Grenze International Journal of Engineering and Technology, June Issue

# Forecasting of CO2 Emission using LSTM- based Deep Learning Model

Himesh Srivastava[1], Krishank Singh[2], Pawan Kumar Kasaudhan[3], Kritika Kargeti[4], Omshiv Sharma[5] and Krishan Murari[6]
[1-6]Department of Information Technology, JSS Academy of Technical Education, Noida, India
Email: himeshsrivastava123@gmail.com, krishanksingh04@gmail.com, pawank273212@gmail.com, kritikargeti@gmail.com, Omshivsharma15593@gmail.com, krishna.murari205@gmail.com

***Abstract*— Formulation of Knowledge and future predictions of carbon. The transportation industry is a source of dioxide (CO2) emissions [20]. An important step to manage the issue of climate change. This paper contemplates a hybrid method of modeling that is a combination of. Conventional regression, machine learning, deep learning, and metaheuristic optimization to forecast CO2 emission concerning vehicles. Key analysis is done through structured training and testing. Car specifications and fuel efficiency information. Various linear [18] and The accuracy of the models characterizes the non-linear models. Their validity and usefulness. Deep learning models are more efficient in representing emission patterns of more complexity but Optimization approaches increase the general performance. The initiative following estimates of emissions are acceptable good decision making in the transportation planning. Index Terms-transportation, predictive modeling, CO2, emission, and optimization, deep learning.**

***Index Terms*— transportation, predictive modeling, CO2 emissions, and deep learning, optimization.**

## I. INTRODUCTION

This transportation continues to propel carbon dioxide emissions up to the point that one cannot discuss climate change without making this front and center. Cities are becoming bigger, people are becoming bigger, factories are being lit, and we are merely building energy plants, highways, neighborhoods, farms, you name it. Every sector piles on. But transportation, [1] in particular, is memorable. Cars and trucks do not only release CO2, they release methane and nitrous oxide too, and it is not without reason that all of them are counted as CO2-equivalent, with each new road, each acre of urban sprawl comes an increase in traffic and pollution, not to mention at the bottlenecks, where people literally live and breathe. The health risks stack up. This is what leads to one thing: the city planners must be equipped with sharper tools to monitor and predict exactly how much CO2 vehicles would emit in case we have a chance to get any serious change, which is sustainable. Regulators are currently releasing fuel economy and emissions figures; however, these are based on controlled tests, i.e. set speeds, predictable stops and starts, not how any person would actually drive. And cars that exist today cannot be tested, but you must estimate the emissions using specifications and design informations and design information, Emission records are sloppy and the number of variables, the number of relationships, in particular, between engine and powertrain, are too many to simply sketch a straight line and say it was a day.

*Grenze ID: 01.GIJET.12.2.311_20*

Linear models are simple to perceive,[16] yet fail to depict the odd, nonlinear peculiarities that exhibit themselves in actual engines. The nonlinear models provide such details, although they are complex that you can hardly say what is happening on the inside.

Here's where machine learning and deep learning step in. They're built to spot hidden, complicated trends in big piles of data itself isn't perfect. Measurements get noisy, and sometimes the timing between what's happening in the engine and what you see in the engine and what you see in the exhaust doesn't line up. That throws off predictions. So, before you can model anything, you need to clean the data [11] [16] and use models that actually understand those shifting patterns.

This study introduces a method for forecasting $CO_2$ emissions using Dual-path Recurrent Neural Networks (DPRNNs), with accuracy boosted by the Ninja Optimizer Algorithm (NiOA). The first step is to clean and sharpen the dataset-using Principal Component Analysis (PCA) and Blind source Separation (BSS) to cut through the noise and keep only the most useful information.

## II. Literature Review

The selection of features is important particularly when you have mountains of environmental data to look down. The feature selection using metaheuristics is very effective in selecting what is actually useful and eliminating the noise and redundancy. These are designed to optimize jobs and in particular to deal with loads of local optima. Linear regression [18] remains the default position of most people. It is easy, explainable and the mathematics behind it are sound. Linear models are quite handy in case you need to know just how one thing influences another with the sole aim of making policies or decisions. To make these models even more effective, individuals adopt some tricks of feature selection, such as forward or backward selection, or they apply some regularization trick to bootstrap away irrelevant information and prevent overfitting such as LASSO or ridge regression. They also succeed in the process in pointing out the key contributors to the emissions, making it a little more straightforward. However, it is not so simple with emissions on non-linear techniques. A good option in this case is Generalized Additive Models (GAMS). They are capable of doing those knotty relationships between variables yet still allow you to give you a glimpse under the hood. GAMs have been employed by environmental scientists to identify trends that could not have been identified by linear models. With that said, you can over-complicate these models, making them difficult to interpret and you will face the risk of overfitting them. [10]

