Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MoE-CB based dynamic resource allocation in 5G

Soham Biswas
Centre for Development of Telematics
Partho Barui
parthobarui445@gmail.com

Centre for Development of Telematics
Gokulraj Pandiyan
Centre for Development of Telematics

Research Article

Keywords: MoE, AI, Resource optimization, 5G

Posted Date: September 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10415152/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

MoE-CB based dynamic resource allocation in 5G

Soham Biswas[1†], Partho Barui[1*†], Gokulraj Pandiyan[1]

[1]Centre for Development of Telematics, Bengaluru, 560100, Karnataka, India.

*Corresponding author(s). E-mail(s): parthobarui445@gmail.com;
Contributing authors: ; sohambiswas61@gmail.com; gokul@cdot.in;
†These authors contributed equally to this work.

Abstract

To ensure the Quality of Service (QoS) of heterogeneous and latency sensitive applications, the dynamic resource allocation is introduced in the 5th generation (5G) networks. This study proposes a Mixture of Experts (MoE) framework with CatBoost (MoE-CB). This is designed to support adaptive and intelligent resource orchestration, making it a suitable candidate for the RAN Intelligent Controller (RIC). The model uses a gating mechanism in conjunction with hybrid CNN-LSTM and CNN-BiGRU expert networks to capture temporal and spatial dependencies in network parameters like latency, bandwidth requirements, and signal strength. A CatBoost regressor is used to model residual nonlinear patterns and further refine the experts' predictions, with a mean squared error (MSE) of 0.000405, root mean squared error (RMSE) of 0.020125, and coefficient of determination (R^2) of 0.9502 on a public 5G resource allocation dataset.

Keywords: MoE, AI, Resource optimization, 5G

1 Introduction

The quick development of 5G wireless networks has brought with it previously unheard of demands for high data rates, low latency, and extensive device connectivity. The dynamic changes in network conditions brought about by the user mobility, traffic bursts, and heterogeneous service requirements make efficient radio resource allocation a crucial challenge [1]. Traditional rule based or static allocation methods cannot adapt well for complex and dynamic environments. This often spurs the use of learning based and data driven optimization techniques. More intelligent and proactive resource management strategies are made possible by the strong potential of ML models to predict traffic patterns and network states [2].

To deploy an intelligent control application with analytical and ML features for adaptive scheduling and policy enforcement, the O-RAN Alliance has introduced the RIC. By advanced management of the network resources, these control systems can predict network demand and help maintain good QoS [3].

In this contest, developing prediction frameworks becomes essential for enabling network behavior. For these reasons, this work explores a hybrid MoE based learning framework to model temporal variations for dynamic allocation [4, 5]. By improving prediction reliability and adaptability, the proposed approach aims to assist programmable entities such as the RIC in maintaining QoS across diverse service scenarios [6].

2 Literature Review

The evaluation of resource allocation techniques in 5G networks a big transition from classical optimization to an intelligent learning framework. A major milestone in the intelligent allocation was presented by Z. Liu et al.(2019). In their paper authors formulated a dynamic resource allocation in ultra dense5G networks as a Markov Decision Process (MDP). To address the problem, the authors proposed a Dueling Deep Q-Network (Dueling DQN), which optimized energy efficiency (EE) and spectral efficiency (SE). This work demonstrated performance improvements over conventional Q-learning and standard DQN. This shows that deep reinforcement learning (DRL) can be a viable solution for dynamic 5G environments [7].

Khan et al. (2019) proposed a hybrid fuzzy logic with genetic algorithm (H-FLGA) for dynamic resource allocation for 5G driven VANETs. They formulated the problem as an optimization task considering QoS metrics such as delay and service requirements. A fuzzy inference system was employed to adaptively tune objective weights, while a genetic algorithm performed the optimization. Although the authors' proposed approach achieved low delay and improved cost performance compared to the conventional GA methods but, it relies on heuristic and iterative search that may create overhead and limit real time scalability [8].

Rico and Merino (2020) presented a comparative study on end to end solutions for reliable low latency communication in 5G networks. The study classifies existing techniques to achieve low latency (1 ms) and high reliability (10-9 error rate). This is essential for applications such as autonomous driving, industrial automation, robotics, and tactile internet services. The survey organizes the literature into three main categories: enhanced end-to-end communication protocols, network-level enabling technologies, and application programming interfaces (APIs) that optimize protocol utilization [9].

Xu, Gui, Gacanin, and Adachi(2021) presented a comprehensive survey on resource allocation in 5G heterogeneous networks (HetNets). The authors focused on the challenges introduced by multi tier cellular architectures and spectrum sharing constraints. The study classifies existing resource allocation approaches in various HetNet scenarios like OFDMA-based, NOMA-based, relay-assisted and multi-antenna systems. Also evaluates them in terms of allocation criteria, optimization techniques and architectural design. The survey highlights the importance of efficient allocation strategies to mitigate inter-cell interference and improve spectrum reuse in dense deployments. Furthermore, the authors identify new research directions and suggest that future wireless networks may use to manage resources more intelligently [10].

Selvamanju and Shalini(2022) proposed a mobile traffic prediction based Archimedes Optimization Algorithm with Deep Belief Network (AOA-DBN) for 5G cellular networks. Their approach integrates a Deep Belief Network (DBN) for time-series traffic forecasting. Where the Archimedes Optimization Algorithm (AOA) is employed to tune the DBN hyperparameters optimally. The framework follows a three-stage process including pre-processing, prediction, and hyperparameter optimization. Simulation results demonstrated improved prediction accuracy with a minimum average MAPE of 0.034 and reduced computational time. However, the model primarily focuses on traffic forecasting rather than direct real-time resource allocation decisions [11].

