

Predictive Network-State Classification in 5G Mobile Communication Networks: A Comparative Random Forest and XGBoost Ensemble Learning Approach

ATU Michael Osadebamwen^{1*}, Fred Izilein², Onyegbadue Ikenna Augustine¹

¹ Department of Electronic and Computer Engineering, Igbinedion University, Okada, Edo State, Nigeria

Corresponding Author: ATU Michael Osadebamwen

Abstract

The rapid densification and service diversification of fifth-generation (5G) mobile communication networks require reliable awareness of network operating conditions for adaptive resource management. This study develops and evaluates a machine-learning approach for classifying network states into Critical, Moderate, and Optimal categories. A controlled synthetic 5G dataset of 10,000 observations was generated with a fixed random seed and explicitly specified radio, traffic, resource, and quality-of-service relationships. A composite Adaptive Optimization Score (AOS) derived the three-state target, while the five AOS constituent indicators were excluded from the predictors to prevent target leakage, leaving 15 radio- and resource-management features. Random Forest (RF) and Extreme Gradient Boosting (XGBoost) models were selected after a compact training-only hyperparameter search and assessed using a stratified 80:20 hold-out split, stratified five-fold cross-validation, confusion-matrix analysis, feature importance, AOS sensitivity testing, and McNemar's paired significance test. On the held-out synthetic test set, RF achieved 95.65% accuracy and XGBoost 96.45%; five-fold mean accuracies were $94.25 \pm 0.60\%$ and $95.36 \pm 0.44\%$, respectively. McNemar's exact test indicated that the held-out difference was statistically significant ($p = 0.0259$). Sensitivity analysis showed that AOS class assignments remained highly concordant under reasonable weight perturbations, although wider threshold shifts materially changed class prevalence. To address external generalizability, an independent robustness experiment was also conducted on 8,736 measurements from the public SUTD private-5G testbed using eight directly measured KPIs and the dataset's native anomaly label. Stratified five-fold accuracy was 88.03% for RF and 86.14% for XGBoost, while leave-one-scenario-out accuracy fell to 64.29% and 66.47%, respectively, demonstrating substantial scenario-shift effects. The results support ensemble learning as a useful predictive-intelligence layer under controlled conditions, while showing that synthetic accuracy should not be interpreted as deployed-network performance without external, scenario-aware validation.

Keywords

5G Mobile Networks; Network-State Classification; Random Forest; XGBoost; Ensemble Machine Learning; Predictive Intelligence; Feature Importance

Date of Submission: 02-10-2026

Date of Acceptance: 12-10-2026

I. Introduction

Fifth-generation (5G) mobile communication networks are designed to support a substantially wider range of service requirements than earlier network generations, spanning enhanced mobile broadband, ultra-reliable low-latency communication, and massive machine-type communication within a single, shared radio and core infrastructure. This diversification, together with dense small-cell deployment, heterogeneous user mobility, and rapidly fluctuating traffic demand, has made the operating condition of a 5G network highly dynamic and difficult to characterize using static, rule-based thresholds. Network operators consequently require a means of continuously and automatically recognizing whether a given segment of the network is operating in a healthy, moderately stressed, or critically degraded condition, so that appropriate management action can be triggered before service quality is visibly affected.

Traditional network management approaches rely on manually configured thresholds applied to individual performance counters, such as flagging a cell as congested once utilization exceeds a fixed percentage. Such approaches struggle to capture the joint, nonlinear interaction among the many radio and resource-management indicators that jointly determine 5G network condition; signal quality, interference, resource utilization, mobility, and scheduling efficiency do not degrade independently, and a network state that appears acceptable on any single indicator may still be operationally critical when indicators are considered jointly. This motivates a data-driven, multivariate classification approach in which a trained model learns the relationship

between a wide feature set describing the current radio and resource condition and a categorical label describing the resulting operating state.

Machine learning, and ensemble tree-based learning in particular, is well suited to this classification problem. Ensemble methods such as Random Forest and Extreme Gradient Boosting (XGBoost) construct predictions from many individual decision trees, allowing them to model complex, nonlinear interactions among heterogeneous network features without requiring an explicit mathematical model of network behavior, while remaining considerably more interpretable, through feature-importance analysis, than deep neural architectures. These properties make ensemble classifiers a practical first step toward predictive network intelligence: a model that can be trained on historical or simulated network observations and subsequently used to classify the current or near-future network state in real time.

Despite the growing body of work applying machine learning to individual 5G prediction problems, including traffic-volume forecasting, congestion prediction, and quality-of-service estimation, the specific combination of a three-class network-state target, a broad radio-plus-resource feature set, explicit leakage prevention, and a controlled Random Forest versus XGBoost comparison remains comparatively underexplored. The contribution claimed here is therefore methodological and integrative rather than a new learning algorithm: the study combines target construction, leakage-aware model development, statistical comparison, sensitivity analysis, and external robustness testing within one 5G network-state classification workflow.

This paper is deliberately scoped to the predictive-intelligence layer that underpins adaptive network optimization. The primary experiment develops the synthetic dataset, AOS-based target, preprocessing and leakage-control pipeline, and Random Forest and XGBoost classifiers, and evaluates them on classification grounds. A secondary experiment uses an independent public 5G measurement dataset to test whether the two ensemble families retain useful discriminatory capability under real measurement noise and scenario shift. The external experiment is an algorithm-level robustness check rather than an exact reproduction of the 15-feature AOS three-class task. The paper does not evaluate, and does not claim, downstream improvements in network throughput, latency, packet loss, spectral efficiency, or energy consumption resulting from optimization actions.

The specific objectives of this study are to: (i) construct a reproducible synthetic 5G network dataset covering representative radio, traffic, and resource-management conditions; (ii) engineer a composite operating-condition score and derive a three-class network-state target while preventing target leakage; (iii) develop and compare Random Forest and XGBoost classifiers; (iv) strengthen model validation through hyperparameter selection, stratified five-fold cross-validation, and paired statistical testing; (v) assess the sensitivity of the AOS labels to alternative weights and class thresholds; (vi) identify the network indicators that most strongly drive classification; and (vii) evaluate external robustness using independently collected public 5G measurements. The study makes the following specific contributions to predictive network-state classification; these are presented as an integrated methodological contribution rather than as claims of novelty in the underlying Random Forest or XGBoost algorithms.