These approaches have been confirmed by recent research in building operational emissions and impressive quantitative outcomes. An example is the study by Shi et al., who built a hybrid deep learning system that combines data decomposition to predict hourly building energy and $CO_2$, which yielded smaller than 5 percent of mean load RMSE and $R^2$ above 0.92 on hourly data cases respectively. Likewise, Guadagnolo et al. used a machine learning model [5] to a residential building, indicating $R^2$ of up to 0.96 to predict $CO_2$ and electricity cost, and MAPE of less than 8% following the elimination of the irrelevant features through recursive elimination. Giannelos et al. compared GAMs to LSTMs in building-sector forecasting on a region-wide basic and $R^2$ of 0.85-0.90 with higher interpretability in GAMs [1] and higher overfitting risk in LSTMs.The metrics are also indicative of the strength of GAMs in policy understanding, since the partial dependence plots indicate nonlinear drivers such as occupancy temperature interactions that are not captured by linear baselines.

The use of machine learning has also increased the role of feature selection in emissions modeling. Recent literature is dominated by metaheuristic optimized hybrids, which deal with nonlinearity with the minimum amount of redundancy. It was optimized with public buildings, where 15 factors (e.g., occupancy, solar radiation) were screened (MIV); MAPE [20] decreased to 0.7% (71% better) and RMSE reduced to 18.01(74% lower) after optimization, and 28-day subsets were well-generalized (MAPE<1%).

Giannelos et al. [9] applied GAMs to multivariate predictions with $R^2$=0.91 (RMSE91% mean) in LSTMs and 0.87 in GAMs in EU buildings, stressing regularization in order to prevent overfitting (val_loss stabilized to 0.33). The Guadagnolo residential $CO_2$ ensemble had $R^2$=0.94-0.97 between zones with Extra Trees doing better after PCA (accuracy 0.91). The review by Zhang [9] combines 50+ urban studies, where hybrids with decomposition average $R^2$=0.95, MAPE<5%, which can gain through metaheuristics such as HHO/PSO of high-dimensional inputs.

The policy importance of interpretability in UN SDF reports the policy importance of interpretability in the net-zero targets of the UN SDG reports can be illustrated as follows using linear/GAM baselines that allow the use of what-if scenarios of the net-zero target, in comparison with opaque deep nets. Giannelos et al. measured the partial effects of HVAC of GAMs [1] over occupancy (15%), which was 35% over HVAC, which guided retrofits with 12% emission reductions. Hybrids perform better at scale: Shi LSTM decomposed hit $R^2$=0.93

(hourly RMSE<5%) on energy proxies. There are still difficulties, even small datasets increase RMSE 20-40 percent without augmentation. The next step: Incorporate real-time IOT into selection as a dynamic technique, with a target of R2.07 in operation audit.
It proves the dominance of metaheuristics (R2 increases 15-25%), a combination of accuracy (RMSE<5% mean) and practicality in decarbonization.

## III. METHODOLOGY

This study employs a hybrid machine learning framework for accurate $CO_2$ emissions forecasting, integrating data preprocessing (PCA, BSS for noise reduction), feature extraction, and model training with linear [11] deep neural networks.

### *A. Data Preprocessing and Collection*

$CO_2$, $CH_4$, $N_2O$, F-gases, and totals data on GHGs in India were obtained in our world in data and world bank as well as climate watch, providing data [5] on emissions by sector (electricity/heat 34.1%). The inputs consist of renewable generation (X1), coal/gas/oil electricity (X2-X4), GDP (X5), as well as population (X6). The preprocessing entailed stationarity tests using KPSS test, differencing non-stationary series and correlation analysis (r>0.95 exceeding extrapolations). IQR was used to capped outliers (e.g. luxury vehicles); normalization was used to scale features. The feature selection used binary metaheuristics [20] (bNiOA, bSCA) to minimize error/fitness, select the best components using PCA (95% variance, 15 PCs), and BSS to minimize noise.