Goswami et al.(2022) proposed a neural-network-based optimal resource allocation method for secure Industrial IoT (IIoT) networks. Their work jointly addresses resource allocation and data security. A convolutional neural network (CNN) is employed to extract optimal channel state information for different applications which enable efficiant and secure utilization of resources. The proposed approach improves flexibility compared to traditional fixed resource allocation schemes and the authors demonstrates the performance of the model through analytical validation and simulations [12].

Gao et al.(2023) proposed a GCN-LSTM-based resource allocation method for 5G networks operating in densely distributed scenarios. Their approach uses Graph Convolutional Networks (GCN) to learn how different network areas are connected to each other. It also uses Long Short-Term Memory (LSTM) networks to extract temporal traffic patterns from historical 5G data. By combining spatial and temporal features, the

model achieves high accuracy in traffic prediction. However, the framework primarily focuses on traffic prediction driven allocation rather than adaptive decision modeling [13].

Yarkina et al.(2024) studied machine learning enhancements for 5G RAN network slicing within an ITU-T compliant framework. They proposed a priority-based resource arbitration mechanism and evaluated several supervised ML techniques to accelerate slice resource allocation while ensuring performance isolation and fairness. Simulation results showed that ML models significantly reduced execution time compared to exact optimization methods, with polynomial regression achieving the best trade-off between accuracy and complexity [14].

Pathak et al.(2024) explored resource allocation optimization in 5G networks using a Variational Quantum Regressor (VQR) within a Quantum Machine Learning (QML) framework. Their study demonstrates that the VQR model can achieve highly accurate predictive performance, significantly reducing mean squared error, where MSE is approximately 0.008. The results highlight the potential of quantum-assisted learning models to enhance prediction accuracy in 5G networks [15].

Sachin Singh(2025) evaluated the performance of multiple machine learning regression models for 5G network resource allocation optimization. Using a 5G resource allocation dataset with eight attributes, the study compared K-Nearest Neighbors (KNN), Random Forest Regressor (RFR), Gradient Boosting Regressor (GBR), and Extra Trees Regressor (ETR) based on metrics such as R-squared, MSE, and RMSE. Experimental results showed that KNN achieved the highest prediction accuracy ($R^2 = 0.9378$) with low error values. However, the study primarily focuses on performance comparison of standalone models without addressing adaptive, multi-expert optimization mechanisms [16].

Karuppiyan et al.(2025) proposed the DRARLCNN model, a hybrid Reinforcement Learning–Convolutional Neural Network framework for dynamic resource allocation in 5G networks. The model integrates CNN-based feature extraction with RL-driven decision-making to optimize latency and Quality of Service (QoS) in dynamic environments. Experimental results demonstrated improved allocation efficiency with performance metrics of $R^2 = 0.517$, MSE = 0.035, and RMSE = 0.188. The primary contribution of authors is combining deep feature learning with adaptive decision mechanisms, offering improved responsiveness compared to traditional models. [17].

Another authors, Gaurav et al.(2025) proposed a resource allocation using a K-Nearest Neighbors (KNN) model with the Osprey Optimization Algorithm (OOA). The model predicts bandwidth and computational resource distribution and the performance was compared to traditional ML approaches such as Linear Regression, Random Forest, and XGBoost. Experimental results shows the predictive accuracy with an MSE of 0.008, MAE of 0.02, and R^2 above 0.9. However [18].

The details of the above-mentioned studies, including their methodologies, objectives, strengths, and limitations, are summarized in Table 1.

3 Research Gap

Although significant advancements have been made in 5G resource allocation through reinforcement learning, metaheuristic optimization, and deep learning-based prediction models, several limitations remain. Many existing deep learning approaches rely on single-architecture models or tightly coupled hybrid structures that learn a unified representation for all traffic patterns. Such monolithic learning may struggle to generalize effectively across heterogeneous temporal behaviors such as stable, bursty, and highly fluctuating network conditions. Furthermore, some works combine spatial and temporal modelling in 5G networks. However, for dynamic resource allocation conditional expert and adaptive gating mechanisms are sill not exployed significantly.

In addition, limited studies investigate the integration of deep sequential feature learning with tree-based gradient boosting refinement to enhance residual nonlinear modeling and robustness against outliers. Therefore, there exists a need for a structured multi-expert framework capable of adaptive specialization, improved generalization across diverse temporal regimes, and enhanced nonlinear refinement for reliable 5G resource allocation prediction. In the Figure 1 the research gap has been demonstrated clearly.

Table 1 Literature Survey on Existing 5G Resource Allocation Approaches

Author	Year	Learning Type	Core Model	Main Objective	Key Strength	Identified Limitation
Liu et al. [7]	2019	DRL	Dueling DQN (MDP formulation)	Joint optimization of energy and spectral efficiency	Demonstrates superior dynamic allocation vs Q-learning and DQN	Large state space and high training complexity
Khan et al. [8]	2019	Metaheuristic + Fuzzy	H-FLGA	Multi-objective QoS-aware allocation in 5G VANETs	Adaptive objective weighting; reduced delay and cost	Iterative evolutionary search limits real-time scalability
Rico & Merino [9]	2020	Survey	End-to-end latency frameworks	Reliability and ultra-low latency analysis	Structured taxonomy of latency solutions	No predictive or intelligent allocation mechanism
Xu et al. [10]	2021	Survey	HetNet RA classification models	Interference mitigation and spectrum reuse	Comprehensive classification of RA strategies	Conceptual frameworks without implementation
Selvamanju & Shalini [11]	2022	Deep Learning	AOA-DBN	Mobile traffic prediction for QoS-aware allocation	High forecasting accuracy (MAPE 0.034) with optimized parameters	Focuses on prediction rather than direct allocation
Goswami et al. [12]	2022	Deep Learning	CNN-based secure allocation	Joint resource allocation and IIoT security optimization	Channel-aware flexible allocation improves efficiency	Limited temporal modeling capability
Gao et al. [13]	2023	Spatiotemporal DL	GCN-LSTM	Traffic-driven allocation in dense 5G scenarios	Captures spatial and temporal correlations	Relies mainly on prediction rather than adaptive decision learning
Yarkina et al. [14]	2024	Supervised ML	Polynomial regression + ML arbitration	Fast network slicing resource arbitration	Reduced execution time with fairness guarantees	Weak sequential and deep modeling capability
Pathak et al. [15]	2024	Quantum ML	Variational Quantum Regressor	Precision allocation prediction	Very low prediction error (MSE ≈ 0.008)	Hardware limitations and experimental maturity
Sachin Singh [16]	2025	ML Regression	KNN, RFR, GBR, ETR	Dataset-based allocation prediction	High standalone prediction accuracy ($R^2 = 0.9378$)	No adaptive or multi-model decision framework
Karuppiyan et al. [17]	2025	Hybrid RL–DL	DRARLCNN	Latency and QoS optimization	Combines deep features with RL-based decisions	Moderate prediction accuracy and single architecture
Gaurav et al. [18]	2025	Metaheuristic + ML	OOA-optimized KNN	Bandwidth and computation allocation prediction	High accuracy ($R^2 > 0.9$) with optimization tuning	Depends on single-model prediction strategy