1.1 Study Contributions

- Leakage-aware target construction: a reproducible procedure for deriving a categorical network-operating-state target from continuous 5G performance indicators while removing the target's constituent variables from model inputs.
- Explicit AOS specification and sensitivity analysis: the weighted AOS formulation and Critical/Moderate/Optimal thresholds are stated mathematically and tested under alternative weight and threshold settings rather than treated as immutable design choices.
- Three-state predictive framework: an operationally interpretable Critical/Moderate/Optimal categorization of network condition based on a broad set of radio, traffic, and resource-management predictors.
- Controlled RF-XGBoost comparison with strengthened validation: both ensemble models are compared under identical data conditions using training-only hyperparameter selection, stratified five-fold cross-validation, held-out testing, and McNemar's paired test.
- Feature-importance interpretation: Gini-based ranking links model behaviour to physically interpretable radio and resource-management variables rather than treating the classifier as a black box.
- Independent external robustness assessment: RF and XGBoost are retrained and evaluated on public real 5G measurements using native anomaly labels and both stratified and leave-one-scenario-out validation, while explicitly distinguishing this robustness test from exact external validation of the AOS three-class task.

II. Related Work

2.1 Machine Learning for 5G Network Prediction

Machine learning has been increasingly applied to prediction problems in cellular and 5G networks as operators seek to move from reactive to proactive network management. Hussien et al. (2025) developed and compared eight predictive models, spanning statistical, machine-learning, and deep-learning families, including AdaBoost and XGBoost, for cellular network traffic prediction using a large real-world call-detail-record dataset,

finding that gradient-boosted tree models offered a favorable balance between predictive accuracy and computational efficiency relative to deep sequence models. This supports the use of tree-based ensembles as a computationally practical choice for network-condition prediction problems, motivating their selection in the present study.

Palaios et al. (2023) examined the broader challenge of machine-learning-based quality-of-service prediction in vehicular 5G communication, highlighting the difficulty of building models that generalize across heterogeneous radio conditions and mobility patterns, and underlining the importance of feature selection and leakage control when the prediction target is itself derived from network performance indicators. This concern with target leakage is directly relevant to network-state classification tasks, where, as in the present study, the categorical label is engineered from the same underlying performance measurements that also inform candidate input features.

Alekseeva et al. (2021) conducted a comparative evaluation of several machine-learning techniques, including gradient boosting and support vector regression, applied to traffic prediction on a real wireless network, and reported that ensemble and boosting methods consistently outperformed simpler regression baselines across the evaluated network segments. These findings reinforce the case for evaluating more than one ensemble algorithm, rather than assuming the superiority of a single technique, an approach adopted in this study through its direct Random Forest-XGBoost comparison.

2.2 Network-State and Congestion Classification

A smaller body of work addresses network condition or congestion as an explicit classification target rather than a continuous forecasting problem. Studies in this direction have applied decision-tree-based classifiers to predict congestion states in wireless and Internet-of-Things-integrated 5G environments from traffic, signal, and node-density features, reporting that tree-based classifiers can reliably distinguish congested from non-congested operating conditions while remaining computationally lightweight enough for near-real-time deployment. This body of work supports the use of a categorical operating-state target, as opposed to forecasting each performance indicator individually, as a tractable and operationally meaningful prediction problem, consistent with the three-class Critical-Moderate-Optimal framing adopted in the present study.

2.3 Ensemble Learning: Random Forest and XGBoost

Random Forest, introduced by Breiman (2001), constructs an ensemble of decision trees trained on bootstrap-resampled subsets of the training data and randomly restricted feature subsets at each split, combining their predictions by majority vote. Its principal strengths for network-condition classification are robustness to overfitting on moderate-sized tabular datasets, native handling of nonlinear feature interactions without feature transformation, and a built-in, Gini-impurity-based feature-importance measure that supports model interpretability (Molnar, 2022). XGBoost, introduced by Chen and Guestrin (2016), instead builds trees sequentially, with each successive tree trained to correct the residual error of the ensemble so far, combined with regularization terms that constrain model complexity. This sequential, gradient-boosted construction typically yields higher predictive accuracy on structured, tabular data than a comparably sized Random Forest, at the cost of greater sensitivity to hyperparameter configuration.

Grabowski and Xu (2024) applied both Random Forest and XGBoost to network intrusion classification, reporting that both ensemble families achieved strong classification performance while offering complementary interpretability characteristics, with Random Forest's native Gini-importance ranking proving straightforward to compute and communicate. This supports the pairing of Random Forest and XGBoost specifically, rather than either algorithm in isolation, as an informative and interpretable basis for network-state classification, the comparative design adopted in this study.

2.4 Research Gap

Taken together, the reviewed literature demonstrates strong precedent for applying ensemble tree-based learning to cellular and 5G prediction problems. What remains useful to examine is the specific integrated workflow adopted here: a broad leakage-controlled feature set used to classify operational state into a small number of actionable categories, a like-for-like RF-XGBoost comparison with statistical validation, explicit testing of target-definition sensitivity, and a transparent distinction between controlled synthetic feasibility and real-measurement robustness. This study addresses that methodological combination without claiming that any individual algorithmic component is itself transformative.

Public real-world 5G measurement datasets also provide an important check on simulation-only conclusions. The SUTD 5G dataset was collected on a private 5G testbed at the Singapore University of Technology and Design using a Samsung Galaxy S22 Ultra and Keysight Nemo measurement tools. It reports RSRP, RSRQ, SINR, modulation-and-coding indices, physical resource blocks, downlink throughput, and manually defined anomaly labels across four coverage and RRU scenarios (FCCLab, 2023). The present study

uses these measurements as an external robustness benchmark while preserving the original synthetic three-state experiment as the primary task.

III. Materials and Methods

This section presents the synthetic-data generation process, external real-world dataset, preprocessing and leakage-control procedures, AOS target construction, model development, hyperparameter selection, validation strategy, and evaluation metrics. Figure 1 summarizes the revised workflow, including both the primary synthetic three-state experiment and the secondary external robustness experiment.