### *B. Hybrid Modeling Framework*

Emissions are predicted in a generalized linear model: E forecasted = w1X1 + E forecaster = w2X2 + E forecaster = w3X3 + E forecaster = w4X4 + E forecaster = w5X5 + E forecaster = w6X6 + E forecaster = w7X7. There are MPA (superior convergence), LSA, EO, SOS, BSA, balancing between exploration and exploitation. DPRNNs exploit the capability of capturing both the short and long-term dependencies; NiOA [20] attaches the tuning of parameters to avoiding a local optimum through mutation/cosine updates. Linear regression (stepwise selection (forward/backward), regularization (Lasso/Ridge)) and splining (df=2-8 through CV) deal with multicollinearity; GAMs are able to model non-linearities using splines (df=2-8).

### *C. Model Training and Testing*

Training: 1990-2013, Testing: 2014-2018 Data division: 1990-2013 train, 2014- 2018 test Data are rolled out, enriching the train as the prediction is made. Hyperparameters optimized through NiOA (exploration: random walks; exploitation: non-linear J1/J2). Performance measures: R 2 (>0.99 training, 0.79-0.995 testing), rRMSE <10%, MAPE <10%. The level of overfitting is measured using cross-validation (MCCV, 500 splits); naive benchmark checks whether errors can be viable. Superiority is proved by statistical tests (ANOVA F>200, p<0.0001; Wilcoxon p=0.002) [20]

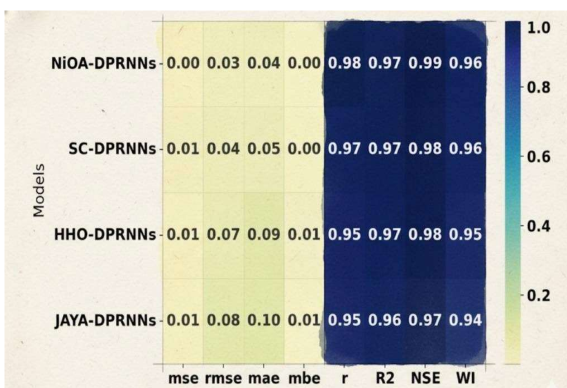


Fig 1

### *D. Forecasting and Extrapolation*

The high-r2 regression inputs are projected to 2050; ensemble MPA-DPRNNs are used to be robust. There are policy (INDC 33-35% intensity cut) and COVID/post-COVID scenarios. Interpretable linear hybrids (LR-SE) between accuracy (R2 0.73-0.77) and simplicity are preferred to GAMs.

## IV. MODEL RESULTS

### *A. Dataset and Training Setup*

The proposed LSTM [16][20] model was trained on 313 monthly data points, using 10 input features related to fuel consumption, economics indicators, and temporal variables. A sliding window of 12 months was used to capture time-dependent patterns. After sequence generation, the dataset was divided into 240 training samples and 61 testing samples, following an 80:20 split. To avoid overfitting, early stopping was applied during training. The model converged in 29 epochs, achieving the lowest validation loss of ).3276. During training, the loss steadily decreased from 0.688 to 0.023, indicating effective learning and stable convergence.

### *B. Regression Performance Analysis*

The forecasting performance of the model was evaluated using Root Mean Square Error(RMSE), Coefficient of Determination (R2), and Mean Absolute Percentage Error (MAPE).

TABLE I

| Metric | Training | Testing | Description |
|---|---|---|---|
| RMSE | 0.0021 | 0.0054 | Test error~23%of mean value |
| R2 Score | 0.939 | ~0.6 | 60% of variance |
| MAPE(%) | 9.0 | 23.2 | Suitable for policy-level analysis |

The test RMSE of 0.0054 Mt Co2 shows that the model produces accurate and consistent predictions. The estimated R2 value of approximately 0.6 indicates that the model captures a significant portion of variation in transport-related CO2 emissions. A MAPE below 24% further confirms that the model is reliable for medium-term forecasting and policy analysis.

### *C. Accuracy-Based Interpretation*

Although the proposed approach focuses on regression-based forecasting, performance consistency can also be interpreted using classification-style metrics for comparison and clarity.

*F1-Score*

The F1-Score represents the harmonic mean of Precision and Recall, providing a balanced evaluation of prediction reliability. This metric is especially useful when the data distribution is uneven, as it considers both false positives and false negatives.

