional Networks (GCN) [15, 19] alongside temporal mechanisms like PatchTST transformers [13], which split multi-site information into isolated tokenized patches. Despite their high accuracy, these multi-site architectures often scale poorly, demand dense monitoring station networks, and fail to translate to mobile end-user devices.

To bridge the gap between high-capacity architectures and resource-constrained deployment environments, edge-computing optimizations have emerged as a critical focus. The foundational knowledge distillation framework [17], which trains a compact student model to mirror a massive teacher network, has been successfully adapted from classification domains to regression-based environmental tasks. State-of-the-art implementations, such as multi-task architectures combining Transformers with Bidirectional Mamba2 [8] or hybrid GCN-Transformer-GRU networks [6], utilize knowledge distillation to compress model parameters by multiple factors while retaining a minimal error margins under 5%. Additionally, lower-level optimizations, such as 8-bit quantization on embedded CNN-BiGRU systems [1], have proved vital in reducing model size by over 60% for low-power IoT applications. Nevertheless, a major gap remains in the existing literature: most edge-optimized models focus strictly on multi-pollutant calculations [14] or isolated indoor scenarios [9], completely omitting user-facing integration. Existing systems rarely provide a unified, production-grade cloud and mobile ecosystem that converts raw spatial-temporal forecasts into real-world public health utilities like proactive health advisories or safe trip route planning [10].

III. PROPOSED SYSTEM

The proposed system, **Breathe**, is designed as an end-to-end, cross-platform air quality intelligence framework that mitigates the computational and functional limitations of existing architectures. Rather than relying solely on server-side computations or static ground installations, Breathe couples a state-of-the-art spatial-temporal deep learning architecture with on-device mobile inference and consumer-centric health metrics. The platform provides real-time, multi-city forecasting for critical pollutants (PM 2.5, PM 10, and NO_2) across 29 cities while maintaining a highly responsive, network-resilient, and personalized user experience.

A. Deep Learning Architecture & Knowledge Distillation

At the core of the system lies a hybrid **GCN-TransGRU** architecture optimized for edge devices. The prediction pipeline is executed through three distinct computational layers:

1. **Spatial Encoding (GCN):** Cities are modeled as nodes within a dynamic city graph. Edge weights are defined via a Gaussian kernel function utilizing the Haversine distance between geographical coordinates. A k-Nearest Neighbor (k-NN) fallback mechanism ensures topological continuity by forcing each city to link to a minimum of three neighbors. Graph Convolutional Networks (GCN) aggregate localized neighborhood features across these nodes to capture regional pollution boundaries and inter-city wind/pollutant propagation.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15690

2. **Temporal Modeling (Transformer Encoder):** The spatially aware embeddings generated by the GCN are passed directly into Transformer encoder layers. Utilizing multi-head self-attention mechanisms, this layer learns macro-level, long-range dependencies and periodic fluctuations across the timeline.
3. **Sequence Refinement (Bidirectional GRU):** Finally, a Bidirectional Gated Recurrent Unit (BiGRU) processes the sequence from both forward and backward temporal orientations, capturing micro-level, short-range sequential patterns without the structural parameter overhead of standard LSTMs.

B. Three-Tier Production Architecture

The operational ecosystem is divided across three cohesive layers to guarantee horizontal scalability, high availability, and structural decoupling:

- **The Model Layer (Edge Inference):** Utilizing the flutter_onnxruntime engine, execution happens directly on the client handset. It processes live, 48-hour environmental data tensors of shape (29, 48, 14) paired with bundled JSON min-max scalers to accurately yield 24-hour lookahead forecasts locally.
- **The Backend Cloud Layer (NestJS Infrastructure):** Built as a modular NestJS API framework, this tier manages the orchestration of data and access control. It features secure JWT bearer authentication paired with email-linked One-Time Passwords (OTP) managed via a high-throughput Redis cache layer. The primary relational storage is managed via a PostgreSQL instance. The architecture leverages an abstract Adapter pattern to fetch environmental data dynamically from external providers like the Open-Meteo API. Location-based operations, geocoding, and reverse geocoding are powered by an independent Nominatim service, while user assets and avatars are managed through an S3-compatible Garage object storage cluster.
- **The Client App Layer (Flutter Mobile Application):** Developed using Flutter following a feature-driven Clean Architecture pattern, the frontend uses the GetIt package for rigid dependency injection. The user interface features a live, color-coded AQI dashboard, real-time breakdown statistics, and an interactive layout driven by OpenStreetMap.