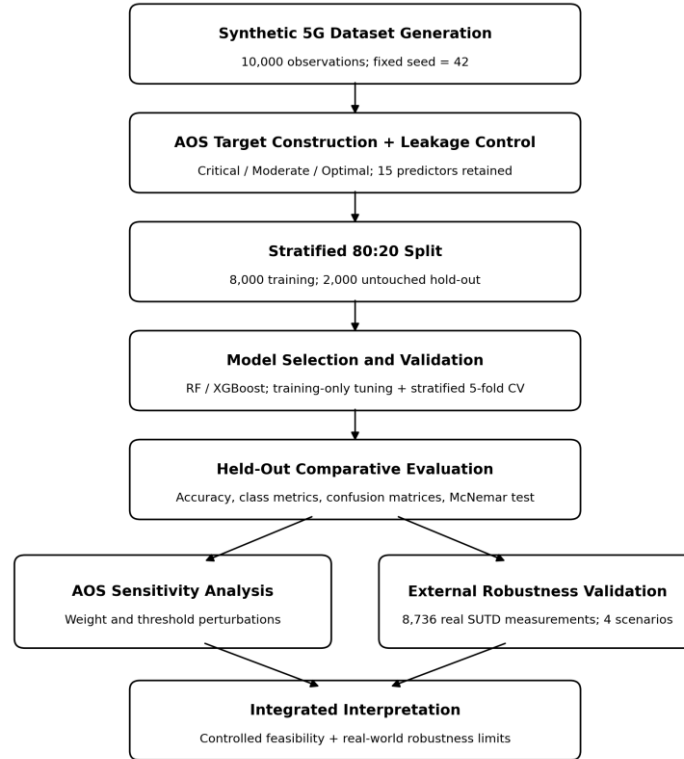


Figure 1. Research Methodology Flowchart for Predictive Network-State Classification.

3.1 Dataset Development

A synthetic 5G network dataset was developed to provide a controlled and reproducible basis for the primary three-state classification experiment. All stochastic values were generated with NumPy's default random-number generator (default_rng) initialized once with seed = 42, and the same generator instance was used sequentially throughout dataset construction in the variable-generation order reported below. The procedure produced 10,000 independent observations. Primary inputs were sampled as follows: user density U was discrete-uniform from 50 to 1,000 users; traffic demand D was uniform on [50, 1000] Mbps; interference power I was uniform on [-110, -70] dBm; signal power S was uniform on [-95, -45] dBm; bandwidth B was sampled with equal probability from {20, 40, 60, 80, 100} MHz; resource utilization R was uniform on [20, 100]%; and mobility speed V was uniform on [0, 120] km/h. SINR was defined as $S - I$.

Dependencies among generated variables were explicitly introduced rather than assuming independent KPI sampling. A congestion factor was defined as $C = 0.45(U/1000) + 0.35(D/1000) + 0.20(R/100)$. Spectral efficiency was generated as $\text{clip}\{\log_2[1 + 10^{(SINR/10)}](1 - 0.35C), 0.1, 8.0\}$. Throughput was $T = \text{clip}\{B \times SE \times (1 - 0.40C) + e_T, 1, \text{infinity}\}$, with $e_T \sim N(0, 8^2)$. Latency was $L = \text{clip}\{5 + 75C + 0.12V - 0.35SINR + e_L, 1, \text{infinity}\}$, with $e_L \sim N(0, 4^2)$. Packet loss was $PL = \text{clip}\{0.2 + 7C - 0.08SINR + 0.015V + e_{PL}, 0, 20\}$, with $e_{PL} \sim N(0, 0.5^2)$. Energy consumption was $E = \text{clip}\{40 + 0.06U + 0.09D + 0.45B + e_E, 20, \text{infinity}\}$, with $e_E \sim N(0, 5^2)$.

Seven additional radio/resource variables were then derived. $RSRP = \text{clip}\{S + e_{RSRP}, -120, -44\}$, $e_{RSRP} \sim N(0, 3^2)$; $RSRQ = \text{clip}\{-3 - 0.12R - 0.004U + e_{RSRQ}, -20, -3\}$, $e_{RSRQ} \sim N(0, 1.5^2)$; $CQI = \text{clip}\{\text{round}[1 + (SINR + 10)14/40], 1, 15\}$; PRB utilization = $\text{clip}\{0.55R + 0.25(D/10) + 0.20(U/10) + e_{PRB}, 0, 100\}$, $e_{PRB} \sim N(0, 4^2)$; cell load = $\text{clip}\{0.45(U/10) + 0.35(D/10) + 0.20R + e_{CL}, 0, 100\}$, $e_{CL} \sim N(0, 3^2)$; handover success = $\text{clip}\{99 - 0.045V - 0.12PL + 0.05SINR + e_H, 80, 100\}$, $e_H \sim N(0, 0.8^2)$; and scheduler efficiency = $\text{clip}\{95 - 0.25(\text{cell load}) - 0.18PL + 0.30CQI + e_S, 40, 100\}$, $e_S \sim N(0, 2^2)$. The Gaussian draws for these additional

variables continued from the same initialized NumPy generator after the primary variables and QoS terms had been generated. Together, the stated distributions, equations, clipping rules, generation order, and fixed seed specify the reproducible synthetic-generation procedure used in the primary experiment.

Table 1. Network Features Used for Model Development

#	Feature	Description / Unit
1	User Density	Number of active users per cell (count)
2	Traffic Demand	Aggregate traffic demand (Mbps)
3	Interference	Interference power (dBm)
4	Signal Power	Received signal power (dBm)
5	SINR	Signal-to-Interference-plus-Noise Ratio (dB)
6	Bandwidth	Available channel bandwidth (MHz)
7	Resource Utilization	Resource-block utilization (%)
8	Mobility	User equipment mobility speed (km/h)
9	RSRP	Reference Signal Received Power (dBm)
10	RSRQ	Reference Signal Received Quality (dB)
11	CQI	Channel Quality Indicator (index)
12	PRB Utilization	Physical Resource Block utilization (%)
13	Cell Load	Cell load (%)
14	Handover Success Rate	Handover success rate (%)
15	Scheduler Efficiency	Scheduler efficiency (%)

The resulting data contain 20 raw network variables. After construction of the AOS target, the five target constituents and the AOS itself were excluded from model inputs, leaving the 15 leakage-free predictors listed in Table 1. The three-class target comprises Critical, Moderate, and Optimal states.

Synthetic generation was retained because it permits controlled stress conditions, exact regeneration, and transparent inspection of target leakage. It does not, however, reproduce all measurement error, temporal dependence, hardware effects, and unanticipated correlations present in live networks. Accordingly, synthetic-model accuracy is interpreted in this paper as controlled feasibility evidence rather than as an estimate of deployed-network performance.