Fig. 1 Research Gap

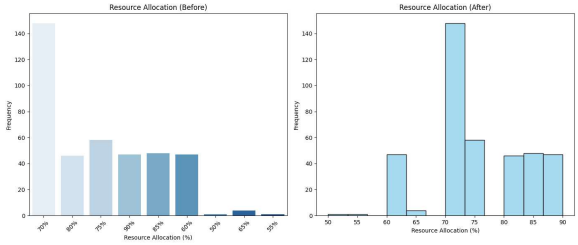


Fig. 2 Histogram before and after preprocessing

To address the limitations of existing resource allocation approaches, this work presents a gated multi-expert model that captures heterogeneous traffic regimes while improving generalization, robustness, and prediction reliability in 5G systems.

4 Dataset Description and Preprocessing

This study uses a renowned public dataset [19] comprising 8 features, reflecting various scenarios of resource allocation in the 5G network. The dataset includes key parameters such as signal strength, required and allocated bandwidth, application types, resource allocation percentages etc. Resource allocation percentage, the target value shows how resources are distributed among different users and applications.

To get the dataset ready for ML based resource allocation prediction, thorough preparation was necessary. The following preprocessing techniques are obtained in this study.

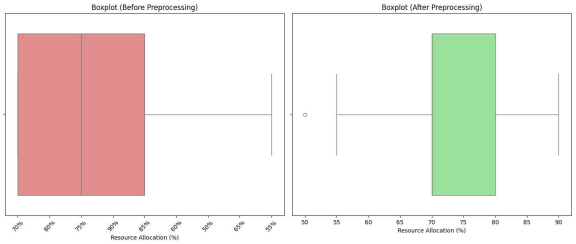


Fig. 3 Boxplot before and after preprocessing

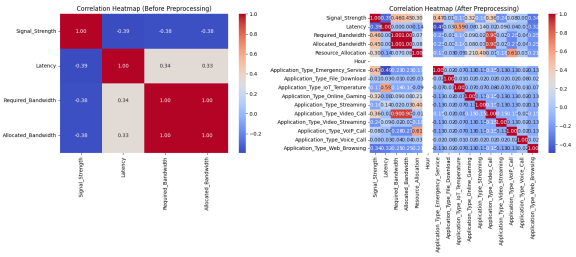


Fig. 4 Correlation heatmap before and after preprocessing

4.1 Conversion of Target Variable

Initially, the column "Resource_Allocation", contained percentage signs (such as "70%"). Regression analysis was made possible by removing them and converting the values to floating point numbers.

4.2 Unit Cleanup and Numeric Conversion

A few of the dataset's characteristics included values with units attached. For example, "Signal_Strength" entries had "dBm," "Latency" values were tagged with "ms," and bandwidth columns like "Required_Bandwidth" and "Allocated_Bandwidth" had mixed units like "Mbps" and "Kbps." All bandwidth measurements were normalised to Mbps after these units were methodically eliminated, producing a collection of clear, consistent numerical characteristics that were ideal for application in ML techniques.

4.3 Temporal Feature Engineering

Time of day effects on network performance were captured by converting the "Timestamp" column to date time format and extracting the hour component as an extra feature.

4.4 Categorical Encoding

The model was able to comprehend the different network service kinds without imposing an arbitrary order by converting the categorical "Application_Type" into one hot encoded binary columns.

4.5 Feature Selection and Cleanup

Irrelevant identifiers such as "Timestamp" and "User_ID" were dropped to prevent leakage and noise in the learning process. Transformation for Sequence Modeling

4.6 Transformation for Sequence Modeling

Sliding windows of time steps (TIME_STEPS) were constructed from the feature set in order to capture temporal relationships in the data. The goal value corresponds to the final step in each window's sequence of input characteristics. Recurrent and attention-based neural networks can learn from past context and dynamically forecast resource allocation.

