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.

Interpretable Machine Learning for Predicting Video Quality of Experience in Mobile Networks

Hafiz M. Asif
Sultan Qaboos University
Abdulraqeb Alhammadi
abdulraqeb.alhammadi@utm.my
University of Technology Malaysia
Naser Tarhuni
Sultan Qaboos University
Mohammad M. Bait-Suwailam
Sultan Qaboos University
ibraheem Shayea
Istanbul Technical University

Article

Keywords:

Posted Date: May 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9624596/v1

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

Additional Declarations: No competing interests reported.

Interpretable Machine Learning for Predicting Video Quality of Experience in Mobile Networks

Hafiz M. Asif[1], Abdulraqeb Alhammadi[2,*], Naser Tarhuni[1,*], Mohammad M. Bait-Suwailam[1], and ibraheem Shayea[3]

[1]Department of Electrical and Computer Engineering, College of Engineering, Sultan Qaboos University, Muscat 123, Oman.
[2]Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia.
[3]Department of Electronics and Communication Engineering, Faculty of Electrical and Electronics Engineering, Istanbul Technical University, Istanbul, 34467, Turkey.
[*]abdulraqeb.alhammadi@utm.my;tarhuni@squ.edu.om

ABSTRACT

Quality of Service (QoS) is a crucial aspect of network performance, focusing on technical aspects. It aids in prioritizing traffic types for voice conversations and video streaming and combines objective network measurements with subjective user expectations for improved service quality. Quality of Experience (QoE) is used to assess and optimize user satisfaction with services such as video streaming, voice communication, gaming, and other network-based applications, helping service providers enhance performance and align with user expectations. Mean Opinion Score (MOS) is a key metric for evaluating QoE based on subjective assessments of service quality, such as audio, video, or overall network performance. This article presents a prediction of MOS using four AI predictive models: Decision Tree, Support Vector Machine (SVM), Ensemble, and Neural Network. The research begins by analyzing datasets to uncover key patterns and correlations that affect network performance and user-perceived quality. Leveraging these insights, AI models are employed to predict the MOS under various feature selection scenarios (MRMR and SHAP), illustrating the influence of different network parameter combinations on user experience. To ensure the reliability and effectiveness of the predicted models, their performance is evaluated using error metrics. The results have demonstrated that selecting the most relevant features (as identified by MRMR and SHAP) has streamlined the models without compromising accuracy, particularly for the Neural Network model. The Neural Network maintains outstanding performance, with an RMSE of 0.009 on validation and 0.007 on test data, and a near-perfect R-squared (0.999) on both datasets.

1 Introduction

Quality of Service (QoS) and Quality of Experience (QoE) are two important criteria that are frequently used to evaluate the performance of mobile broadband networks. In order to guarantee that services fulfill certain requirements, QoS focuses on the technical facets of network performance, including latency, jitter, throughput, and packet loss. Especially for applications like voice conversations and video streaming, these indicators aid in prioritizing various traffic types to preserve peak performance. QoS techniques were developed in 4G to effectively support a range of services. These mechanisms have developed with 5G to support more complex use cases, such as huge machine-type communications and ultra-reliable low-latency communications, allowing for more accurate and adaptable service delivery. By fusing objective network measurements with subjective elements like user expectations and context, QoE, on the other hand, focuses on user perceptions of service quality[1,2]. These services require not only superior technical performance but also adaptive approaches that consider varying user environments and device capabilities, making QoE a crucial factor in the development and optimization of modern networks. An AI-driven QoE prediction and sustainability framework integrates with 5G network data analytics, network slicing, and mobile edge computing technologies to gather real-time cross-layer performance data and dynamically adjust network resources as needed[3].

One important statistic for evaluating QoE in communication systems is the Mean Opinion Score (MOS). It was first proposed by the International Telecommunication Union (ITU), which utilized end-user input on network performance as a criterion to assess network providers[4]. User-perceived service quality is measured by the MOS using a scale of 1 (poor) to 5 (excellent). Users' subjective assessments of their experiences with services such as

video, audio and multimedia applications are commonly utilized for determining MOS. By converting technical measurements like latency or packet loss into a more user-centric assessment, this score offers insightful information on how users see service quality[5]. By doing this, MOS assists companies in determining how well their offerings satisfy customer expectations.

In the context of 4G and 5G networks, MOS is crucial for optimizing service delivery in several studies[6,7]. As users increasingly rely on services such as high-definition video streaming, voice communication, and immersive applications like virtual reality, understanding their perception of quality becomes essential. MOS, combined with objective QoS data, enables network providers to fine-tune their systems to ensure consistent and satisfactory user experience. Moreover, modern approaches like machine learning are being utilized to predict MOS in real-time, allowing networks to adjust dynamically to changing conditions and maintain high levels of QoE.

Mobile network operators rely on drive testing as a vital tool to evaluate and enhance network performance, ensuring optimal QoE for users. This process involves outfitting vehicles with specialized equipment to measure key network parameters such as signal strength, coverage, call quality, and data speeds across various locations[8]. Drive testing provides real-world insights into network behavior, helping operators assess MOS by linking technical performance metrics with user-perceived service quality. Through these insights, operators can identify problem areas and implement targeted improvements, delivering more reliable and consistent service to their customers.

This paper offers several key contributions to the field. First, we study the statistical analysis of datasets and investigate its impact on network performance and user-perceived quality, identifying key patterns and correlations. Second, we utilize AI models (Fine Tree, SVM, Boosted Tree, and Neural Network) to predict the MOS under various scenarios of feature selection such as Minimum Redundancy Maximum Relevance (MRMR) and SHapley Additive exPlanations (SHAP), demonstrating how different combinations of network parameters influence MOS prediction. Third, we evaluate the applied AI models using several performance metrics, such as Root Mean Square Error (RMSE), Mean Square Error (MSE), Mean Absolute Error (MAE), and R-squared, to validate their effectiveness and reliability in predicting MOS across diverse conditions. These contributions provide a comprehensive approach to prediction MOS that is related to QoE in mobile networks using network performance parameters.

The reminder of this article is organized as follows: Section 2 discusses related work on MOS prediction using various methods. Section 3 presents data collection and features, and description. Section 4 outlines the research methodology including four main parts: data analysis, features engineering, performance metrics. The AI model predictions and test results are presented and discussed in Section 5. Finally, Section 6 concludes the article.

2 Related Work

Prediction of MOS provides valuable insights into user experience, allowing for the evaluation of how network performance influences overall satisfaction across different conditions. In[9], the authors evaluated quality of LTE for video conferencing using MOS to figure out user experience due to video conferencing requiring high bandwidth and fast internet. The study in[7] presented a novel QoE prediction model for video streaming services over 4G networks utilizing network performance parameters. The model applied several ML algorithms such as random forest, gradient tree boosting, adaptive boosting, extra tree and support vector regression to predict the service's MOS, utilizing real mobile operator drive test data that includes both application layer and physical layer metrics.

In[10,11], the authors presented a Bayesian model designed to estimate the MOS for video streaming services over specific cellular networks. MOS is widely recognized as a key measure of quality of experience. The proposed model incorporates multiple input factors, including round-trip time, stalling load, and bit rates. Its performance was assessed and analyzed using different test data sizes and various evaluation metrics. The authors in[12] tackled the challenges in modeling and evaluating QoE for video streaming services. It introduces a novel full-reference QoE model to predict user MOS for streamed videos, with a focus on YouTube. Instead of relying solely on packet analysis, the study uses screen captures of streamed videos for an immersive user experience assessment. The approach leverages FFmpeg and VMAF, open-source tools for video processing and quality evaluation, respectively. In[13], the study introduced user-centric QoS to QoE mapping model using backpropagation neural network. The model incorporates network parameters and MOS. It applies a combination of algorithms for network optimization, with enhancements made to the optimization algorithm based on the internal characteristics of the QoS/QoE mapping process.

The study in[14] tackles the issue by proposing an innovative hybrid XGBStackQoE analytical model that utilizes a two-tier layering approach. The first level integrates several Machine Learning (ML) models through a layer-one

Hybrid XGBStackQoE model. Each ML model at this level is trained on the complete training dataset. In the second level, the Hybrid XGBStackQoE model is trained using the outputs (meta-features) generated by the first-level ML models.

