

**CHENNAI – PONDICHERRY**

**Toward Better Statistical Validation of Machine Learning-Based Multimedia Quality Estimators**

**Abstract**

Objective assessment of multimedia quality using machine learning (ML) has been gaining popularity especially in the context of both traditional (e.g., terrestrial and satellite broadcast) and advance (such as over-the-top media services, IPTV) broadcast services. Being data-driven, these methods obviously rely on training to find the optimal model parameters. Therefore, to statistically compare and validate such ML-based quality predictors, the current approach randomly splits the given data into training and test sets and obtains a performance measure (for instance mean squared error, correlation coefficient etc.). The process is repeated a large number of times and parametric tests (e.g., t test) are then employed to statistically compare mean (or median) prediction accuracies. However, the current approach suffers from a few limitations (related to the qualitative aspects of training and testing data, the use of improper sample size for statistical testing, possibly dependent sample observations, and a lack of focus on quantifying the learning ability of the ML-based objective quality predictor) which have not been addressed in literature. Therefore, the main goal of this paper is to shed light on the said limitations both from practical and theoretical perspectives wherever applicable, and in the process propose an alternate approach to overcome some of them. As a major advantage, the proposed guidelines not only help in a theoretically more grounded statistical comparison but also provide useful insights into how well the ML-based objective quality predictors exploit data structure for learning. We demonstrate the added value of the proposed set of guidelines on standard datasets by comparing the performance of few existing ML-based quality estimators. A software implementation of the presented guidelines is also made publicly available to enable researchers and developers to test and compare different models in a repeatable manner.