3.1.1 External Real-World Dataset

External robustness was evaluated using the public SUTD 5G dataset (FCCLab, 2023), which contains measurements collected by walk test on a private 5G deployment at the Singapore University of Technology and Design. The four uploaded scenario files contain 8,736 observations in total: Level 4 with all RRUs active (2,327), Level 5 with all RRUs active (2,031), Level 6 with one active RRU (2,156), and Level 6 with all RRUs active (2,222). The dataset provides manually defined binary anomaly labels (lab_anom) that are independent of the AOS construction used in the primary experiment.

Eight directly measured KPIs were used as external predictors: RSRP, RSRQ, SINR, PDSCH MCS, PUSCH MCS, PDSCH PRBs, PUSCH PRBs, and downlink throughput. Time, cell identifiers, object identifiers, C-RNTI, corridor tags, the anomaly label, auxiliary fault labels, RRU-status labels, and the composite Label field were excluded from the predictor set to reduce leakage and scenario-identity shortcuts. Because the external feature schema and native binary target differ from the 15-feature AOS three-class task, this analysis is explicitly treated as an external algorithm-level robustness assessment, not as an exact external replication of the primary classification problem.

3.2 Data Preprocessing

For the synthetic dataset, data types, missing values, and duplicate observations were checked before feature engineering; no missing values or duplicate records were present. Although a Min-Max-normalized copy of the dataset was created during the broader research workflow for exploratory analysis, the final RF and XGBoost models reported in this paper were trained on the original numerical predictor scales because tree-based split rules do not require feature scaling. For the external SUTD experiment, the small number of missing SINR/MCS/PRB measurements was handled by median imputation fitted on the training portion only and then applied to the corresponding validation/test portion, preventing information leakage across folds or scenarios.

3.3 Feature Engineering and Target-Class Derivation

To convert the multidimensional performance profile of each synthetic observation into a single operating-condition indicator, the Adaptive Optimization Score (AOS) was defined on a 0-100 scale as $AOS = 100[0.30(T/T_{max}) + 0.25(1 - L/L_{max}) + 0.15(1 - PL/PL_{max}) + 0.20(SE/SE_{max}) + 0.10(1 - E/E_{max})]$. Here, T_{max} , L_{max} , PL_{max} , SE_{max} , and E_{max} are the maximum observed values of throughput, latency, packet loss, spectral efficiency, and energy consumption, respectively, computed from the generated 10,000-observation

synthetic dataset before class derivation. Throughput and spectral efficiency therefore contribute positively, while latency, packet loss, and energy consumption are inverted because lower raw values are preferable.

The continuous AOS was discretized into three states: AOS ≥ 80 was labelled Optimal, $60 \leq \text{AOS} < 80$ was labelled Moderate, and AOS < 60 was labelled Critical. Because the AOS and class labels are computed directly from throughput, latency, packet loss, spectral efficiency, and energy consumption, those five variables and the AOS itself were removed before training. The resulting input set therefore contains only the 15 predictors listed in Table 1.

The leakage-free feature set was partitioned into training and testing subsets using a stratified 80:20 split with `random_state = 42`, yielding an untouched held-out test set of 2,000 observations. Hyperparameter selection and five-fold validation were conducted using the 8,000-observation training partition only; the test set was reserved for the final comparison. To test the analyst-defined AOS parameters, four alternative weight schemes and both narrow and wider threshold shifts were also evaluated by recomputing the labels and measuring class-assignment agreement and Cohen's kappa relative to the baseline AOS specification.

3.4 Model Development

3.4.1 Random Forest

Random Forest is an ensemble bagging algorithm in which independently bootstrapped trees are combined by majority vote (Breiman, 2001). Candidate settings were compared on a stratified inner validation subset of the training data using macro F1-score, which was chosen because the Optimal class is comparatively small. The selected RF configuration used 200 trees, maximum depth 30, `max_features = 0.7`, minimum leaf size 1, balanced class weighting, and `random_state = 42`. This configuration was then evaluated by stratified five-fold cross-validation on the full training partition before final refitting and held-out testing.

3.4.2 Extreme Gradient Boosting (XGBoost)

XGBoost constructs trees sequentially so that each new tree reduces the residual classification error of the preceding ensemble (Chen & Guestrin, 2016). Candidate combinations of tree count, depth, learning rate, row subsampling, and column subsampling were compared on the same inner training validation partition using macro F1-score. The selected configuration used 300 estimators, maximum depth 4, learning rate 0.05, `subsample = 0.90`, `colsample_bytree = 0.90`, the multi-class probabilistic objective, and `random_state = 42`. It was then subjected to the same five-fold training-set validation and final held-out evaluation as RF.

3.4.3 Hyperparameter Selection and Validation

Random Forest and XGBoost were retained because they provide a controlled bagging-versus-boosting comparison for structured network telemetry. Hyperparameter selection was performed strictly within the training partition and did not use the held-out test labels. After model selection, stratified five-fold cross-validation (`shuffle = True`, `random_state = 42`) quantified variation in accuracy, macro F1, weighted F1, and balanced accuracy across training folds. Final models were then refitted on all 8,000 training observations and evaluated once on the 2,000-observation held-out test set. This separation prevents the test partition from influencing model selection.

3.5 Evaluation Metrics

Evaluation used accuracy, weighted precision, weighted recall, weighted F1-score, macro F1-score, balanced accuracy, confusion matrices, and feature importance. Macro F1 and balanced accuracy were included because class prevalence is unequal. For the primary held-out comparison, McNemar's exact paired test was applied to RF and XGBoost correct/incorrect outcomes on the same 2,000 observations; $p < 0.05$ was treated as evidence that the difference in error patterns was unlikely to be due to chance alone. External real-world evaluation additionally used ROC-AUC. Two validation schemes were used externally: stratified five-fold cross-validation across the pooled measurements and leave-one-scenario-out (LOSO) validation, in which each complete physical measurement scenario was withheld in turn to assess scenario transfer.

Table 2. Classification Evaluation Metrics

Metric	Formula	Interpretation
Accuracy	$(TP + TN) / (TP + TN + FP + FN)$	Overall proportion of observations correctly classified across all three states.
Precision (weighted)	$TP / (TP + FP)$, averaged and weighted by class support	Of all observations predicted as a given state, the proportion that truly belong to it.
Recall (weighted)	$TP / (TP + FN)$, averaged and weighted by class support	Of all observations that truly belong to a given state, the proportion correctly identified.
F1-score (weighted)	$2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$	Harmonic balance of precision and recall.
Confusion Matrix	Cross-tabulation of predicted vs. actual class	Reveals which specific state pairs are most frequently confused by the model.
Feature Importance	Gini-impurity-based importance (Random Forest)	Ranks input features by their average contribution to reducing classification impurity across all trees.
Macro F1	Unweighted mean of class-specific F1 scores	Gives equal importance to Critical, Moderate, and Optimal classes.
Balanced Accuracy	Mean recall across classes	Reduces dominance of the majority Critical class.