4.7 Dataset Visual Analysis

Figure. 2 displays the Resource Allocation variable's distribution as category string values (e.g., "70%", "80%") before cleaning. Without exposing underlying patterns or numerical correlations, this format groups values. Figure. 3 provides a visualization of the distribution and central tendency for resource allocation. It is clearly visible that after conversion to a numeric format, it reveals a clear spread, median, quartiles and potential outliers within resource allocation. This transformation enables rigorous statistical analysis and supports ML applications by providing an accurate, interpretable summary of data variability and central tendency. Figure. 4 presents correlation heatmaps of the dataset before preprocessing and after preprocessing. The heatmap produced before preprocessing is limited to original numeric features and shows only basic relationships in the raw data. After preprocessing including cleaning, feature engineering and encoding the heatmap incorporates all numeric and transformed categorical variables, revealing more detailed network

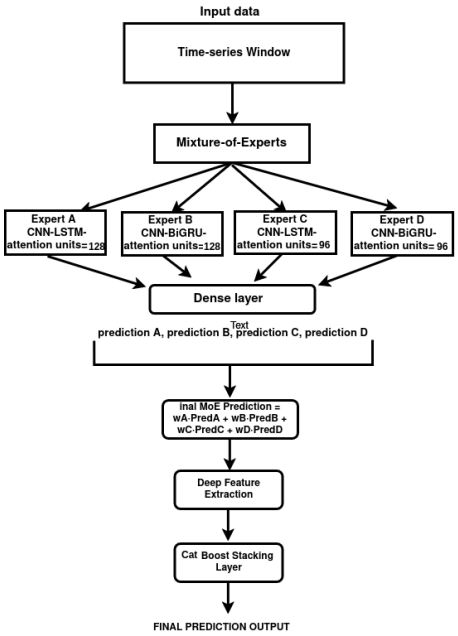


Fig. 5 Proposed model architecture

of associations. This demonstrates how preprocessing has been performed to enhance feature interoperability and analytical capability.

5 Proposed Models

In this section, the proposed model has been discussed. This hybrid model is explicitly designed to address the challenges posed by dynamic 5G resource allocation. The model extracts temporal and spatial information from important network parameters, including Signal Strength, Latency, and Bandwidth Requirements, by utilizing various neural network experts (CNN-LSTM and CNN-BiGRU). To improve prediction accuracy, a gating network is used to integrate the forecasts of these experts, and the resulting deep features are then stacked using an CatBoost regressor. Figure 5 presents the model overview.

5.1 Input Representation

Assume that $X = \{x_1, x_2, \ldots, x_T\}$ represents an input sequence inside a T-length sliding window. By converting unstructured sequential data into organized temporal segments, the time series windowing process allows the model to capture both short-term and long-term fluctuations.

Algorithm 1 MoE-CB Based Dynamic Resource Allocation

Require: Dataset D, window size T
Ensure: Final predicted allocation $\hat{Y}$

Load dataset D
Clean, normalize, and encode features
Convert target values to numeric form
Temporal Sequence Construction
Initialize sequence set $S \leftarrow \emptyset$
for $i = 1$ to $N - T + 1$ **do**
    $X_i \leftarrow D[i : i + T - 1]$
    $y_i \leftarrow$ target at last timestep
    $S \leftarrow S \cup (X_i, y_i)$
end for
Extract last-step feature vector
$X_{\text{last}} \leftarrow x_T$ for each sequence
Split S into training and testing sets
Standardize feature values
Mixture-of-Experts Learning
for each expert $E_i \in \{A, B, C, D\}$ **do**
    $F_i \leftarrow$ CNN feature extraction
    $T_i \leftarrow$ temporal modeling (LSTM/BiGRU)
    $C_i \leftarrow$ attention aggregation
    $\hat{y}_i \leftarrow$ dense prediction
    Store deep representation f_i
end for
Gating Network
$P \leftarrow$ global pooling of sequence input
$G \leftarrow \text{Concatenate}(P, X_{\text{last}})$
$w \leftarrow \text{Softmax}(\text{MLP}(G))$
Mixture Prediction
$Y_{MoE} \leftarrow 0$
for $i = 1$ to 4 **do**
    $Y_{MoE} \leftarrow Y_{MoE} + w_i \cdot \hat{y}_i$
end for
Feature Stacking for CatBoost
$F_{stack} \leftarrow \text{Concatenate}(X_{\text{last}}, Y_{MoE})$
Final Prediction
Train CatBoost using stacked training features F_{stack}
$\hat{Y} \leftarrow \text{CatBoost}(F_{stack})$
return $\hat{Y}$

5.2 Mixture-of-Experts Layer

The learning task is divided among four independent experts according to the MoE framework [20]. Every expert generates a unique forecast based on the same information X.

$$X = \{x_1, x_2, \ldots, x_T\} \tag{1}$$

$$y_i = f_i(X), \ i \in \{A, B, C, D\} \tag{2}$$

where f_i stands for a single expert network. By permitting conditional computing, the MoE aims to prevent overfitting, enhance specialization, and lessen model bias. MoE only engages the most pertinent experts for a particular pattern, in contrast to monolithic deep models.

5.3 Expert Networks

Hybrid CNN-LSTM & CNN-BiGRU with Attention networks are used in the construction of each expert. While LSTM and BiGRU layers model sequential dependencies and temporal dynamics, CNN layers extract local spatial dependencies and feature representations.

5.4 Gating Network and Weighted Prediction

A gating MLP computes adaptive importance weights W_A, W_B, W_C, W_D such that,

$$w_A + w_B + w_C + w_D = 1, w_i \geq 0 \tag{3}$$

And the final MoE aggregated output is expressed as:

$$Y_{\text{MoE}} = w_A y_A + w_B y_B + w_C y_C + w_D y_D \tag{4}$$

5.5 CatBoost Stacking Layer

To improve prediction accuracy, the aggregated MoE output is combined with the last step input features to form the stacked feature representation:

$$F_{\text{stack}} = \text{Concatenate}(X_{\text{last}}, Y_{\text{MoE}}) \text{ where, } X_{\text{last}} = x_T \tag{5}$$

The final resource allocation prediction is obtained using a CatBoost regressor:

$$Y_{\text{final}} = \text{CatBoost}(F_{\text{stack}}) \tag{6}$$

Algorithm 1 describes the proposed model workflow in details and the details of the model layers are shown in Table 2.