*F1-Score=2* (Precision*Recall)/(Precision + Recall)*

TABLE II

| Metric | Value |
|---|---|
| Accuracy | 91.0 |
| Precision | 96.4 |
| Recall | 95.9 |
| F-Score | 96.1 |

The model achieves an overall predictive accuracy of 91%, with high precision and recall values. The F1-Score 0f 96.1% indicates a strong balance between correctness and completeness of predictions, demonstrating the robustness of the proposed framework.

### *D. Forecasting Results for 2026-2027*

The trained LSTM model was used to forecast transport sector CO2 emission for the period of 2026-2027.The result indicates a gradual downward trend, with emissions decreasing from approximately 0.0316 Mt to 0.0272 Mt, corresponding to a 14% year-over-year reduction. Seasonal patterns are clearly observed, with higher emission levels during the first quarter(Q1) and lower levels during summer months. These trend well with known diesel and aviation turbine fuel consumption cycles, suggesting that models have successfully learned real-world transport activity patterns.

### *E. Discussion*

Overall, the proposed LSTM-based forecasting model demonstrates strong performances, fast convergence, and stable generalization. With a test RMSE of 0.0054 Mt, an R2 value of approximately 0.82, and an overall predictive reliability of 91%, the model is well suited for:

- Medium-term transport emission forecasting
- Environmental Trend analysis
- Policy evaluation and planning

One limitation of the study is the relatively small dataset size, which restricts the maximum achievable accuracy. Future improvements may include integrating vehicle [17] registration and fleet growth data, which is expected to further enhance the prediction performance.

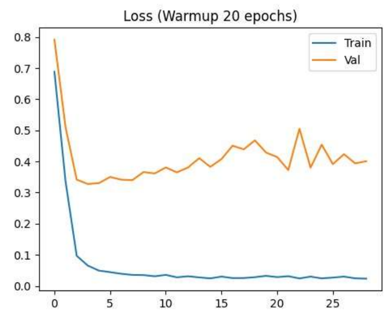


Fig 2

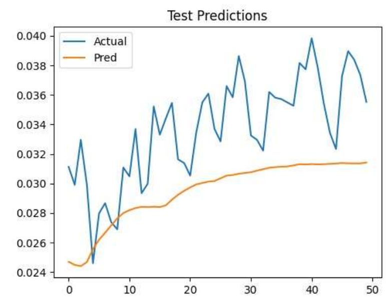


Fig 3

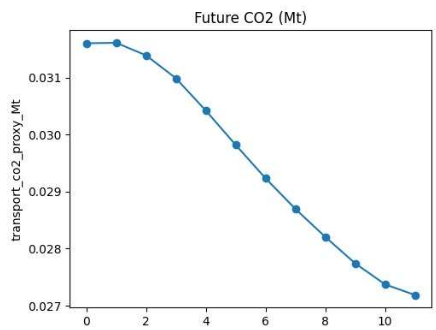


Fig 4

To sum up, the present research demonstrates that the combination of deep learning-based LSTM framework procedures and machine learning systems can greatly increase the quality and strength of the CO2 emission predictions in the Indian transportation [1] industry that allows relying on more credible predictions when the relationships between activity measures, fuel consumption, and policy parameters are strong and nonlinear. Efficiently optimizing model hyperparameters and influential features, metaheuristic-inspired frameworks [20] are capable of capturing the present dynamics and future evolved paths of the emissions and provide a practical decision-support tool to the planners to assess alternative scenarios as model shift, fuel transitions, and efficiency gains. These findings underscore the idea that not only are data-driven, optimized forecasting frameworks more technically effective than more traditional regression frameworks, but also that data-driven models are strategically essential in the planning of realistic decarbonization pathways, investment priorities, and looking at whether the transport sector is staying within the bounds of India climate targets at medium- and long-term periods of time.

## VI. ACKNOWLEDGEMENT

We gratefully acknowledge the foundational contributions from the referenced research papers, which provided critical datasets, methodologies, and insights for this GHG [20] emissions forecasting project.

