**CONCLUSION:**

With the growing demands for more immersive quality of experience from consumers, quality monitoring in multimedia content delivery especially via broadcast services assumes a significant role in today’s scenario. To that end, ML based quality predictors offer a plausible solution. Moreover, promising results from related disciplines such as computer vision and the availability of required hardware (e.g., GPU-accelerated computing) have opened up possibilities of developing efficient ML based implementations of quality predictors. However, proper validation and benchmarking of such ML based quality estimators is important prior to deployment. In that context, the main goal of the paper was to highlight few drawbacks associated with the current approach of statistical comparison and validation. These stem primarily from lack of considerations to theoretical and practical aspects of statistical testing. Therefore, the main goal of the paper was to raise awareness about some of the identified issues in the current approach. We also provided theoretical analysis concerning dependent (correlated) sample observations. Further, we discussed several other limitations related to sample size, the lack of assessment of the magnitude of treatment effect and an almost exclusive reliance on p values to compare ML based quality predictors. We also argued that assessment of learning ability is an important aspect to validate such learning based predictors, and discussed the use of a permutation test to that end.

**FUTURE WORK:**

Essentially, the proposed guidelines treat statistical comparison of ML based quality estimators as a multi-dimensional problem. Accordingly, we seek to assess the predictors more holistically in terms of their local performance on specific test conditions, their learning ability and the magnitude of treatment effect (in order to quantify the practical significance of the observed differences). In contrast, the current approach tends to reduce this task to binary and global statistical decision making, and does not reveal systematic weaknesses (or strengths) of the predictors. In order to provide a tool for practical use, software implementing the proposed guidelines is made publicly available.