Table 2 MoE Model Layer Details

Layer (type)	Output Shape	Param #	Connected to
seq_input (InputLayer)	(None, 5, 15)	0	-
conv1d_104 (Conv1D)	(None, 5, 64)	2,944	seq_input[0][0]
conv1d_106 (Conv1D)	(None, 5, 64)	2,944	seq_input[0][0]
conv1d_108 (Conv1D)	(None, 5, 32)	1,472	seq_input[0][0]
conv1d_110 (Conv1D)	(None, 5, 32)	1,472	seq_input[0][0]
batch_normalization (BatchNormalization)	(None, 5, 64)	256	conv1d_104[0][0]
batch_normalization_1 (BatchNormalization)	(None, 5, 64)	256	conv1d_106[0][0]
batch_normalization_2 (BatchNormalization)	(None, 5, 32)	128	conv1d_108[0][0]
batch_normalization_3 (BatchNormalization)	(None, 5, 32)	128	conv1d_110[0][0]
conv1d_105 (Conv1D)	(None, 5, 64)	12,352	batch_normalization[0][0]
conv1d_107 (Conv1D)	(None, 5, 64)	12,352	batch_normalization_1[0][0]
conv1d_109 (Conv1D)	(None, 5, 32)	3,104	batch_normalization_2[0][0]
conv1d_111 (Conv1D)	(None, 5, 32)	3,104	batch_normalization_3[0][0]
lstm_57 (LSTM)	(None, 5, 128)	98,816	conv1d_105[0][0]
lstm_58 (LSTM)	(None, 5, 128)	98,816	conv1d_107[0][0]
lstm_59 (LSTM)	(None, 5, 96)	49,536	conv1d_109[0][0]
lstm_60 (LSTM)	(None, 5, 96)	49,536	conv1d_111[0][0]
dropout_56 (Dropout)	(None, 5, 128)	0	lstm_57[0][0]
dropout_57 (Dropout)	(None, 5, 128)	0	lstm_58[0][0]
dropout_58 (Dropout)	(None, 5, 96)	0	lstm_59[0][0]
dropout_59 (Dropout)	(None, 5, 96)	0	lstm_60[0][0]
global_average_pooling (GlobalAveragePooling1D)	(None, 15)	0	seq_input[0][0]
last_input (InputLayer)	(None, 15)	0	-
attention_layer_56 (AttentionLayer)	(None, 128)	8,256	dropout_56[0][0]
attention_layer_57 (AttentionLayer)	(None, 128)	8,256	dropout_57[0][0]
attention_layer_58 (AttentionLayer)	(None, 96)	6,208	dropout_58[0][0]
attention_layer_59 (AttentionLayer)	(None, 96)	6,208	dropout_59[0][0]
concatenate_30 (Concatenate)	(None, 30)	0	global_average_pooling[0][0], last_input[0][0]
dense_90 (Dense)	(None, 64)	8,256	attention_layer_56[0][0]
dense_91 (Dense)	(None, 64)	8,256	attention_layer_57[0][0]
dense_92 (Dense)	(None, 64)	6,208	attention_layer_58[0][0]
dense_93 (Dense)	(None, 64)	6,208	attention_layer_59[0][0]
dense_94 (Dense)	(None, 128)	3,968	concatenate_30[0][0]
featA (Dense)	(None, 32)	2,080	dense_90[0][0]
featB (Dense)	(None, 32)	2,080	dense_91[0][0]
featC (Dense)	(None, 32)	2,080	dense_92[0][0]
featD (Dense)	(None, 32)	2,080	dense_93[0][0]
dropout_60 (Dropout)	(None, 128)	0	dense_94[0][0]
predA (Dense)	(None, 1)	33	featA[0][0]
predB (Dense)	(None, 1)	33	featB[0][0]
predC (Dense)	(None, 1)	33	featC[0][0]
predD (Dense)	(None, 1)	33	featD[0][0]
dense_95 (Dense)	(None, 64)	8,256	dropout_60[0][0]
concatenate_31 (Concatenate)	(None, 4)	0	predA[0][0], predB[0][0], predC[0][0], predD[0][0]
gate (Dense)	(None, 4)	260	dense_95[0][0]
multiply_15 (Multiply)	(None, 4)	0	concatenate_31[0][0], gate[0][0]
Y_MoE (Lambda)	(None, 1)	0	multiply_15[0][0]

6 Justification of the Proposed Model

The suggested hybrid MoE architecture is intended to tackle the main issues that arise in intricate real world, time series forecasting problems, including noise, nonlinear patterns, abrupt fluctuations, and diverse dataset behavior. Because traditional single deep models (LSTM, GRU, or CNN alone) learn a single representation for all data patterns, they frequently fail to generalize across various temporal dynamics. In cases when the relationships change over time, the model performance could be decreased.

To overcome these limitations, we introduce a MoE model where four specialized expert networks learn complementary temporal structures. Rather than forcing a single model to learn all possible variations, MoE enables conditional selective computation, improving accuracy and adaptability. Each expert integrates CNN for feature extraction and either LSTM or BiGRU for temporal dependency modeling, allowing the network to simultaneously learn both global long term and local short term dependencies. The use of attention mechanisms further enhances interpretability and precision by prioritizing the most relevant time steps and reducing noise effects.

The gating MLP plays a crucial role by assigning adaptive weights to each expert based on the input characteristics. Instead of averaging predictions blindly, the gating function (Equation 4) ensures that the most appropriate expert dominates the final response. This is important because different sequence patterns (stable, fluctuating, bursty) cannot be handled equally well by a single recurrent architecture. Therefore, gating solves the core problem of model specialization and dynamic adaptation. Although DL models produce powerful feature representations, they sometimes struggle to capture residual nonlinear relationships or poorly represented outlier behaviors. For this reason, a CatBoost stacking layer is employed to refine the MoE output. CatBoost, with its gradient-boosted decision tree structure, efficiently captures remaining nonlinear patterns and strengthens generalization. This combination fuses deep sequential learning with strong classical ML optimization, offering a more stable and high precision final prediction.

