This paper involves data engineering to produce data collection, preprocessing and modeling training. This evaluates different data sampling strategies to optimize model training, while ensuring data stability and efficiency.

Three datasets were utilized, PPumadyn-8nm, Pumadyn-32nm and an Electric Motor Temperature Dataset. The first two datasets consist of simulated data from a Puma 560 robotic arm, filled with kinematic and dynamic parameters. The Electric Motor Temperature dataset contains Sensor-based dataset for predicting surface temperatures in a permanent magnet synchronous motor.

The Machine Learning Models used were Extreme Gradient Boosting (XGBoost) and LightGBM (LGBM), both gradient boosting techniques for regression tasks. Bagging techniques were used to help stabilize prediction and accuracy by combining multiple models/learners.

Three data sampling methods were used, Random Sampling, Latin Hypercube Sampling (LHS), Query-by-Committee (QBC). For Random Sampling, data points are chosen randomly. LHS uses a structured sampling method to establish comprehensive coverage of the feature space. QBC uses an active learning approach where an ensemble of models selects new data points that improve learning efficiency.

The datasets were divided into training and test sets, with cross validation (CV) used for 10 folds. QBC relies on an unlabeled pool of data, with LGB, models choose new training samples. Prediction performance was measured using the variance score.

The findings in this paper showed that QBC sampling with bagging outperforms other strategies in terms of accuracy and stability. Random sampling is viable when experimental costs are low. LHS performs well but requires predefined sample counts, making it less flexible.

QBC is the most effective choice, for incremental learning.

In contrast, my approach emphasizes data preprocessing and normalization. I propose using the Box-Cox transformation to address data skewness and improve its suitability for analysis. I also suggest using Kernel Density Estimation (KDE) to smooth the data and better visualize its distribution. A key element of my method is the removal of unrealistic outliers, such as those caused by simulation artifacts, like holes positioned too close to the beam's edge, which lead to unrealistic stress concentrations. I also mentioned standardization when comparing different simulation runs to eliminate scale and unit discrepancies.