IV. Results and Discussion

This section reports the strengthened primary synthetic classification results, including cross-validation, held-out testing, paired statistical comparison, AOS sensitivity, and feature interpretation, followed by the independent external real-world robustness experiment. No downstream network-optimization performance claim is made in this paper.

4.1 Synthetic Classification Performance

Table 3 reports the final held-out performance after training-only hyperparameter selection. RF achieved 95.65% accuracy, while XGBoost achieved 96.45%. The weighted metrics remained closely aligned with overall accuracy for both models.

Table 3. Comparative Classification Performance of Random Forest and XGBoost

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	95.65	95.61	95.65	95.59
XGBoost	96.45	96.42	96.45	96.43

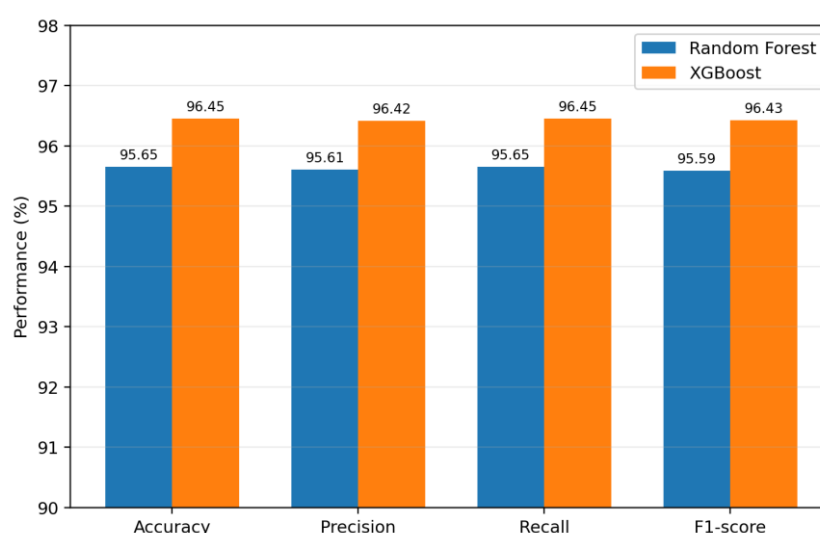


Figure 2. Comparative Classification Performance of Random Forest and XGBoost on the Held-Out Synthetic Test Set.

Figure 2 visualizes the held-out results in Table 3. The difference is modest in absolute terms but consistent across accuracy, weighted precision, weighted recall, and weighted F1-score.

Both classifiers also remained strong under stratified five-fold cross-validation of the training partition. RF achieved $94.25 \pm 0.60\%$ mean accuracy, $88.53 \pm 0.80\%$ macro F1, and $86.62 \pm 1.82\%$ balanced accuracy. XGBoost achieved $95.36 \pm 0.44\%$ mean accuracy, $90.96 \pm 0.87\%$ macro F1, and $89.43 \pm 0.97\%$ balanced accuracy. On the final held-out test observations, RF was correct where XGBoost was wrong in 15 cases, while XGBoost was correct where RF was wrong in 31 cases. McNemar's exact test gave $p = 0.0259$, indicating a statistically significant difference at the 5% level for this controlled synthetic experiment. The effect size remains small and should not be interpreted as universal algorithmic superiority.

4.2 Confusion-Matrix and Class-Wise Performance

Figure 3 illustrates the distribution of the three network-state classes within the 2,000-observation test set. The Critical class is the majority class (67.2%), followed by Moderate (29.0%) and Optimal (3.9%), reflecting the class imbalance considered in the macro-F1 and balanced-accuracy analyses.

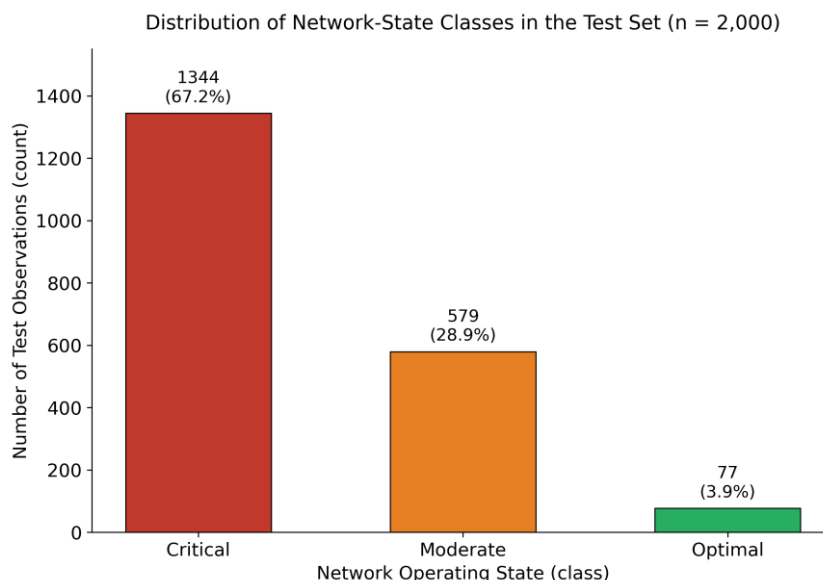


Figure 3. Distribution of Network-State Classes in the Synthetic Test Set (n = 2,000).

Table 4 reports class-wise precision, recall, F1-score, and support for the selected XGBoost model. Critical-state recall remained high, while the small Optimal class remained the most difficult state to detect.