7 Experimental Results

In this section, a comprehensive evaluation of the proposed framework is presented and compared with conventional ML and DL models. The evaluation is conducted using MSE, RMSE, and the R^2. The baseline ML models include Linear Regression, Random Forest, Gradient Boosting, XGBoost, CatBoost and Support Vector Regression (SVR) with an RBF kernel.

As shown in Table 3, ensemble-based models generally outperform linear approaches due to their ability to capture nonlinear relationships in the data. Among the baseline methods, SVR (RBF) achieves the highest R^2 value of 0.9469, followed by Gradient Boosting with an R^2 of 0.9401. CatBoost almost achieves R^2 of 0.9366. Linear Regression exhibits comparatively lower performance, indicating its limited capacity to model complex patterns in dynamic 5G resource allocation scenarios. To evaluate the effectiveness of deep learning architectures, Artificial Neural Networks (ANN) and Deep Neural Networks (DNN) are also considered. The ANN model demonstrates poor generalization performance with a relatively high MSE of 0.00310 and a low R^2 of 0.6460. Although the DNN model improves upon ANN, achieving an MSE of 0.00107 and an R^2 of 0.8779, it still underperforms compared to traditional ensemble-based machine learning approaches.

The proposed MoE-CB framework achieves the best overall performance among all evaluated models. It attains the lowest MSE of 0.000405, the lowest RMSE of 0.020125, and the highest R^2 value of 0.9502. Its also visible that proposed MoE also performed less than the proposed model. This represents a clear improvement over the strongest baseline model (SVR), demonstrating superior predictive accuracy and enhanced generalization capability. Also the MoE without stack demostrate poor performance. Overall, the experimental results validate the effectiveness of the proposed MoE-CB framework. By combining specialized expert learning with ensemble optimization, the model achieves robust prediction accuracy and establishes a strong benchmark for intelligent resource allocation in next-generation 5G networks. Figure: 6, 7 &8 demonstrate the experimental result of different models along with the proposed model.

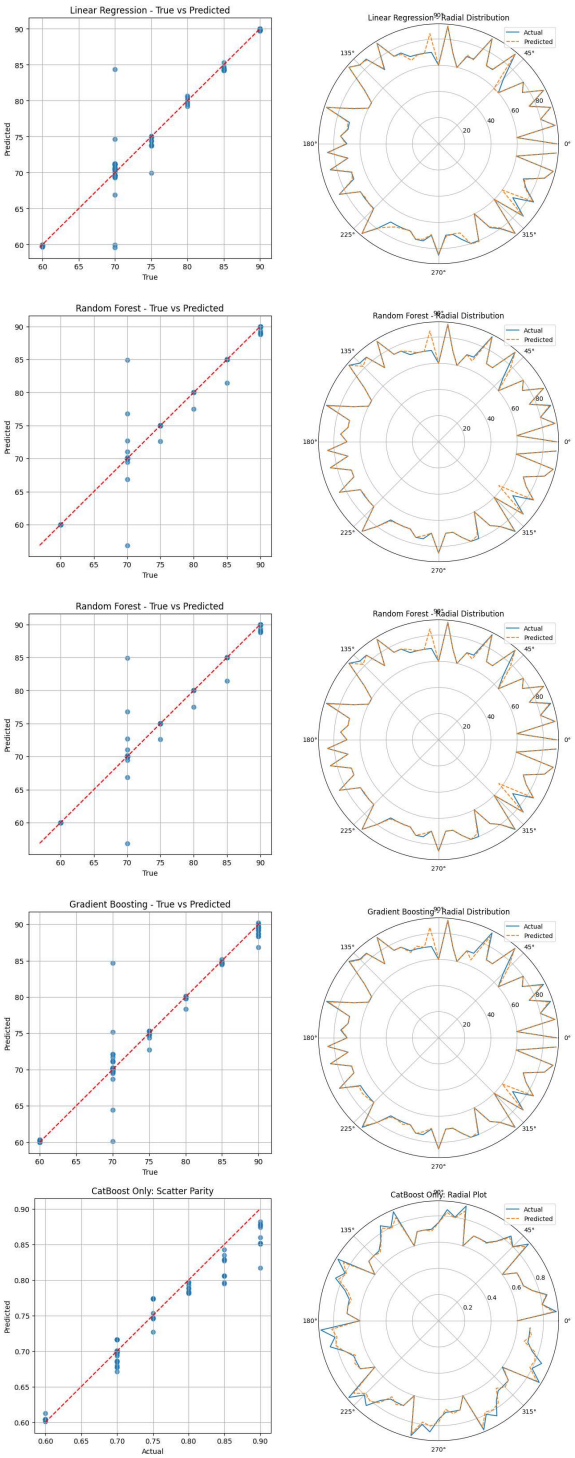


Fig. 6 Comparison of models under different conditions.