Authors in[15] analyzed real-time testbed data from the 5G mobile network, focusing on calculating the inter-MOS and QoS metrics for both 5G Voice over LTE (VoLTE) and Voice over New Radio (VoNR) services within the IMS framework. It evaluates QoS metrics across various communication stages between the user equipment for both 5G stand-alone and 5G non-stand-alone networks. Similar in work[16], the authors evaluated voice quality in VoLTE calls on Telecom Italia Mobile's radio network using MOS to identify how radio propagation affects speech quality. Tests were conducted in a controlled environment, simulating electromagnetic conditions within a reverberation chamber, by varying signal strength and interference. In[17], the authors contributed to ongoing research by developing a system for automatically predicting the MOS provided by human listeners. This system was designed specifically for the recent VoiceMOS Challenge. Building on the success of fusion systems in similar competitions, our approach employs an ensemble that combines the outputs of seven distinct models through interpolation. The work in[18] assessed the quality of voice calls in mobile cellular networks through drive tests conducted with specialized measurement terminals. Three scenarios were examined: mobile-to-mobile calls on the GSM network (CS domain), VoLTE calls on a 4G network (PS domain), and WhatsApp calls on the same 4G network (PS domain). Voice quality was compared across scenarios using MOS. The uniqueness of this research lies in its real-world measurements conducted in Bucharest, Romania, providing valuable insights to guide local network enhancements and policymaking.

The research in[19] presented a novel method to improve VoIP call QoS during handovers from LTE to Wi-Fi. The proposed solution incorporates two MOS prediction models, namely multilayer perception and recurrent neural networks, into a software-defined network, controller, optimized using the grey wolf optimizer. This approach enables the software-defined networking to accurately predict and route mobile devices and traffic to the most optimal nearby Wi-Fi access points.

3 Data collection and Description

3.1 Data collection

This research aims to develop AI-driven models that predict the QoE through MOS. The dataset used in this study was obtained from previously published measurement campaigns conducted using commercial mobile smartphones and drive-test experiments over Mobile Network Operators (MNO)[20]. The collected data included anonymized QoE indicators and MOS assessments from prior studies. The present work only utilized the previously collected and anonymized dataset for ML modelling and analysis and did not involve new experiments with human participants.

The data collection was designed to capture user experiences across different times, including weekdays and weekends, as well as during peak and off-peak hours. This temporal analysis helps in understanding the variations in Mobile Broadband (MBB) performance based on user activity patterns.

For the experimental testing of video-streaming MBB services, network performance data was collected using the Speedvideo application during drive tests in cities, inter-city roads, and highways. Each testing cycle involved YouTube video streaming at low resolution (720p) and high resolution (1080p), with the number of cycles configured by the user. The car speed was maintained at 60 km/h for city and inter-city roads and 80 km/h for highways. Mobile phones in the car communicated with base stations along the route to gather. The collected data from the measurement application was transmitted to a centralized database for storage and further processing. Fig. 1 illustrates the procedures for data collection, transmission, and analysis.

The data collected consists of 53 features, specifically measured Key Performance Indicators (KPIs) for video streaming services using the Speedvideo application. The KPIs included:

- Initial buffering success ratio
- Cellular signal strength
- Ping round-trip time (RTT) delay
- Video total download rate
- Initial buffering latency (user perceived delay)
- Total rebuffering latency

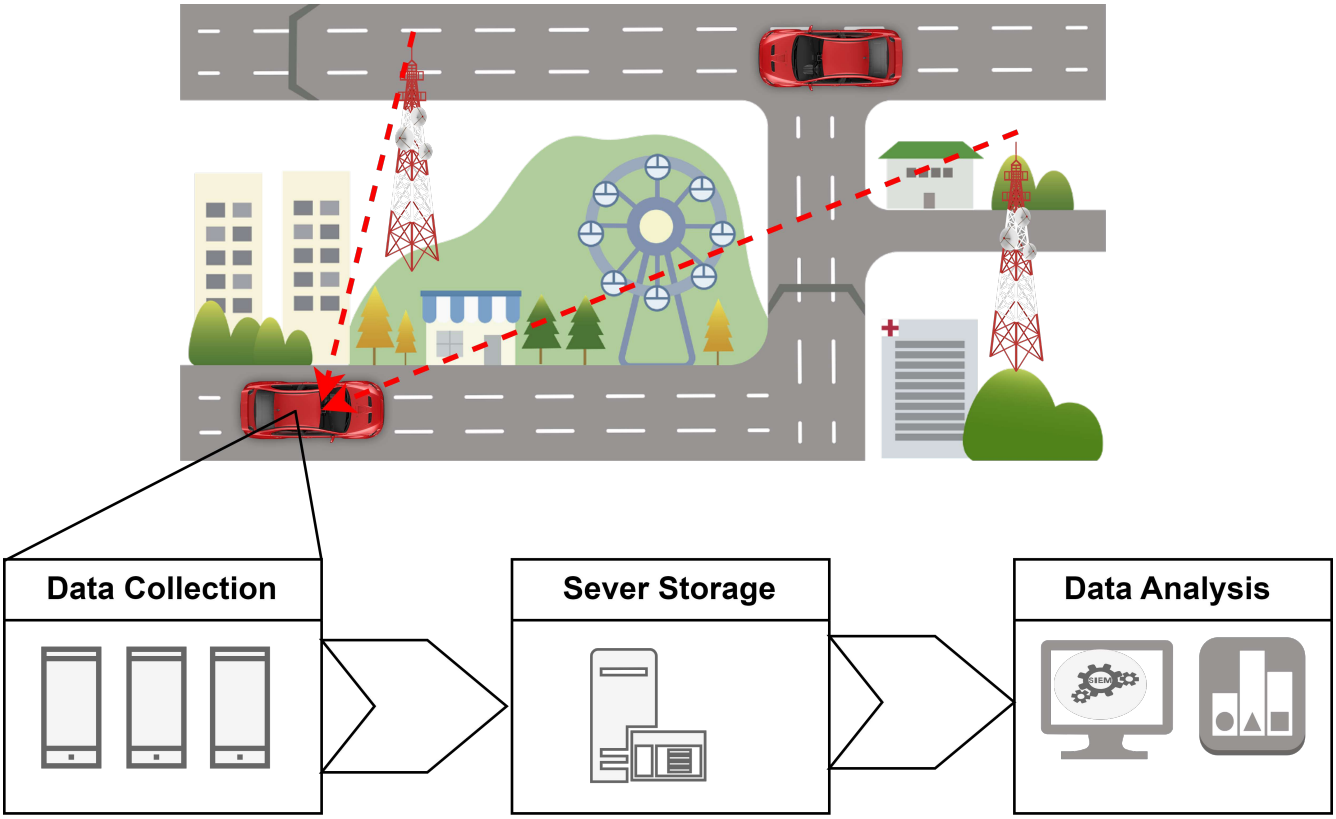


Figure 1. Procedures for data collection

- Mobile video MOS score

We added more eight features to that considered in the previous study which are: Loading Score, Bit Rate, Stalling Ratio, Quality Score, Stalling Score, Network Type, City and Carrier. These features are most important for MOS.

3.2 Data Analysis

Table 1 provides descriptive indicators of 10 input parameters. The statistical analysis includes 12 measures such as Mean, Standard Deviation, Confidence Level (95.0%), Variance, Skewness, Kurtosis, Mode, Minimum. Median, Maximum and Range. The table presents descriptive statistics for various network performance metrics. Key insights include:

- Cellular Signal Strength has a mean of approximately -83.3 dBm, with a narrow 95% confidence interval, indicating consistent signal levels. It also shows moderate skewness and kurtosis, implying a fairly normal distribution.
- Mobile vMOS and Quality Score metrics exhibit low variance and small ranges, suggesting stable user experiences. Their skewness and kurtosis indicate distributions close to normal.
- Initial Buffering Latency and Total Rebuffering Latency have high variance and wide ranges, highlighting significant variability in user experiences regarding video streaming delays.
- Video end-to-end (E2E) RTT demonstrates high variability with a large range, indicating inconsistent latency in video transmission.
- Loading Score and Stalling Score have low variance and a small range, indicating consistent loading and stalling performance across observations. Both metrics show slightly negative skewness, meaning that most values are clustered around higher performance levels.
- Stalling Ratio has a relatively higher variance and range compared to other metrics, with a positively skewed distribution and moderate kurtosis. This indicates that while stalling is rare for most users, some experience

significantly higher stalling events.

- Video Total DL Rate (kbps) shows moderate variance and a small range, suggesting consistent download rates across users. The negative skewness indicates that most users achieve relatively higher download rates, with few experiencing lower rates.

Overall, metrics like buffering latency and rebuffering latency exhibit significant spread and variability, while metrics like mobile vMOS and Quality Score are more consistent.

4 Research Methodology

This section outlines data analysis (statistical measures), features engineering (including feature selection), AI models, and methodological choices evaluation. Therefore, the research methodology includes three main parts: First part, the data analysis was conducted through extensive descriptive statistics. The second part discusses the features of engineering using three different coefficients, namely, Pearson correlation coefficient (PCC), Minimum Redundancy Maximum Relevance (MRMR) and SHapley Additive exPlanations (SHAP). Finally, in the third part, the regression metrics were presented.