C. Functional Capabilities and Public Health Utility

Breathe translates raw predictive analytics into actionable utilities through three major client modules. First, the **Health Advisory Module** evaluates forecasted pollutant bounds against standardized AQI thresholds to output contextual guidelines (e.g., mask advisories, outdoor exercise limits) uniquely customized to user health profiles (such as asthma or chronic respiratory conditions). Second, the **Trip Planning Module** allows commuters to input specific routes and destinations, projecting expected air pollution along the route onto the interactive map layer. This enables users to select optimal, low-exposure travel windows. Finally, the app ensures **Offline Resilience** by caching historical queries and the last-loaded data frames natively in device memory, ensuring continuous utility even during total network blackouts.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The empirical evaluation of the **Breathe** system assesses both the predictive precision of the underlying deep learning architecture and its operational efficiency under deployment-oriented conditions. The forecasting performance is evaluated using standard regression criteria, including the Coefficient of Determination (R^2), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Symmetric Mean Absolute Percentage Error (sMAPE), computed across the normalized pollutant space. Additionally, to evaluate the system's public health utility, continuous PM 2.5 predictions are binned into categorical intervals matching the official Indian Air Quality Index (AQI) bands. This categorization allows for the computation of discrete classification metrics, specifically Precision, Recall, and the F1Score, to measure how accurately the system identifies hazardous environmental conditions.

A. Quantitative Performance and Convergence Analysis

The training regimen of the hybrid GCN-TransGRU framework was structured as a two-stage process. First, the high-capacity teacher model, containing 2,523,203 parameters, was trained and achieved convergence in 42 epochs using an early stopping mechanism. Second, the compact student model, containing 466,115 parameters, was trained within a Knowledge Distillation (KD) framework over 68 epochs. The student network was optimized to replicate the soft output probabilities and structural embeddings of the teacher network.

The experimental logs demonstrate steady convergence for both models, with loss values decreasing sharply during initial epochs before stabilizing. On a held-out chronological test split (covering data from the 2022–2024 window), the distilled student model achieved a final validation loss of 0.00041, closely mirroring the teacher's best validation loss of 0.00039. In terms of predictive capacity, the compressed student network achieved a test R^2 score of 94.5%, marginally outperforming the teacher network's R^2 score of 94.4%. This indicates that the knowledge distillation regularizes the student model, preventing overfitting on noisy spatial-temporal nodes. The student model yielded an overall test MAE of

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026

DOI: 10.17148/IJARCCE.2026.15690

0.0151, an RMSE of 0.0272, and an sMAPE of 22.1% in the normalized pollutant space, confirming a high degree of fidelity across multi-city forecasting tasks.

Metric	Teacher	Student
Test R^2 (%)	94.4	94.5
Test MAE	0.0155	0.0151
Test RMSE	0.0281	0.0272
Parameters	2,523,203	466,115
Best val loss	0.00039	0.00041
Training epochs	42	68 (KD)

Table4.1: Result Analysis

B. Error Distribution and Confusion Matrix Analysis

Analysis of the residual distributions (Prediction vs. Actual) indicates that the errors for both the student and teacher models are tightly clustered around zero in a near-Gaussian distribution, establishing unbiased regression traits across the normalized scale. To evaluate categorical performance, the binned PM 2.5 output was analyzed via an Indian AQI confusion matrix. The student model demonstrates high precision and recall when identifying "Good" and "Satisfactory" air quality categories, which are vital for confirming safe outdoor exercise and travel windows.

Minor misclassifications primarily occur between adjacent, high-pollution bands—such as "Moderate" versus "Poor" or "Very Poor" versus "Severe." This behavior stems from boundary conditions where pollutant concentrations are numerically and visually contiguous, leading the network to occasionally favor a neighboring classification. This minor overlap is an acceptable trade-off for edge-deployed warning modules, as the system consistently identifies the overall trend of air quality degradation.

C. Deployment Efficiency and Comparative Benchmarking

Breathe addresses the challenge of making advanced deep learning execution practical on edge architectures. When benchmarked against state-of-the-art systems, the GCN-TransGRU architecture demonstrates a highly efficient accuracy-to-compute footprint. Traditional sequence networks (such as multi-chain LSTMs [2] or Holt-Winters hybrids [5]) can reach R^2 values of up to 99%, but they require heavy server-side hardware that prevents deployment on mobile applications. Conversely, edge-optimized implementations often limit their scope to single-city tabular spaces [11] or indoor settings [9]. Breathe achieves a balance by maintaining an R^2 above 94% across 29 geographically diverse Indian cities while compressing the execution footprint by 5.4 times compared to the teacher model. Converted to a 1.51 MB ONNX float32 binary, the student model executes batched input tensors of shape (29, 48, 14) in under one second on standard mobile GPUs via flutter_onnxruntime.