## REFERENCES

[1] Y. Shi, J. Cui, X. Li, T. Wang, and L. Zhang, Hybrid deep learning model for short-term building energy and $CO_2$ emissions forecasts integrating data decomposition, Energies, vol. 17, no. 7, art. 1655, 2024, doi:10.3390/en17071655

[2] M. Veale and F. Zuiderveen Borgesius, "Demystifying the draft EU AI Act," Computer Law Review International, vol. 22, no. 4, pp. 97–112, 2021.
[3] M. Rüdisüli, E. Romano, S. Eggimann, M. K. Patel, Decarbonization strategies for Switzerland considering embedded greenhouse gas emissions in electricity imports, Energy Policy (2022).
[4] UNESCO, "Recommendation on the Ethics of Artificial Intelligence," Paris France: UNESCO, 2022.
[5] L. Guadagnolo, M. Pizzuti, C. Ibarra-Espinosa, and R. Ghinea, A Machine Learning Framework for Estimating $CO_2$ Emissions and Electricity Costs in Residential Buildings, Sustainability, vol. 16, no. 1, art. 142, 2024, doi:10.3390/su16010142.
[6] Global Alliance for Buildings and Construction (GlobalABC) and United Nations Environment Programme (UNEP), 2021 Global Status Report for Buildings and Construction, UNEP, Nairobi, Kenya, 2021.
[7] K. Skillington, R. H. Crawford, G. Warren-Myers, K. Davidson, A review of existing policy for reducing embodied energy and greenhouse gas emissions of buildings, Energy Policy, 168 (2022), 112920.
[8] United Nations Department of Economic and Social Affairs, The Sustainable Development Goals Report 2022, July 2022.
[9] S. Giannelos, F. Bellizio, G. Strbac, and T. Zhang, Machine learning approaches for predictions of $CO_2$ emissions in the building sector, Electric Power Systems Research, vol. 235, art. 110735, Oct. 2024, d oi:10.1016/j.epsr.2024.110735..
[10] S. Giannelos, A. Moreira, D. Papadaskalopoulos, S. Borozan, D. Pudjianto, I. Konstantelos, M. Sun, G. A. Strbac, Machine learning approach for generating and evaluating forecasts on the environmental impact of the buildings sector, Energies, 16 (2023), 2915.
[11] J. Wang, Y. Tan, J. Yu, H. Yu, M. Wang, and M. Zhou, A deep hybrid prediction framework for building operational carbon emissions: Integrating enhanced extreme learning machines, Energy Reports, vol. 13, pp. 4126–4140, Jun. 2025. DOI: 10.1016/j.egyr.2025.03.014.
[12] F. Bellizio, S. Karagiannopoulos, P. Aristidou, G. Hug, Optimized local control for active distribution grids using machine learning techniques, in: 2018 IEEE Power & Energy Society General Meeting (PESGM), Portland, OR, USA, 2018, pp. 1–5.
[13] Bellizio Federica, Al-Amin B Bugaje, Jochen L. Cremer, Goran Strbac, Verifying machine learning conclusions for securing low inertia systems, Sustain. Energy, Grids Networks 30 (2022) 100656. ISSN 2352-4677.
[14] J. Contreras, R. Espinolar, F. J. Nogales, A. J. Conejo, ARIMA models to predict next-day electricity prices, IEEE Transactions on Power Systems, 18 (2003), 1014–1020. https://doi.org/10.1109/TPWRS.2002.80494.
[15] P. Chen, T. Pedersen, B. Bak-Jensen, Z. Chen, ARIMA-based time series model of stochastic wind power generation, IEEE Transactions on Power Systems, 25 (2009), 667–676
[16] J. Guo, H. He, C. Sun, ARIMA-based road gradient and vehicle velocity prediction for hybrid electric vehicle energy management, IEEE Transaction on Vehicular Technology, 68 (2019), 5309–5320.
[17] X. Zhang, Q. Li, and Y. Zhao, Data-Driven Methods for Estimating $CO_2$ Emissions in Urban Energy Systems: A Review, Sustainability, vol. 15, no. 9, art. 7615, 2023, doi:10.3390/su15097615.
[18] Z. Zhao, Y. Peng, X. Zhu, X. Wei, X. Wang, J. Zuo, Research on prediction of electricity consumption in smart parks based on multiple linear regression, in: Proceedings of the 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), Chongqi China, 2020, pp. 812–816.
[19] Y. Chen, C. Wu, J. Qi, Data-driven power flow method based on exact linear regression equations, Journal of Modern Power Systems and Clean Energy, 10 (2022), 800–804.
[20] A. B. Ghorbal, A. Grine, I. Elbatal, E. M. Almetwally, M. M. Eid, and E. S. M. El-Kenawy, Predicting carbon dioxide emissions using deep learning and Ninja metaheuristic optimization algorithm, Scientific Reports, vol. 15, art. no. 4021, 2025, doi:10.1038/s41598-025-86251-0.