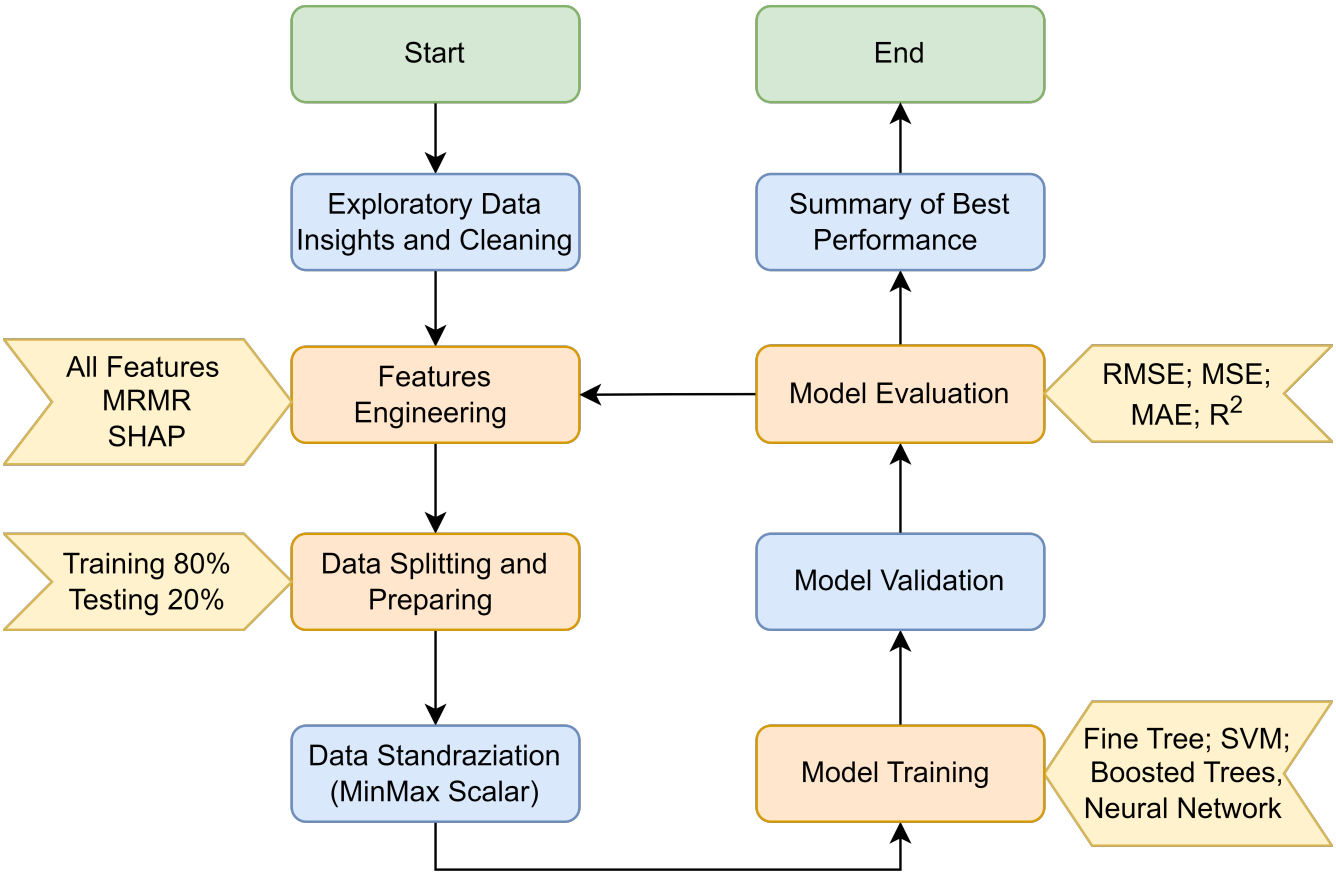


Figure 2. General workflow of research methodology

Fig. 2 shows a flowchart illustrating the general workflow of the methodology in detailed steps. The methodology depicted follows a structured machine learning workflow:

1. Exploratory Data Insights and Cleaning: Initial data exploration to understand distributions, identify missing values, and correct inconsistencies.

2. Features Engineering: Feature selection is performed using techniques like PCC, MRMR and SHAP to identify the most impactful features.

Table 1. Descriptive Statistics of Input Parameters

Parameter	Mean	Std Dev	Lower 95% CI	Upper 95% CI	Variance	Skewness	Kurtosis	Mode	Min	Median	Max	Range
Cellular Signal Strength (dBm)	-83.32	15.76	-84.14	-82.50	248.36	1.73	3.97	-91	-120	-87	-31	89
Mobile vMOS	2.67	1.36	2.60	2.74	1.85	-1.00	-0.16	1	-1	3.38	3.98	4.98
Quality Score	3.84	0.99	3.79	3.89	0.99	-4.36	18.53	4.28	-1	3.78	4.28	5.28
Loading Score	2.87	1.59	2.79	2.96	2.53	-1.06	-0.42	-1	-1	3.77	4.27	5.27
Stalling Score	3.51	2.26	3.40	3.63	5.09	-0.95	-0.97	5	-1	5	5	6
Stalling Ratio	2.11	5.18	1.84	2.38	26.85	2.23	3.13	0	0	0	18.62	18.62
Initial Buffering Latency (ms)	5805.13	9693.49	5300.34	6309.91	93963700	2.48	5.97	0	0	1436	60961	60961
Total Rebuffering Latency (ms)	5230.03	9938.81	4712.47	5747.59	98780000	1.65	1.02	0	0	0	35295	35295
Video Total DL Rate (kbps)	1593.44	606.44	1561.86	1625.02	367769.73	-0.49	-0.13	1417	3	1483	2701	2698
Video E2E RTT (ms)	425.13	866.69	380.00	470.26	751148.80	2.72	7.23	55	-1	57	4967	4968

3. Data Splitting and Preparing: The dataset is split into training (80%) and testing (20%) subsets to evaluate model performance on unseen data.
4. Data Standardization: MinMax scaling is applied to normalize feature values, ensuring all features contribute equally to model training.
5. Model Training: Several algorithms, including Regression Trees (Fine Tree), Support Vector Machine (SVM), Ensemble of Trees (Boosted Trees), and Neural Network, are trained to find the best-performing model.
6. Model Validation: Models are validated to prevent overfitting and ensure reliable performance.
7. Model Evaluation: Model performance is evaluated using metrics such as RMSE, MSE, MAE, and R^2.
8. Summary of Best Performance: The workflow concludes with a summary of the best-performing model based on evaluation metrics.

4.1 Feature Selections

4.1.1 PCC

The PCC measures the linear relationship between two variables, making it useful in feature selection to identify features that are strongly correlated with the target variable, while also helping to eliminate features that are highly correlated with each other to reduce redundancy. The mathematical expression for the PCC between two variables x and y is given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where x_i and y_i are the individual data points for the two variables, $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively, and n is the number of data points.

Alternatively, this can also be expressed in terms of covariance and standard deviations:

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \tag{2}$$

where $\text{cov}(x, y)$ is the covariance between x and y, and σ_x and σ_y are the standard deviations.

4.1.2 MRMR

The MRMR algorithm is a feature selection method that identifies an optimal subset of features that balance two objectives: maximizing relevance to the response variable and minimizing redundancy among features. It uses mutual information to quantify relevance and redundancy, as defined below:

Relevance (V_S):- The average mutual information between features in the set and the response variable :

$$V_S = \frac{1}{|S|} \sum_{x_i \in S} I(x, y) \tag{3}$$

Redundancy (W_S): The average pairwise mutual information between features within the set :

$$W_S = \frac{1}{|S|^2} \sum_{x_i, x_j \in S} I(x, z) \tag{4}$$

The algorithm aims to maximize V_S while minimizing W_S, which is computationally challenging since evaluating all subsets S from the entire feature set Ω involves $2^{|\Omega|}$ combinations. To address this, MRMR uses a forward selection heuristic with the Mutual Information Quotient (MIQ):

$$\text{MIQ}_x = \frac{V_x}{W_x} \tag{5}$$

where $V_x = I(x, y)$ is the relevance of a feature x, and

$$W_x = \frac{1}{|S|} \sum_{z \in S} I(x, z) \tag{6}$$

is the redundancy of x with respect to features in the current subset S.