Table 4. XGBoost Class-Wise Classification Performance

Network State	Precision (%)	Recall (%)	F1-Score (%)	Support
Critical	98.14	98.36	98.25	1,344
Moderate	93.49	94.30	93.90	579
Optimal	88.41	79.22	83.56	77
Overall Accuracy	96.45	-	-	2,000
Macro Average	93.35	90.63	91.90	2,000
Weighted Average	96.42	96.45	96.43	2,000

Table 5. XGBoost Confusion Matrix (rows = actual, columns = predicted)

Actual \ Predicted	Critical	Moderate	Optimal	Row Total
Critical	1,322	22	0	1,344
Moderate	25	546	8	579
Optimal	0	16	61	77

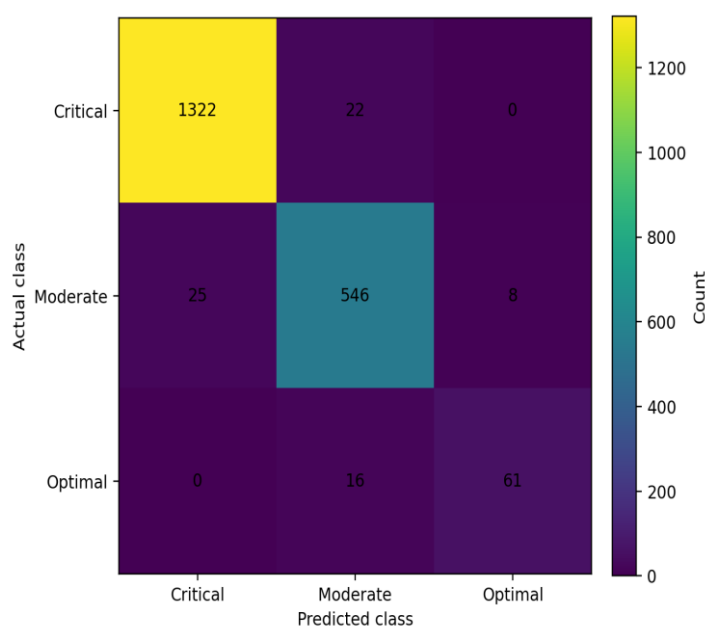


Figure 4. XGBoost Confusion Matrix Heatmap for the Held-Out Synthetic Test Set (n = 2,000).

Figure 4 presents the Table 5 confusion matrix as a heatmap. Errors remain concentrated between adjacent operating states, with no direct Critical-to-Optimal or Optimal-to-Critical confusion.

The Critical class was classified with 98.36% recall: 1,322 of 1,344 Critical observations were correctly identified and 22 were classified as Moderate. The Moderate class achieved 94.30% recall, with 546 of 579 observations correctly identified, 25 classified as Critical, and 8 as Optimal. The Optimal class achieved 79.22% recall, with 61 of 77 observations correctly identified and 16 classified as Moderate. The lower Optimal-class recall reflects its small support and reinforces the value of macro F1 and balanced accuracy alongside overall accuracy (He & Garcia, 2009). Importantly, all errors involving the Optimal class remained adjacent-state errors rather than extreme Critical-Optimal confusions.

4.3 Comparative Discussion

The primary results indicate that both Random Forest and XGBoost are strong classifiers under the controlled synthetic task, while XGBoost retains a small but statistically supported advantage after training-only hyperparameter selection. Five-fold validation shows that this difference is not confined to a single train-test partition, but the absolute margin remains modest.

Because the held-out accuracy gap is only 0.80 percentage points, model selection should consider operational factors in addition to raw accuracy. RF offers parallel training and direct Gini-importance interpretation, while XGBoost provides somewhat stronger macro F1 and balanced accuracy in the synthetic experiment. The McNemar result supports a difference on this test set, but the external experiment in Section 4.6 shows that relative performance changes with measurement scenario and validation design; the synthetic ranking should therefore not be generalized uncritically to all deployments.

For AI-driven optimization in 5G mobile networks, the classification layer is useful only insofar as it provides reliable state awareness to a downstream controller. The present results support feasibility of that predictive layer under controlled conditions, particularly for the Critical class, but this paper does not demonstrate the performance of any downstream resource-allocation action. External robustness and deployment limitations are therefore treated separately below.

4.4 Feature-Importance Analysis

Gini-based feature-importance scores were extracted from the trained Random Forest model to provide an interpretable ranking of the network parameters most strongly associated with the predicted operating state. Table 6 presents the resulting ranking of the most influential features identified by the analysis.

Table 6. Random Forest Feature-Importance Ranking (Gini-Based)

Rank	Feature	Importance Score
1	SINR (dB)	0.299140
2	Bandwidth (MHz)	0.211416
3	Cell Load (%)	0.204178
4	Channel Quality Indicator (CQI)	0.079000
5	Scheduler Efficiency (%)	0.068383
6	Handover Success Rate (%)	0.020705
7	Traffic Demand (Mbps)	0.020497
8	User Density (count)	0.016732
9	Mobility (km/h)	0.016723
10	PRB Utilization (%)	0.014165
11	Signal Power (dBm)	0.012080
12	Interference (dBm)	0.010642
13	RSRQ (dB)	0.009514
14	RSRP (dBm)	0.009071
15	Resource Utilization (%)	0.007754

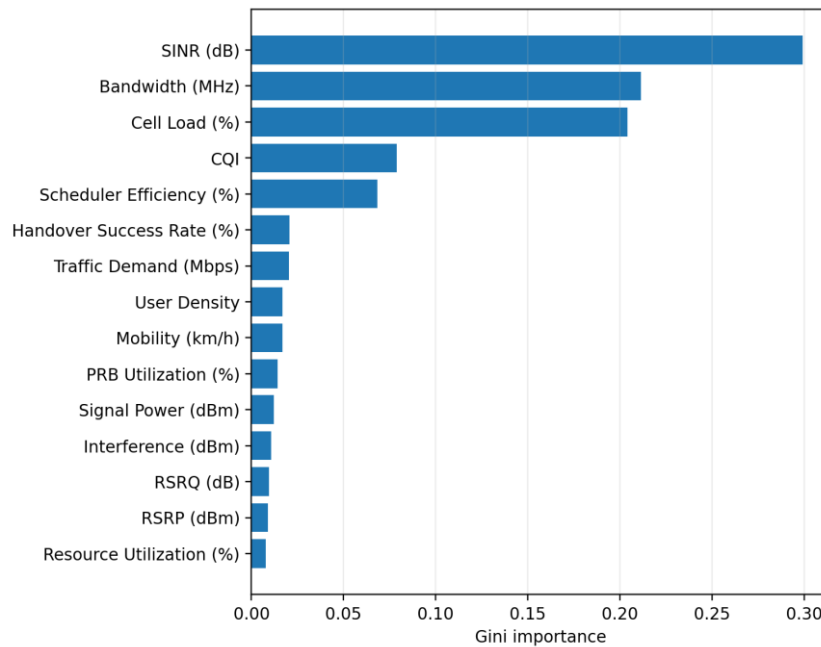


Figure 5. Random Forest Feature-Importance Ranking (Gini-Based) for the Selected RF Model.