D. Operational Limitations and Future Direction

Despite its high predictive accuracy and low inference latency, the proposed system exhibits specific operational dependencies:

1. **Data Stream Integrity:** The structural reliance on a spatial city graph requires an uninterrupted streaming input across all 29 cities. Missing sensor reporting or incomplete 48-hour historical windows from CPCB-style telemetry can weaken the spatial graph signal and degrade localized prediction quality.
2. **Temporal Distribution Shift:** Because the primary training configuration was bound to the 2022–2024 window, changes in urban topography or environmental policies from 2025 onward introduce temporal drift. Evaluating the model against an out-of-time 2025 holdout split verified reasonable generalization, though the overall R^2 score shifted downward to 0.927.
3. **Forecasting Horizon Constraints:** The platform is currently optimized for next-hour, single-step multi-pollutant outputs (PM 2.5, PM 10, and NO 2), meaning multi-day lookahead trajectories are not yet supported.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 6, June 2026
DOI: 10.17148/IJARCCE.2026.15690

V. CONCLUSION

The proposed system Breathe: Air Quality Prediction Using Embedded Machine Learning and Deep Learning Models with Quantization Techniques was successfully designed and implemented to forecast urban air quality using spatial temporal deep learning techniques. The system utilizes a hybrid GCN-TransGRU architecture combined with Knowledge Distillation and INT8 quantization to accurately predict PM2.5, PM10, and NO2 levels across 29 Indian cities (Agartala, Ahmedabad, Aizawl, Bengaluru, Bhopal, Bhubaneswar, Chandigarh, Chennai, Dehradun, Delhi, Gangtok, Gurugram, Guwahati, Hyderabad, Imphal, Itanagar, Jaipur, Kohima, Kolkata, Lucknow, Mumbai, Panaji, Patna, Raipur, Ranchi, Shillong, Shimla, Thiruvananthapuram, and Visakhapatnam).

The results demonstrate that the system achieves a test R^2 of 94.48% using the distilled student model with a compression ratio of 5.4 times, making it a reliable and efficient tool for real time air quality forecasting. By automating multi-city pollutant prediction, the system eliminates the need for manual data analysis and helps users take timely health precautions. The inclusion of additional features such as health advisories, trip planning, and weather integration further enhances the practical usefulness of the system in everyday applications.

Overall, the project highlights the potential of embedded deep learning in transforming traditional air quality monitoring into intelligent, accessible, and data-driven approaches. The system is efficient, deployable, and practical for mobile integration. With further improvements such as expanded datasets, real-time sensor feeds, and multi-horizon forecasting, Breathe can evolve into a powerful environmental intelligence platform for modern urban health management.

REFERENCES

[1]. A. Mazinani, D. Antonucci, D. P. Pau, L. Davoli, and G. Ferrari, "Air Quality Prediction via Embedded ML/DL and Quantized Models," IEEE Access, vol. 13, pp. 123678–123695, 2025, doi: 10.1109/ACCESS.2025.3603920.
[2]. K. Chatterjee, S. S. Kumar, R. P. Kumar, et al., "Future Air Quality Prediction Using Long Short-Term Memory Based on Hyper Heuristic Multi Chain Model," IEEE Access, vol. 12, pp. 123678–123705, 2024, doi: 10.1109/ACCESS.2024.3441109.
[3]. Y. Cao, D. Zhang, S. Ding, W. Zhong, and C. Yan, "A Hybrid Air Quality Prediction Model Based on Empirical Mode Decomposition," Tsinghua Science and Technology, vol. 29, no. 1, pp. 99–111, 2024, doi: 10.26599/TST.2022.9010060.
[4]. F. Naz, M. Fahim, A. A. Cheema, B. D. E. McNiven, T.-V. Cao, R. Hunter, and T. Q. Duong, "Air Quality and Healthy Ageing: Predictive Modeling of Pollutants Using CNN Quantum-LSTM," IEEE Access, vol. 13, pp. 94212–94227, 2025, doi: 10.1109/ACCESS.2025.3570526.
[5]. T. Jayanth, A. Manimaran, V. R. K. Reddy, and R. N., "Enhancing Air Quality Prediction Through Holt–Winters Smoothing and Transformer-BiGRU With Bayesian Optimization," IEEE Access, vol. 13, pp. 180756–180780, 2025, doi: 10.1109/ACCESS.2025.3621231.
[6]. S. Kumar, V. Kour, A. Raj, et al., "Optimizing Air Pollution Forecasting Models Through Knowledge Distillation: A Novel GCN and TRANS-GRU Methodology for Indian Cities," IEEE Access, vol. 13, 2025, doi: 10.1109/ACCESS.2025.3546504.
[7]. M. Ahmed, S. Islam, M. H. Sulaiman, M. M. Hassan, and T. Bhuiyan, "MetaForecaster: A PSO-Driven Neural Model for Sustainable Industrial Air Quality Management," IEEE Access, vol. 13, pp. 121670–121681, 2025, doi: 10.1109/ACCESS.2025.3587716.
[8]. Z. A. Xie, C. O. Chow, J. H. Chuah, and W. J. K. R., "Knowledge-Distilled Multi-Task Model With Enhanced Transformer and Bidirectional Mamba2 for Air Quality Forecasting," IEEE Access, vol. 13, pp. 158870–158880, 2025, doi: 10.1109/ACCESS.2025.3595679.