Algorithm 1 MRMR Feature Selection

Require: Feature set Ω, target variable y
Ensure: Selected feature subset S
1: Initialize $S \leftarrow \emptyset$
2: Compute relevance for all features:

$$V_x = I(x, y), \quad \forall x \in \Omega$$

3: Select feature with maximum relevance:

$$x^* = \arg\max_{x \in \Omega} V_x$$

4: $S \leftarrow S \cup \{x^*\}$
5: **while** there exist features in $S^c = \Omega \setminus S$ with nonzero relevance **do**
6: Compute redundancy for each feature:

$$W_x = \frac{1}{|S|} \sum_{z \in S} I(x, z)$$

7: **if** exists feature with zero redundancy **then**
8: Select $x^* = \arg\max_{x \in S^c} V_x$
9: **else**
10: Compute MIQ:

$$\text{MIQ}_x = \frac{I(x, y)}{\frac{1}{|S|} \sum_{z \in S} I(x, z)}$$

11: Select:

$$x^* = \arg\max_{x \in S^c} \text{MIQ}_x$$

12: **end if**
13: $S \leftarrow S \cup \{x^*\}$
14: **end while**
15: Add remaining features with zero relevance to S in random order
16: **return** S

Computational Efficiency:

The MRMR algorithm reduces the computational cost from $\mathcal{O}(2^{|\Omega|})$ to $\mathcal{O}(|\Omega| \cdot |S|)$ for the forward addition process. However, when ranking all features, the computation cost becomes $\mathcal{O}(|\Omega|^2)$.

Outputs: - Feature Indices: A ranked list of features in decreasing order of importance. - Scores: Importance values for each feature, where a larger score indicates higher importance. A significant drop in scores between

consecutive features reflects confidence in feature selection. The MRMR algorithm is particularly effective in reducing dimensionality while retaining features that are most informative and least redundant, making it suitable for tasks such as classification and regression.

4.1.3 SHAP

The SHAP framework is a method for interpreting machine learning models by attributing importance values, known as Shapley values, to features. It is based on concepts from cooperative game theory, particularly the Shapley value, which assigns contributions to players in a game. In SHAP, features are treated as "players," and the goal is to fairly allocate the contribution of each feature to the model's prediction. SHAP assumes an additive explanation model:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i \tag{7}$$

where M denotes the number of simplified input features. $z' \in \{0,1\}^M$ represents a simplified binary input vector representing the presence ($z'_i = 1$) or absence ($z'_i = 0$) of a feature. $\phi_i \in \mathbb{R}$ represents the coefficient (Shapley value) for feature i, representing its contribution to the prediction.

The coefficients ϕ_i satisfy three properties:

1. **Local Accuracy:** The explanation model's output matches the original model's prediction for the given input.
2. **Missingness:** Features that are absent ($z'_i = 0$) contribute nothing ($\phi_i = 0$).
3. **Consistency:** If a feature's contribution increases in any model, its importance (Shapley value) does not decrease.

Shapley Value Formula:

Under these properties, the Shapley value for a feature i is given by:

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!\,(M - |z'| - 1)!}{M!} \left[f_x(z') - f_x(z' \setminus i) \right] \tag{8}$$

where $|z'|$ represents the number of non-zero entries in z' (active features), and $z' \setminus i$ represents the subset of z' excluding feature i. $f_x(z')$ is the model's prediction when only the features represented in z' are included. This formula computes the average marginal contribution of feature i across all possible subsets of features.

Shapley values quantify the importance of each feature by evaluating its contribution to the model's prediction. It considers all possible combinations of feature inclusion to fairly distribute the "credit" for the prediction among the features.

4.2 Performance Metrics

In this study, four different metrics were measured to evaluate the prediction performance of each AI model in estimating the vMOS: RMSE, MSE, MAE, and R^2. These metrics provide a comprehensive assessment of the models' accuracy, reliability, and ability to capture the variance in the predicted values.

1. Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{9}$$

where y_i represents actual values, $\hat{y}_i$ predicted values, and n the number of observations.

2. Mean Squared Error (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{10}$$

3. Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{11}$$

4. Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{12}$$

where $\bar{y}$ is the mean of actual values.

These metrics collectively quantify the error magnitude, variability, and the goodness of fit between the predicted and actual values, providing a robust evaluation of the AI models' predictive capabilities.

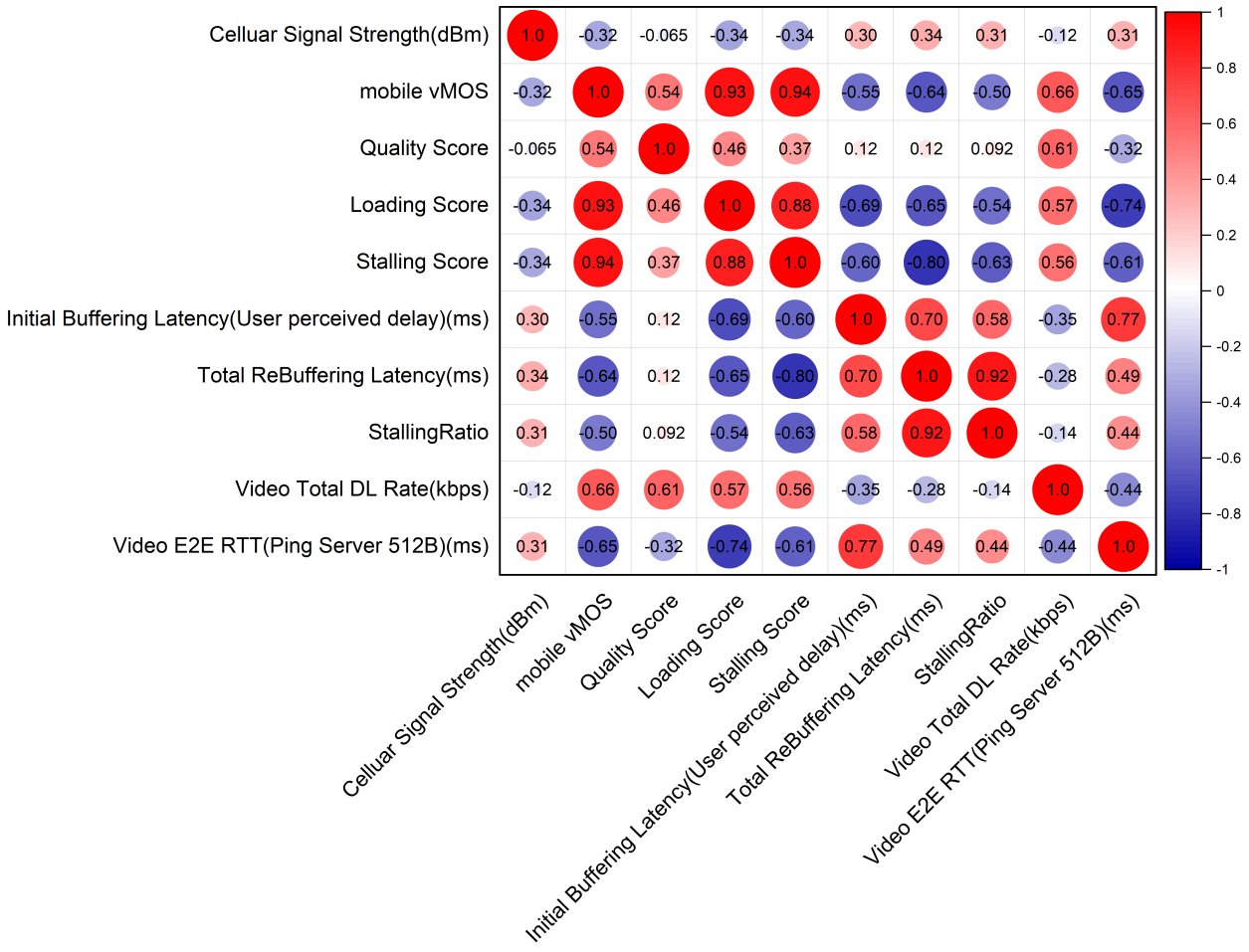


Figure 3. PCC of features

5 Results and Discussion

Figure 3 illustrates correlation relationships between various QoE metrics (features) for mobile video streaming. Each cell in the matrix represents the PCC between two variables, with color and size indicating the strength and direction of the correlation. Red circles represent positive correlations, while blue circles represent negative

correlations. Larger circles indicate stronger correlations (either positive or negative), while smaller circles represent weaker correlations. The scale on the right shows correlation values from -1 (strong negative correlation) to +1 (strong positive correlation).