Figure 5 presents the ranking in Table 6 as a horizontal bar chart.

SINR (0.299) is the most influential RF predictor, followed by bandwidth (0.211) and cell load (0.204). CQI (0.079) and scheduler efficiency (0.068) form the next tier. Together, these variables indicate that the classifier relies on a combination of radio quality, available capacity, congestion, and resource-management efficiency rather than a single input dimension.

The remaining variables contribute smaller but non-zero importance. Because Gini importance can be distributed among correlated predictors, the ranking is interpreted as a global descriptive summary of the selected RF model rather than as a causal ordering of network variables.

Taken together, the feature analysis supports the physical interpretability of the predictive layer, while the cross-validation, paired significance test, sensitivity analysis, and external robustness experiment place stronger bounds on how the high synthetic accuracy should be interpreted.

4.5 AOS Sensitivity Analysis

The AOS weights and 60/80 boundaries are analyst-defined design parameters, so their sensitivity was tested explicitly. Relative to the baseline labels, alternative weight schemes produced 90.50-95.23% class-assignment agreement (Cohen's kappa = 0.780-0.900). Narrow threshold changes to 58/78 and 62/82 retained 94.37% and 94.89% agreement, respectively (kappa approximately 0.88), whereas wider shifts to 55/75 and 65/85 reduced agreement to 86.04% and 87.78%. The result shows that the three-state construction is reasonably stable to moderate perturbations but that class prevalence, especially the small Optimal class, is sensitive to larger boundary changes. The baseline weights and thresholds are therefore retained as a transparent design choice rather than presented as uniquely optimal.

Table 7. Sensitivity of AOS Class Assignments to Alternative Weight and Threshold Specifications.

Perturbation	Critical (n)	Moderate (n)	Optimal (n)	Agreement (%)	Cohen's kappa
Equal weights	6,427	3,301	272	93.23	0.856
Throughput-latency emphasis	7,415	2,331	254	91.75	0.810
Reliability emphasis	6,390	3,181	429	95.23	0.900
Energy emphasis	7,351	2,542	107	90.50	0.780
Thresholds 55/75	5,783	3,374	843	86.04	0.727
Thresholds 65/85	7,678	2,202	120	87.78	0.709

4.6 External Real-World Robustness Validation

The external experiment used 8,736 real measurements from four SUTD private-5G walk-test scenarios and the dataset's native lab_anom target (FCCLab, 2023). Eight measured radio/resource KPIs were used, with median imputation learned from training data only. Under ordinary stratified five-fold validation across pooled

measurements, RF achieved $88.03 \pm 1.26\%$ accuracy and XGBoost $86.14 \pm 0.93\%$. Because adjacent walk-test observations can share scenario-specific characteristics, a stricter leave-one-scenario-out (LOSO) protocol was also applied. Mean LOSO accuracy fell to $64.29 \pm 16.01\%$ for RF and $66.47 \pm 18.07\%$ for XGBoost, with large between-scenario variation. This drop demonstrates that cross-scenario transfer is substantially more difficult than within-distribution classification and directly supports the reviewer's concern that synthetic accuracy can overstate real-world generalizability.

Table 8. External SUTD Real-World Robustness Results (Mean $\pm$ SD).

Validation	Model	Accuracy (%)	Balanced Acc. (%)	Macro F1 (%)	ROC-AUC (%)
Stratified 5-fold	Random Forest	88.03 ± 1.26	87.66 ± 1.28	87.86 ± 1.28	95.35 ± 0.87
Stratified 5-fold	XGBoost	86.14 ± 0.93	85.74 ± 0.95	85.93 ± 0.94	93.81 ± 0.78
Leave-one-scenario-out	Random Forest	64.29 ± 16.01	68.21 ± 10.90	61.78 ± 18.42	84.89 ± 6.63
Leave-one-scenario-out	XGBoost	66.47 ± 18.07	69.69 ± 12.22	63.74 ± 20.52	88.08 ± 3.98

The external task is intentionally not presented as a direct validation of the synthetic Critical/Moderate/Optimal AOS classifier because the public dataset has a different feature schema and a native binary anomaly label. Instead, it tests whether the same ensemble-learning families retain useful predictive capability on independently collected 5G measurements. The much lower LOSO performance relative to the synthetic experiment establishes a clearer boundary on the study's generalizability and motivates multi-site and operator-level validation as future work.

4.7 Comparison with Prior Studies

Table 9 situates the present study against the primary comparative literature already discussed in Section 2, while Sections 4.5–4.6 provide the new sensitivity and external-validation evidence requested during review.

Table 9. Comparison of the Present Study with Primary Related Work

Study	Method	Application	Key Findings	Limitation	Difference from Present Study
Hussien et al. (2025)	8 models incl. AdaBoost, XGBoost, deep sequence models	Cellular traffic prediction (real CDR data)	Gradient-boosted trees balanced accuracy and efficiency vs. deep models	Continuous traffic-volume target, not an operating-state classification	Present study targets a 3-class operating-state label from a broad radio+resource feature set, not raw traffic volume
Palaios et al. (2023)	ML-based QoS prediction	Vehicular 5G QoS estimation	Highlighted generalization difficulty across radio conditions; flagged leakage risk when targets are derived from performance indicators	Single-domain (vehicular) focus; leakage risk noted but not resolved with an explicit removal procedure	Present study implements and reports an explicit leakage-removal procedure (Section 3.3) directly motivated by this concern
Alekseeva et al. (2021)	Multiple ML techniques incl. gradient boosting, SVR	Traffic prediction on a real wireless network	Ensemble/boosting methods outperformed simpler regression baselines	Single-algorithm-family comparison per technique class; no bagging-vs-boosting ensemble comparison	Present study runs a direct, identical-conditions RF (bagging) vs. XGBoost (boosting) comparison
Grabowski & Xu (2024)	Random Forest, XGBoost	Network intrusion classification	Both ensembles performed strongly; RF's native Gini-importance easiest to compute/communicate	Security/intrusion domain, not network operating-condition classification	Present study applies the same RF/XGBoost pairing to 5G operating-state classification, a distinct application domain and label structure

The comparison indicates that the individual elements used here have clear precedent: Random Forest and XGBoost are established classifiers, categorical network-condition targets have been studied previously, and leakage control is a known requirement for derived targets. The contribution of the present study is therefore the integrated application and validation of these elements for a three-state 5G operating-condition task, strengthened by sensitivity analysis and an independent real-measurement robustness assessment. No claim is made that the combination constitutes a transformative new learning algorithm.