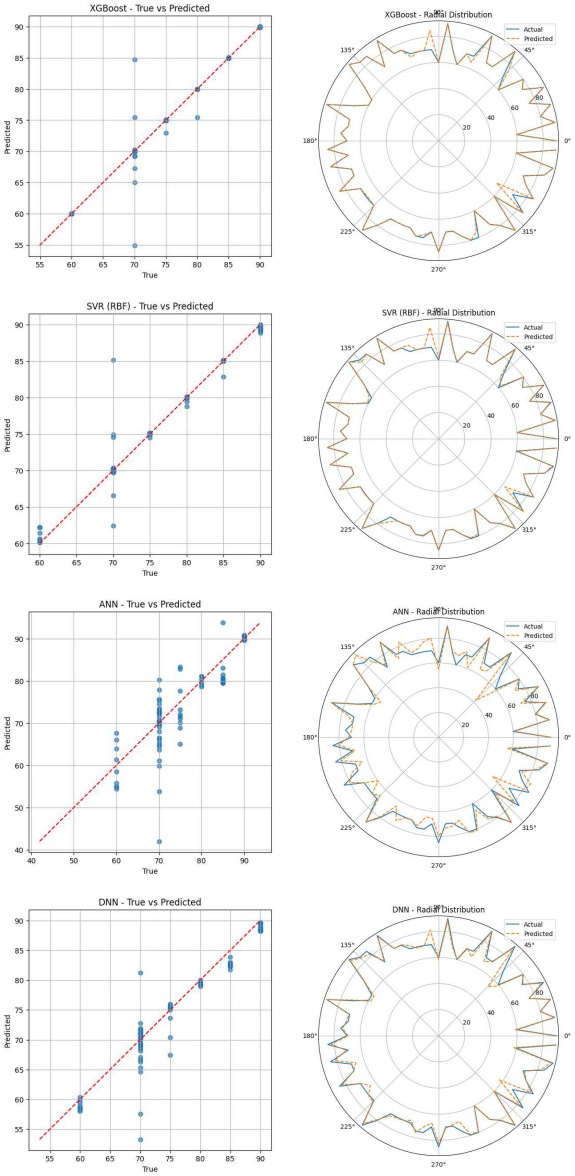


Fig. 7 Comparison of models under different conditions.

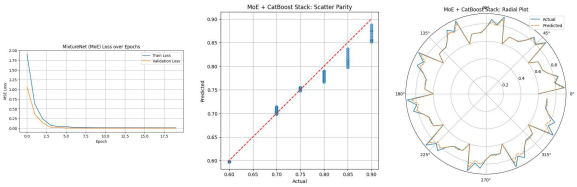


Fig. 8 Results of the proposed MoE-CB model.

8 State of Arts

Table 4 presents a comparison between the proposed MoE-CB model and recent studies conducted on the same 5G Resource Allocation dataset. The study by Singh reported strong results using KNN ($R^2 = 0.9378$, MSE = 0.0034), while Gaurav et al. further enhanced KNN with the Osprey Optimization Algorithm, achieving an MSE of 0.008 and R^2 above 0.9. Although these approaches demonstrate the effectiveness of distance-based learning methods, the proposed MoE-CB model achieves lower prediction error (MSE = 0.000405) and improved goodness of fit ($R^2 = 0.9502$), indicating enhanced modeling of nonlinear relationships within the dataset.

In comparison, the hybrid RL-CNN framework reports substantially lower regression performance ($R^2 = 0.517$, MSE = 0.035), suggesting that highly complex architectures may not generalize effectively on relatively small structured datasets. Overall, the results highlight the advantage of ensemble-based expert learning in delivering accurate, stable, and computationally efficient predictions for 5G resource allocation tasks.

9 Conclusion

This paper presents a reliable and scalable ML framework for intelligent resource allocation using a MoE based CatBoost (MoE-CB) architecture. Unlike the traditional model, the proposed method uses multiple expert models instead of a single one. This allows the model to fit in the dynamic 5G environments efficiently. Extensive experiments on the 5G resource allocation dataset show that the proposed MoE-CB model performs better than existing methods.

Extensive experimentation on the 5G resource allocation dataset demonstrates that the proposed MoE-CB model consistently outperforms existing state of the art methods, including KNN, RL-CNN, and hybrid optimization-based approaches. The model achieves a significantly lower MSE and RMSE, while attaining the highest coefficient of determination. These results indicate superior predictive stability, reduced estimation variance, and improved generalization capability.

The comparative analysis further highlights that while traditional models may provide acceptable accuracy, they struggle to maintain consistent performance under varying network conditions. In contrast, the proposed MoE-CB framework demonstrates strong robustness and adaptability, making it more suitable for real-time 5G resource orchestration scenarios.

From a practical perspective, the lightweight computational structure of CatBoost combined with expert specialization makes the model feasible for deployment in intelligent RAN controllers and data driven network optimization platforms.

10 Statements & Declarations

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Author Contributions

Soham Biswas and Partho Barui contributed to the conception and design. Gokulraj Pandiyan guided and provided suggestions. Code design and material preparation were done by Soham Biswas. Data collection and result analysis were performed by Partho Barui. The first draft of the manuscript was written by Soham Biswas and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

[1] Sufyan, A., Khan, K.B., Khashan, O.A., Mir, T., Mir, U.: From 5g to beyond 5g: A comprehensive survey of wireless network evolution, challenges, and promising technologies. Electronics **12**(10) (2023) https://doi.org/10.3390/electronics12102200

[2] Das, L., Sahoo, B.M., Rana, A., Dadhich, K., Sharma, S., Yadav, S.A.: In: Rai, A., Kumar Singh, D., Sehgal, A., Cengiz, K. (eds.) Application of AI & ML in 5G Communication, pp. 149–170. Springer, Singapore (2023). https://doi.org/

Table 3 Performance comparison of ML, DL, and proposed ensemble models (Target normalized to 0–1)

Model	MSE	RMSE	R^2
Machine Learning Models			
Linear Regression	0.000620	0.024902	0.929227
Random Forest	0.000612	0.024734	0.930179
Gradient Boosting	0.000525	0.022907	0.940113
XGBoost	0.000669	0.025859	0.923684
SVR (RBF)	0.000465	0.021570	0.946899
CatBoost	0.000515	0.022700	0.936663
Deep Learning Models			
ANN	0.003102	0.055694	0.645993
DNN	0.001070	0.032704	0.877932
Proposed Model MoE-CB			
MoE(without CatBoost stack)	0.000574	0.023950	0.9294
MoE-CB	**0.000405**	**0.020125**	**0.9502**