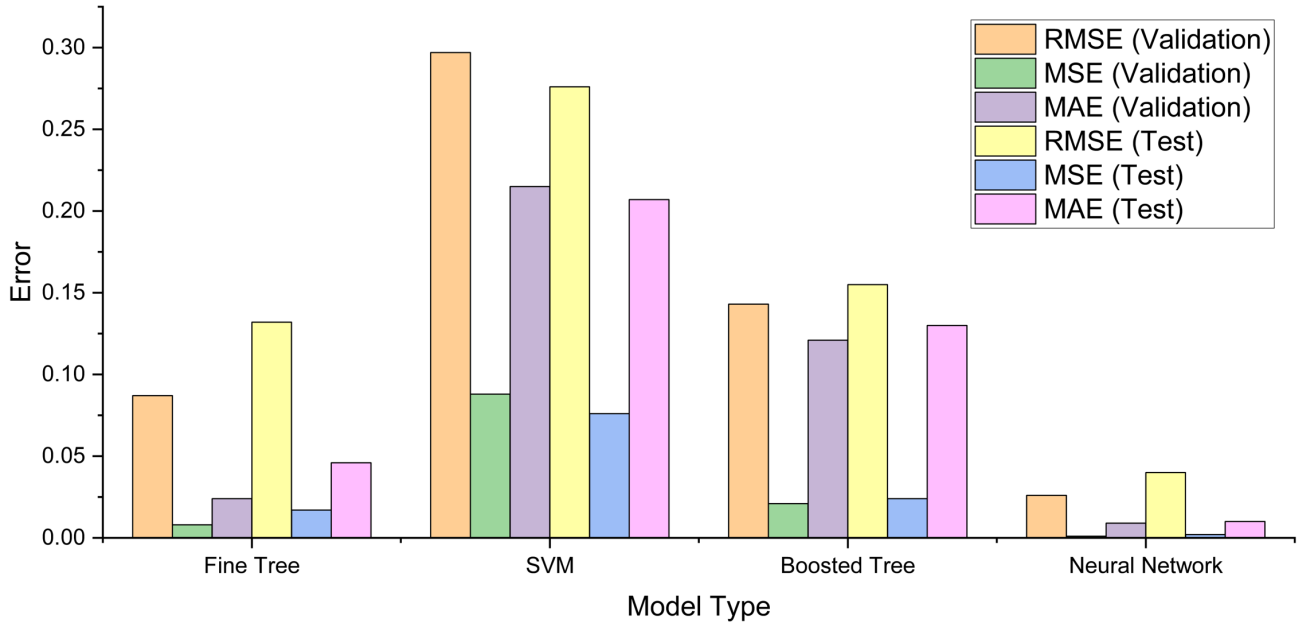


Figure 4. Evaluation validation and test datasets across several models on PCC selection

The correlation plot shows that user-perceived video quality (vMOS) is closely tied to loading and stalling scores, with high positive correlations indicating that improvements in these areas enhance user experience. Cellular signal strength has a negative correlation with these same QoE metrics, suggesting that weaker signals lead to longer loading times and increased stalling, reducing perceived quality. Metrics such as buffering latency and stalling ratio are strongly correlated, meaning initial delays typically lead to more frequent buffering. Higher download rates improve loading and stalling, while lower E2E RTT are generally beneficial for overall streaming quality. From PCC, we selected the features with the strongest linear correlation to the target variable (vMOS) which are: Quality Score, Loading Score, Stalling Score, Bit Rate, and Video Total DL Rate.

Figure 4 evaluates the performance of four models on validation and test datasets various error metrics. The Fine Tree model shows low validation errors but higher test errors, indicating overfitting. SVM performs poorly, with significantly higher errors on both datasets, reflecting poor generalization. The Boosted Tree model has moderate validation and test errors, showing better generalization than SVM but less precision than the Neural Network. The Neural Network outperforms all other models, exhibiting the lowest errors across both datasets, demonstrating strong accuracy and generalization. Table 2 shows the Neural Network model achieves the best performance across all metrics during validation and test training. The test training shows the lowest RMSE (0.040), MSE (0.002), and MAE (0.010), along with the highest R-squared (0.999).

Table 2: Summarizes the performance of four different models based on PCC selection

	Validation				**Test**			
Model Type	**RMSE**	**MSE**	**MAE**	**RSquared**	**RMSE**	**MSE**	**MAE**	**RSquared**
Fine Tree	0.087	0.008	0.024	0.995	0.132	0.017	0.046	0.987
SVM	0.297	0.088	0.215	0.947	0.276	0.076	0.207	0.945
Boosted Tree	0.143	0.021	0.121	0.988	0.155	0.024	0.130	0.983
Neural Network	**0.026**	0.001	0.009	1.000	**0.040**	0.002	0.010	0.999

5.1 MRMR Feature selection

MRMR method is employed to select features that are highly relevant to the target variable while minimizing redundancy among them. This approach ensures the model focuses on the most impactful features, improving prediction accuracy and reducing computational complexity. Figure 5 shows the importance scores of the selected features. Loading scores emerge as the most influential, followed by (scores higher than 1) quality score and stalling ratio, highlighting their strong association with the target variable. Other significant features include video total DL rate, initial buffering latency, bit rate, network type, stalling score and video E2E RTT which contribute to understanding network performance. Less impactful features, such as total rebuffering latency, initial buffering success, city and carrier and cellular signal strength have lower importance scores (scores lower than 1), indicating minimal contribution to the predictive model. This process ensures the selection of diverse and highly relevant features, enhancing the efficiency and accuracy of the model.

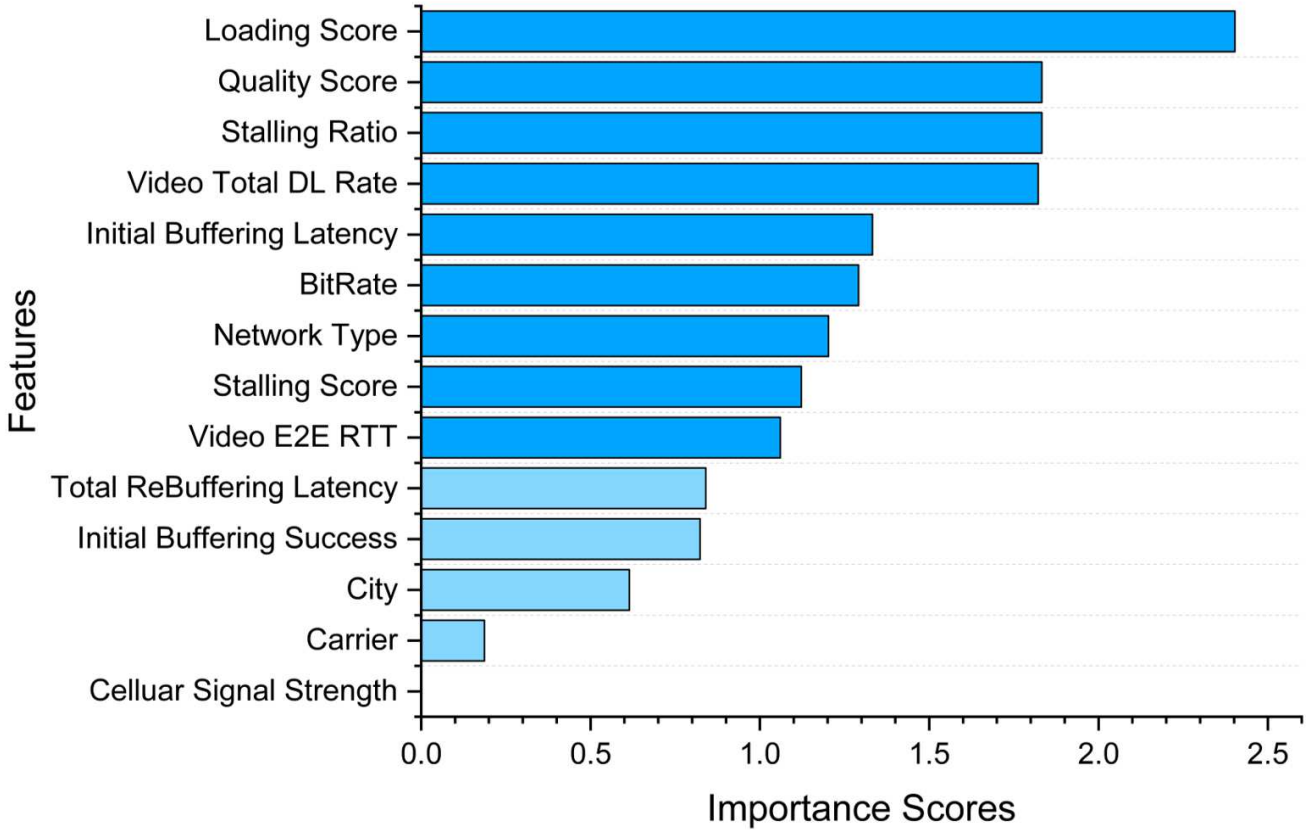


Figure 5. Importance score of MRMR feature selections

Four performance metrics are utilized to evaluate vMOS predictions: RMSE, MAE, MSE and R-squared were computed to enhance the accuracy and efficiency of vMOS predictions as shown in Figure 6. These metrics were employed across parameter configurations and AI predictive models: Decision Tree, SVM, Ensemble, and Neural Network to optimize the prediction models. Table 3 summarizes the performance of four different models.