4.8 Limitations of the Study

Several limitations continue to bound the interpretation and generalizability of the findings even after the additional validation performed in response to review.

- Controlled synthetic primary task: the main Critical/Moderate/Optimal experiment is generated from analytically specified relationships. Its 95-96% accuracy therefore demonstrates feasibility under known conditions and should not be treated as expected field performance.

- External-validation scope: the added SUTD experiment uses genuine private-5G measurements, but it is not commercial operator telemetry and does not reproduce the original 15-feature AOS three-class schema. It validates algorithm-level robustness, not exact deployment performance of the primary model.
- Class imbalance: the synthetic Critical, Moderate, and Optimal classes are unevenly represented, particularly the small Optimal class. Macro F1, balanced accuracy, stratification, and RF class weighting mitigate but do not remove this limitation.
- Non-temporal modeling: observations are classified as snapshots. The models do not explicitly learn temporal state transitions or concept drift, and sequential dependence in real walk-test data is addressed only indirectly through scenario-wise validation.
- Restricted algorithm set: the comparison is limited to two tree-based ensemble families and therefore addresses a bagging-versus-boosting question rather than the full space of classifiers or temporal deep-learning models.
- No downstream closed-loop validation: the study evaluates state classification and external anomaly discrimination, not the network-level effect of optimization actions triggered by those predictions.
- Analyst-defined AOS parameters: sensitivity testing shows substantial agreement under moderate changes, but wider threshold shifts alter class prevalence. The baseline weights and 60/80 boundaries remain design assumptions rather than operator-derived service-level thresholds.
- Validation breadth: five-fold cross-validation and McNemar testing strengthen the synthetic comparison, while LOSO testing exposes scenario shift in the external data. Broader multi-site, multi-vendor, and operator-level datasets are still required before deployment-level generalization can be claimed.

4.9 Future Research Directions

Building on the classification framework and findings reported here, several directions merit further investigation.

- Multi-site and operator-level external validation using real 5G telemetry with a feature schema and target definition that can reproduce the Critical/Moderate/Optimal task directly, rather than the binary anomaly proxy used for the present external robustness experiment.
- Extension to online or incremental learning, allowing the classifier to update as network conditions and traffic patterns drift over time, rather than relying on a fixed, statically trained model.
- Temporal and sequential modeling using architectures such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks, and attention-based Transformer models, to exploit network-state history when it is available, rather than classifying each observation independently.
- Explainable-AI techniques beyond Gini-based feature importance, such as SHAP or LIME, to provide instance-level explanations of individual classification decisions in addition to the global feature ranking reported in Section 4.4.
- Integration with O-RAN intelligent-controller architectures, situating the classifier as a candidate xApp or rApp component within a standardized near-real-time or non-real-time control loop.
- Closed-loop optimization integration, in which the classifier's predicted state is consumed by an adaptive resource-allocation mechanism and the resulting throughput, latency, packet loss, spectral efficiency, and energy outcomes are evaluated directly; this remains a distinct extension of the present classification-focused study.

V. Conclusion

This study developed and evaluated a leakage-aware machine-learning approach for classifying 5G network operating state into Critical, Moderate, and Optimal categories using a controlled synthetic dataset of 10,000 observations. The revised methodology fully specifies the synthetic-generation equations, input distributions, Gaussian noise terms, clipping rules, and random seed, and it strengthens evaluation through training-only hyperparameter selection, stratified five-fold cross-validation, AOS sensitivity analysis, and paired statistical testing.

On the held-out synthetic test set, Random Forest achieved 95.65% accuracy and XGBoost 96.45%; five-fold mean accuracies were $94.25 \pm 0.60\%$ and $95.36 \pm 0.44\%$, respectively. McNemar's exact test ($p = 0.0259$) supported a small but statistically significant difference on the controlled test set. XGBoost detected the Critical class with 98.36% recall, while the minority Optimal class remained more difficult at 79.22% recall. AOS sensitivity analysis showed strong agreement under moderate weight changes but greater sensitivity to wider threshold shifts, confirming that the scoring parameters should be regarded as explicit design assumptions rather than uniquely optimal constants.

External evaluation on 8,736 independent SUTD private-5G measurements provided a more realistic boundary on these findings. Although pooled stratified five-fold accuracy remained 88.03% for RF and 86.14% for XGBoost, leave-one-scenario-out accuracy fell to 64.29% and 66.47%, demonstrating substantial scenario-shift effects. The combined evidence therefore supports ensemble learning as a useful and interpretable predictive-intelligence layer while showing that the 95-96% synthetic accuracy is controlled feasibility evidence, not a

deployment-performance estimate. Future work should validate the exact three-state formulation on multi-site operator telemetry and then test the classifier inside a fully implemented closed-loop optimization system.

References

- [1]. Alekseeva, D., Stepanov, N., Veprev, A., Sharapova, A., Lohan, E. S., & Samuylov, A. (2021). Comparison of machine learning techniques applied to traffic prediction of real wireless network. *IEEE Access*, 9, 159495-159514. <https://doi.org/10.1109/ACCESS.2021.3129850>
- [2]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- [3]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). Association for Computing Machinery. <https://doi.org/10.1145/2939672.2939785>
- [4]. Grabowski, A., & Xu, S. (2024). Explainable machine learning for network intrusion detection. *Journal of Cybersecurity Education, Research and Practice*, 2025(1), Article 25. <https://digitalcommons.kennesaw.edu/jcerp/vol2025/iss1/25/>
- [5]. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284. <https://doi.org/10.1109/TKDE.2008.239>
- [6]. Hussien, A. A., Nashaat, H., & Abdel-Kader, R. F. (2025). Machine learning techniques for spatiotemporal traffic prediction in 5G cellular networks. *Discover Applied Sciences*, 7(10), Article 1047. <https://doi.org/10.1007/s42452-025-06746-3>
- [7]. Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Leanpub.
- [8]. Palaos, A., Riihijarvi, J., Mahonen, P., Vaeth, M., & Griesser, R. (2023). Machine learning for QoS prediction in vehicular communication: Challenges and solution approaches. *IEEE Access*, 11, 92459-92477. <https://doi.org/10.1109/ACCESS.2023.3308668>
- [9]. Future Communications and Connectivities Lab (FCCLab). (2023). SUTD 5G dataset [Data set]. Singapore University of Technology and Design. GitHub. https://github.com/FCCLab/sutd_5g_dataset_2023