Table 4 State-of-the-Art Comparison on 5G Resource Allocation Dataset

Authors	Model	MSE	RMSE	R^2
Singh, Sachin [16]	KNN	0.0034	0.0583	0.9378
Karuppiyan et al. [17]	RL-CNN	0.0350	0.1880	0.5170
Gaurav et al. [18]	KNN + OOA	0.0080	0.0894	>0.9000
Proposed Work	**MoE-CB**	**0.000405**	**0.020125**	**0.9502**

10.1007/978-981-99-0109-8_9 . https://doi.org/10.1007/978-981-99-0109-8_9

[3] Balasubramanian, B., Daniels, E.S., Hiltunen, M., Jana, R., Joshi, K., Sivaraj, R., Tran, T.X., Wang, C.: Ric: A ran intelligent controller platform for ai-enabled cellular networks. IEEE Internet Computing **25**(2), 7–17 (2021) https://doi.org/10.1109/MIC.2021.3062487

[4] Cai, W., Jiang, J., Wang, F., Tang, J., Kim, S., Huang, J.: A survey on mixture of experts. Authorea Preprints (2024)

[5] Xu, Y., Wang, J., Zhang, R., Zhao, C., Niyato, D., Kang, J., Xiong, Z., Qian, B., Zhou, H., Mao, S., Jamalipour, A., Shen, X., Kim, D.I.: Mixture of experts for decentralized generative ai and reinforcement learning in wireless networks: A comprehensive survey. IEEE Communications Surveys & Tutorials **28**, 4051–4085 (2026) https://doi.org/10.1109/COMST.2025.3641591

[6] Behravesh, R., Coroama, D., Nikroo, M., Gholami, M., Mitra, R.N.: Demo xapp: Qos-based radio resource optimization on carrier-grade o-ran testbed. In: 2025 IEEE 11th International Conference on Network Softwarization (NetSoft), pp. 498–500 (2025). https://doi.org/10.1109/NetSoft64993.2025.11080593

[7] Liu, Z., Chen, X., Chen, Y., Li, Z.: Deep reinforcement learning based dynamic resource allocation in 5g ultra-dense networks. In: 2019 IEEE International Conference on Smart Internet of Things (SmartIoT), pp. 168–174 (2019). https://doi.org/

10.1109/SmartIoT.2019.00034

[8] Khan, A.A., Abolhasan, M., Ni, W., Lipman, J., Jamalipour, A.: A hybrid-fuzzy logic guided genetic algorithm (h-flga) approach for resource optimization in 5g vanets. IEEE Transactions on Vehicular Technology **68**(7), 6964–6974 (2019) https://doi.org/10.1109/TVT.2019.2915194

[9] Rico, D., Merino, P.: A survey of end-to-end solutions for reliable low-latency communications in 5g networks. IEEE Access **8**, 192808–192834 (2020) https://doi.org/10.1109/ACCESS.2020.3032726

[10] Xu, Y., Gui, G., Gacanin, H., Adachi, F.: A survey on resource allocation for 5g heterogeneous networks: Current research, future trends, and challenges. IEEE Communications Surveys and Tutorials **23**(2), 668–695 (2021) https://doi.org/10.1109/COMST.2021.3059896

[11] Selvamanju, E., Shalini, V.B.: Archimedes optimization algorithm with deep belief network based mobile network traffic prediction for 5g cellular networks. In: 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 370–376 (2022). https://doi.org/10.1109/ICSSIT53264.2022.9716552

[12] Goswami, P., Mukherjee, A., Maiti, M., Tyagi, S.K.S., Yang, L.: A neural-network-based optimal resource allocation method for secure iiot network. IEEE Internet of Things Journal **9**(4), 2538–2544 (2022) https://doi.org/10.1109/JIOT.2021.3084636

[13] Gao, X., Wang, J., Zhou, M.: The research of resource allocation method based on gcn-lstm in 5g network. IEEE Communications Letters **27**(3), 926–930 (2023) https://doi.org/10.1109/LCOMM.2022.3224213

[14] Yarkina, N., Gaydamaka, A., Moltchanov, D., Koucheryavy, Y.: Performance assessment of an itu-t compliant machine learning enhancements for 5g ran network slicing. IEEE Transactions on Mobile Computing **23**(1), 719–736 (2024) https://doi.org/10.1109/TMC.2022.3228286

[15] Pathak, P., Oad, V., Prajapati, A., Innan, N.: Resource allocation optimization in 5g networks using variational quantum regressor. In: 2024 International Conference on Quantum Communications, Networking, and Computing (QCNC), pp. 101–105 (2024). https://doi.org/10.1109/QCNC62729.2024.00025

[16] Singh, S.: Performance evaluation of machine learning regression models for 5g network resource allocation optimization. In: 2025 International Conference on Multimedia Computing, Networking and Applications (MCNA), pp. 109–114 (2025). https://doi.org/10.1109/MCNA65829.2025.11124386

[17] Karuppiyan, M., Subramani, H., Raju, S.K., Prakasam, M.M.A.: Dynamic resource allocation in 5g networks using hybrid rl-cnn model for optimized latency and quality of service. Network: Computation in Neural Systems **36**(2), 249–273 (2025) https://doi.org/10.1080/0954898X.2024.2334282 https://doi.org/10.1080/0954898X.2024.2334282. PMID: 38594948

[18] Gaurav, A., Gupta, B.B., Chui, K.T.: Ai-enhanced resource allocation in 5g-advanced networks using knn with osprey optimization for next-generation mobile services. In: 2024 IEEE Conference on Standards for Communications and Networking (CSCN), pp. 322–323 (2024). https://doi.org/10.1109/CSCN63874.2024.10849762

[19] 5G Resource Allocation Dataset:Optimizing Band — kaggle.com. https://www.kaggle.com/datasets/omarsobhy14/5g-quality-of-service/data. [Accessed 14-10-2025]

[20] Wu, T., Wong, N.: A unified view for attention and moe