Table 3: Summarizes the performance of four different models based on MRMR selection

	Validation				**Test**			
Model Type	**RMSE**	**MSE**	**MAE**	**RSquared**	**RMSE**	**MSE**	**MAE**	**RSquared**
Fine Tree	0.093	0.009	0.028	0.995	0.064	0.004	0.021	0.997
SVM	0.395	0.156	0.230	0.904	0.445	0.198	0.249	0.868
Boosted Tree	0.142	0.020	0.122	0.988	0.135	0.018	0.120	0.988
Neural Network	**0.065**	0.004	0.020	0.997	**0.040**	0.002	0.013	0.999

Neural Network: This model achieved the best performance across all metrics during validation and test training. The test training shows the lowest RMSE (0.040), MSE (0.002), and MAE (0.013), along with the highest R-squared (0.999). These values indicate that the neural network model has minimal prediction errors and nearly perfect correlation with the actual values, making it the most accurate among the evaluated models.

Fine Tree: The decision tree model also performed well during test training, with an RMSE of 0.064, MSE of 0.004, MAE of 0.021 and an R-squared of 0.997. Although not as accurate as the neural network, it still achieved a high degree of correlation with actual values and relatively low errors. Its simplicity and interpretability might make it a suitable choice for cases where slightly lower accuracy is acceptable.

Boosted Tree: The Ensemble model had an RMSE of 0.135, MSE of 0.018, MAE of 0.120, and R-squared of 0.988. While it improved on SVM, it did not reach the accuracy of the decision tree or neural network. ensembles generally offer good stability and robustness, though this model's performance indicates it might need further tuning to compete with the neural network's accuracy.

SVM: This model showed the highest error rates among all models, with an RMSE of 0.445, MSE of 0.198, MAE of 0.249, and R-squared of 0.868. This suggests that SVM is not well-suited for this particular task or dataset, possibly due to its sensitivity to parameter selection and feature scaling.

The performance of the Neural Network model is similar for both MRMR and PCC feature selection methods, despite PCC using five features and MRMR using nine features. could be attributed to the fact that PCC selects features with the strongest linear correlation to the target variable, ensuring that the most relevant information is retained. On the other hand, MRMR considers both relevance and redundancy, selecting features that contribute complementary information. In this case, the additional features chosen by MRMR may not significantly improve the model's performance, as the core predictive information is already captured by the fewer, highly correlated features selected by PCC. This highlights the importance of feature relevance over sheer quantity in determining model performance.

In summary, the neural network is the most accurate model for vMOS prediction in this study, followed by the fine tree and boosted tree models. The SVM model's comparatively poor performance suggests it may not be the best fit for this problem without significant tuning.

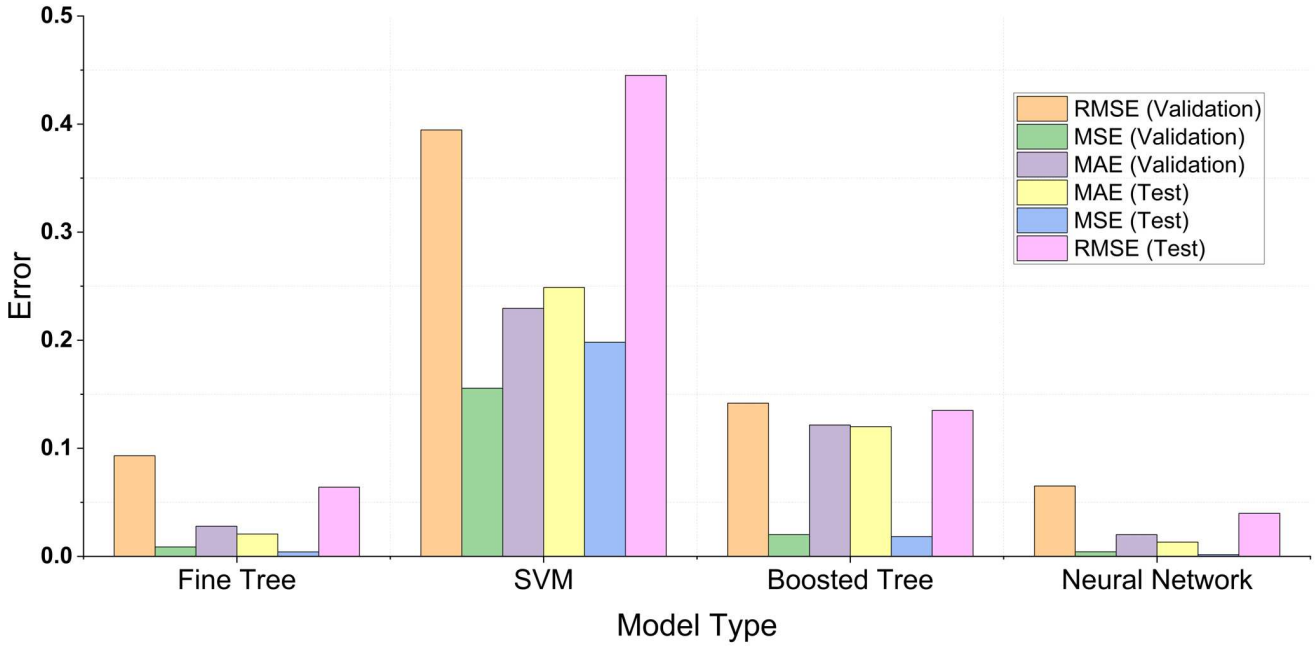
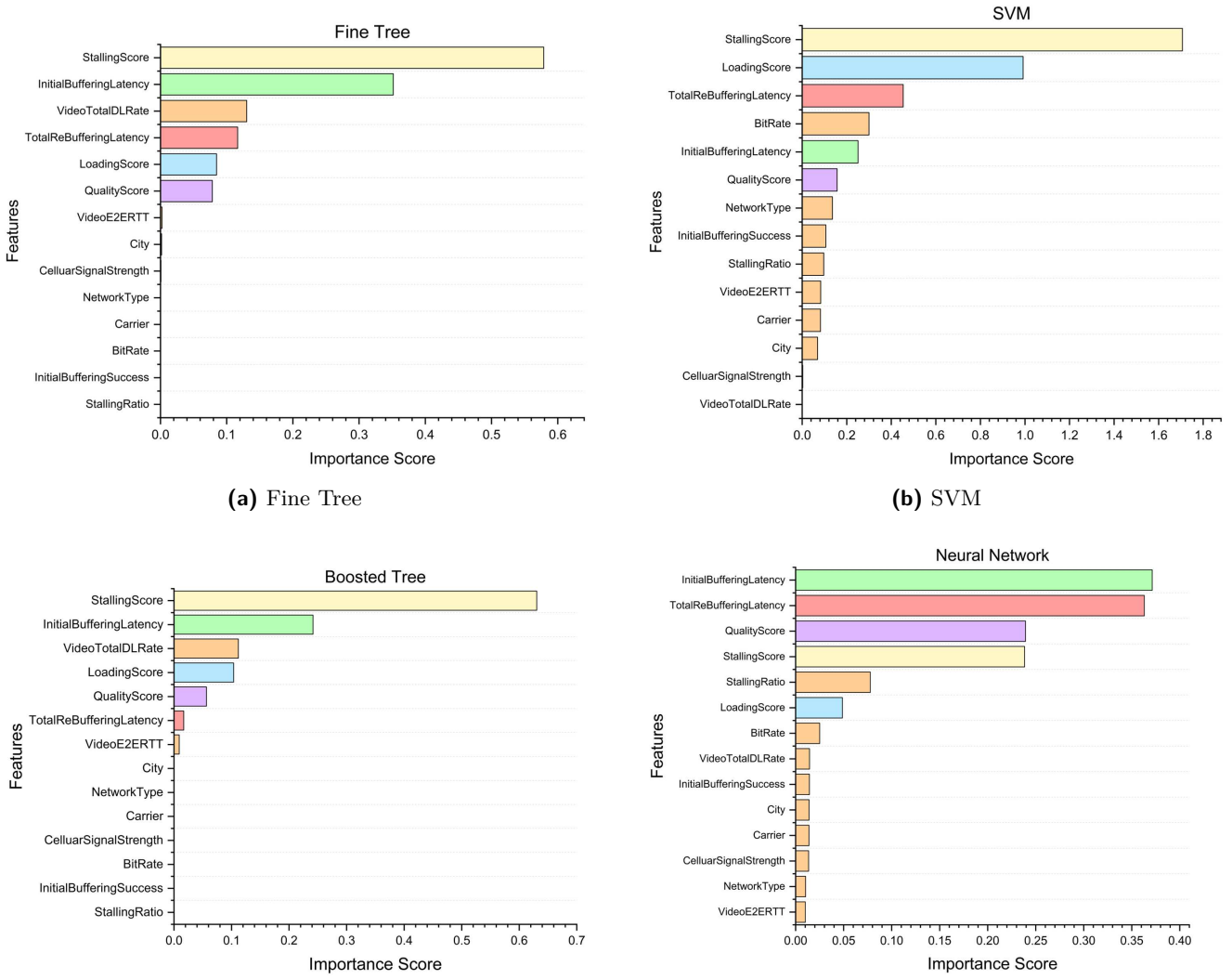


Figure 6. Evaluation validation and test datasets across several models for MRMR

5.2 SHAP Feature selection

We used SHAP selection for all models to identify the most influential features. Afterward, we selected the common features across all SHAP models to ensure consistency and generalizability. Figure 7 illustrates the SHAP feature selection results for all models. The five most important features are: (Quality Score, Loading Score, Stalling Score, Initial Buffering Latency, and Total ReBuffering Latency).



(a) Fine Tree **(b)** SVM

(c) Boosted Tree **(d)** Neural Network

Figure 7. Features selection across all SHAP models (a) Fine Tree (b) SVM (c) Boosted Tree (d) Neural Network

From Figure 8, the neural network maintains outstanding performance, with an RMSE of 0.009 on validation and 0.007 on test data, and a perfect R-squared (1.000) on both datasets. This model appears largely unaffected by the feature reduction, indicating that the selected features capture the critical information for vMOS prediction very effectively. The extremely low errors (MSE of approximately 0 on validation and test) show that the neural network continues to be highly accurate and stable. Table 4 summarizes the performance of four different models for SHAP.

The Fine Tree also shows strong performance with the reduced feature set, achieving an RMSE of 0.102 on validation and 0.070 on test, and an R-squared of 0.993 on validation and 0.997 on test. These results indicate that the decision tree generalizes well and has improved its accuracy slightly compared to previous results, suggesting that the selected features are highly relevant for this model.

The Boosted Tree model remains consistent, with an RMSE of 0.148 on validation and 0.135on test data, and an R-squared of 0.986 on validation and 0.989 on test. The performance is stable but slightly lower than the Fine Tree

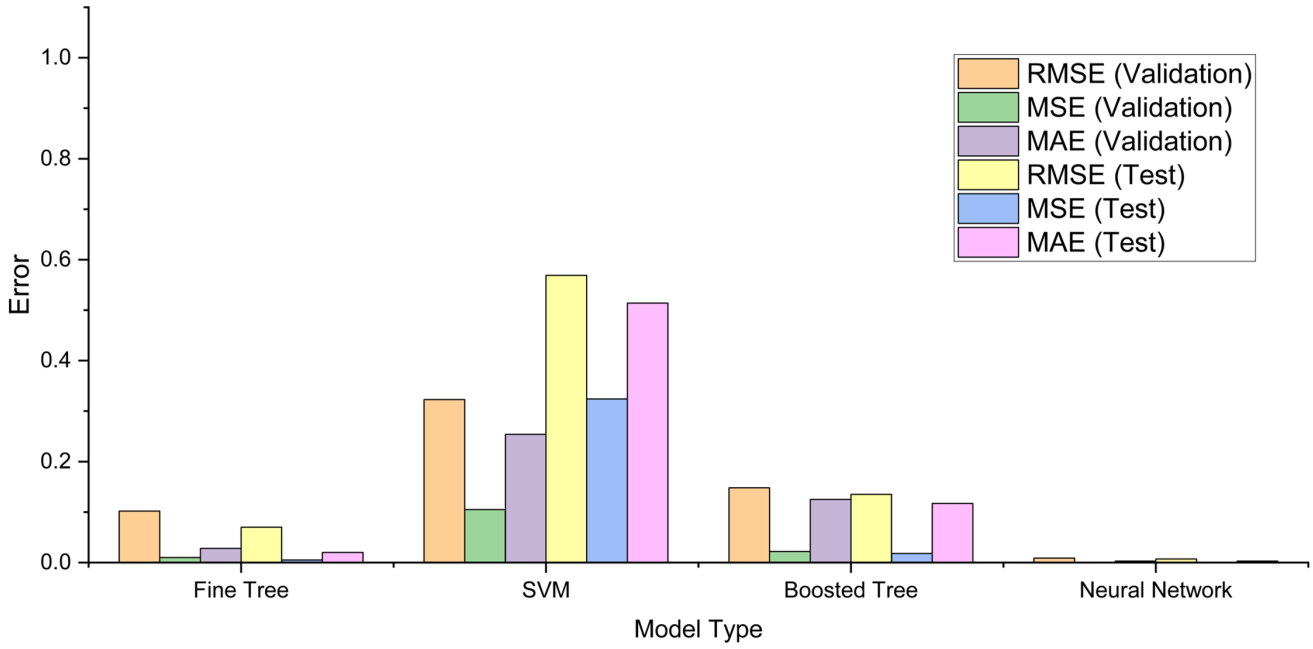


Figure 8. Evaluation validation and test datasets across several models with various error metrics

and Neural Network, indicating that while the model performs well with the selected features, it may not be as well-suited to this reduced feature set as the top-performing models.

SVM has the highest RMSE (0.323 on validation and 0.569 on test) and the lowest R-squared (0.933 on validation and 0.811 on test) among all models, even after feature selection. This suggests that SVM is less effective for this task, possibly due to limitations in handling the complexity or distribution of the data with only five features.

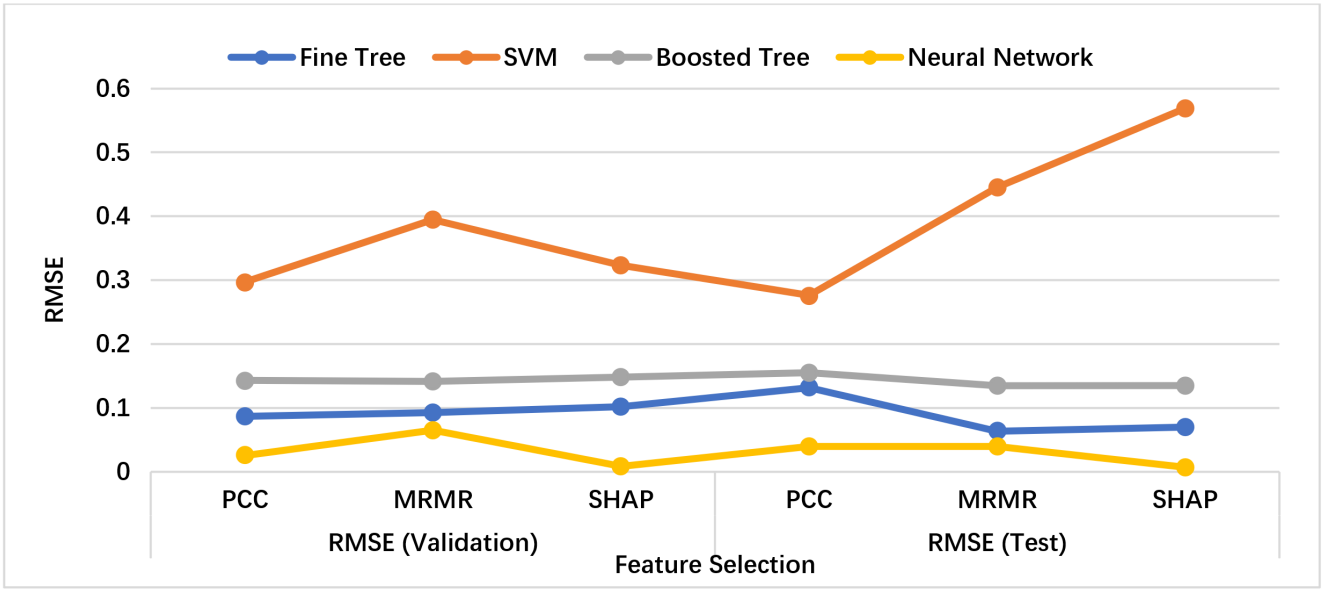
Table 4: Summarizes the performance of four different models for SHAP

	Validation				**Test**			
Model Type	**RMSE**	**MSE**	**MAE**	**RSquared**	**RMSE**	**MSE**	**MAE**	**RSquared**
Fine Tree	0.102	0.010	0.028	0.993	0.070	0.005	0.020	0.997
SVM	0.323	0.105	0.254	0.933	0.569	0.324	0.514	0.811
Boosted Tree	0.148	0.022	0.125	0.986	0.135	0.018	0.117	0.989
Neural Network	**0.009**	0.000	0.003	1.000	0.007	0.000	0.003	1.000

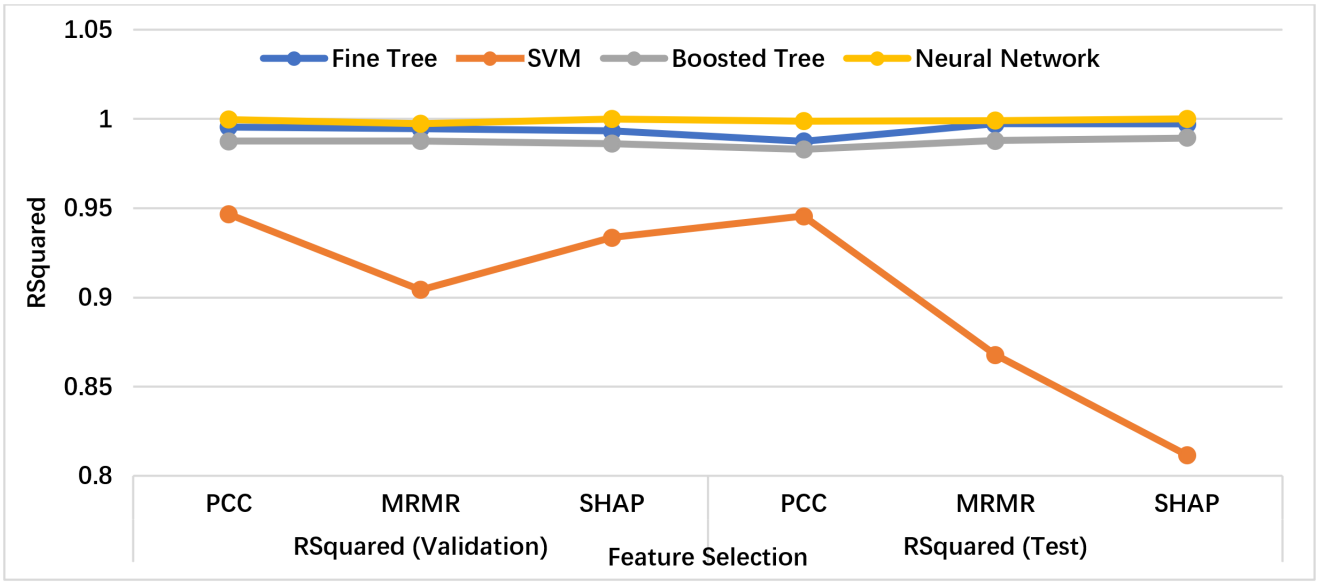
Neural Network remains the most accurate model, showing minimal degradation in performance despite the reduction in features. This suggests that the selected features are highly representative for predicting MOS, making the Neural Network an ideal choice. Fine Tree also performs very well with the reduced feature set, showing strong generalizability and accuracy.

Boosted Tree retains reasonable performance but shows slightly less adaptability to the reduced feature set. SVM remains the least effective model, with the highest errors and lowest R-squared, indicating it may not be well-suited to this problem even with fewer features.

Figure 9 illustrates the performance of four models based on RMSE and R-squared metrics for validation and test sets across three feature selection methods: PCC, MRMR, and SHAP. The Neural Network outperforms the other models, offering the best trade-off between RMSE and R-squared metrics across all feature selection methods for both validation and test sets. In contrast, SVM obtained poorer performance particularly on the test set compared to the other models, especially in SHAP method. Additionally, feature selection methods have a more pronounced impact on SVM compared to the other models.



(a)



(b)

Figure 9. Performance of four models across feature selections for validation and test sets (a) RMSE (b) R-squared metrics

6 Conclusions

This study presented a multifaceted approach to study QoE in mobile networks by analyzing statistical data, predicting the MOS under various feature selection scenarios, and rigorously evaluating AI models using several performance metrics. By identifying key patterns and correlations and demonstrating how network parameters influence MOS, the findings offer valuable insights for optimizing user experience and guiding future advancements in the field. The results have demonstrated that selecting the most relevant features (as identified by MRMR and SHAP) has streamlined the models without compromising accuracy, particularly for the Neural Network and Fine Tree models. This demonstrates the effectiveness of the feature selection process in retaining essential information for accurate MOS prediction. These insights offer valuable guidance for future research and practical implementations in the field.

Data Availability

The datasets analysed during the current study are not publicly available due to licensing and privacy restrictions but are available from the corresponding author on reasonable request.

References

1. Dahj, J. N. M., Ogudo, K. A. & Boonzaaier, L. A hybrid analytical concept to qoe index evaluation: Enhancing embb service detection in 5g sa networks. *J. Netw. Comput. Appl.* **221**, 103765 (2024).

2. Fowler, S. & Souihi, S. Optimizing qoe for virtual reality games on mobile edge networks. In *2024 9th International Conference on Fog and Mobile Edge Computing (FMEC)* (IEEE, 2024).

3. Kao, H.-W. & Wu, E. H.-K. Qoe sustainability on 5g and beyond 5g networks. *IEEE Wirel. Commun.* **30**, 118–125 (2023).

4. International Telecommunication Union. Series p: Telephone transmission quality, telephone installations, local line networks, vocabulary for performance, quality of service and quality of experience. t p.10/g.100. Tech. Rep., International Telecommunication Union Telecommunication Standardization Sector, Geneva, Switzerland (2017).

5. International Telecommunication Union. Series p: Terminals and subjective and objective assessment methods: Methods for objective and subjective assessment of speech and video quality: Mean opinion score (mos) terminology, itu-t rec. p.800.1. Tech. Rep., Telecommunication Standardization Sector ITU (2016).

6. Daengsi, T., Sirawongphatsara, P. & Wuttidittachotti, P. Proposed qoe models associated with delay and jitter using subjective approach and applications for 4g and 5g networks. In *2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet)* (IEEE, 2021).

7. Moura, D. *et al.* A no-reference video streaming qoe estimator based on physical layer 4g radio measurements. In *2020 IEEE Wireless Communications and Networking Conference (WCNC)* (IEEE, 2020).

8. El-Saleh, A. A. *et al.* Measurements and analyses of 4g/5g mobile broadband networks: an overview and a case study. *Wirel. Commun. Mob. Comput.* **2023**, 6205689 (2023).

9. Pratiwi, D. & Edward, I. J. M. Analysis efficiency network performance of 4g lte in video conference applications. In *2022 16th International Conference on Telecommunication Systems, Services, and Applications (TSSA)* (IEEE, 2022).

10. El-Saleh, A. *et al.* Mean opinion score estimation for mobile broadband networks using bayesian networks. *CMC-Computers, Mater. & Continua* **72** (2022).

11. Alhammadi, A., El-Saleh, A. & Shayea, I. Mos prediction for mobile broadband networks using bayesian artificial intelligence. In *2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST)* (IEEE, 2021).

12. Aguilar, A. P. *et al.* Objective qoe prediction for video streaming services: A novel full-reference methodology. In *2024 IEEE 99th Vehicular Technology Conference (VTC2024-Spring)* (IEEE, 2024).

13. Wu, R. *et al.* Qos to qoe mapping based on optimised bp neural networks. In *2023 8th International Conference on Image, Vision and Computing (ICIVC)* (IEEE, 2023).

14. Ajeyprasaath, K. & Vetrivelan, P. A hybrid machine learning approach for improvised qoe in video services over 5g wireless networks. *Comput. Mater. & Continua* **78** (2024).

15. Arnez, J. J. A. & De Souza, C. B. B. Analysis of inter-mean-opinion-score (imos) and qos metrics for 5g voice over lte (volte) and voice over new radio (vonr) services over ip-multimedia-subsystem (ims) in 5g mobile networks. In *2023 6th International Conference on Advanced Communication Technologies and Networking (CommNet)* (IEEE, 2023).

16. Barazzetta, M. *et al.* Testing of volte mean opinion score in reverberation chambers. *IET Sci. Meas. & Technol.* **14**, 949–954 (2020).

17. Kunešová, M. *et al.* Ensemble of deep neural network models for mos prediction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2023).

18. Leonte, S. *et al.* Voice quality evaluation in a mobile cellular network: In situ mean opinion score measurements. *Sensors* **24**, 6630 (2024).

19. Mouhassine, N. & Moughit, M. A mean opinion score prediction model for voip calls offloading handover from lte to wifi. *Clust. Comput.* 1–19 (2024).

20. Shayea, I. *et al.* Performance analysis of mobile broadband networks with 5g trends and beyond: Urban areas scope in malaysia. *IEEE Access* (2021).

Acknowledgements

This work was funded by Omantel under the project number EG/SQU-OT/25/02.

Author contributions statement

Conceptualization, H.M.A., A.A., N.T., M.M.B.-S, and I.S.; methodology, H.M.A., A.A., N.T., and I.S.; writing original draft preparation, A.A.; writing review and editing, H.M.A., N.T., M.M.B.-S.and I.S.; supervision, H.M.A. and N.T.; project administration, H.M.A. All authors reviewed the manuscript.