1. Stability of Rankings (Quantitative Agreement)

CMAPSS

- **Random Forest**: 1-trial and 25-trial results are almost identical, with 9 out of 10 features the same (90% overlap). Only Sensors 12 and 14 switch positions, showing the model is very stable.

- **XGBoost**: Also stable with 9 out of 10 features identical (90% overlap). Sensor 13 drops in the 25-trial average, indicating only minor volatility.

SECOM

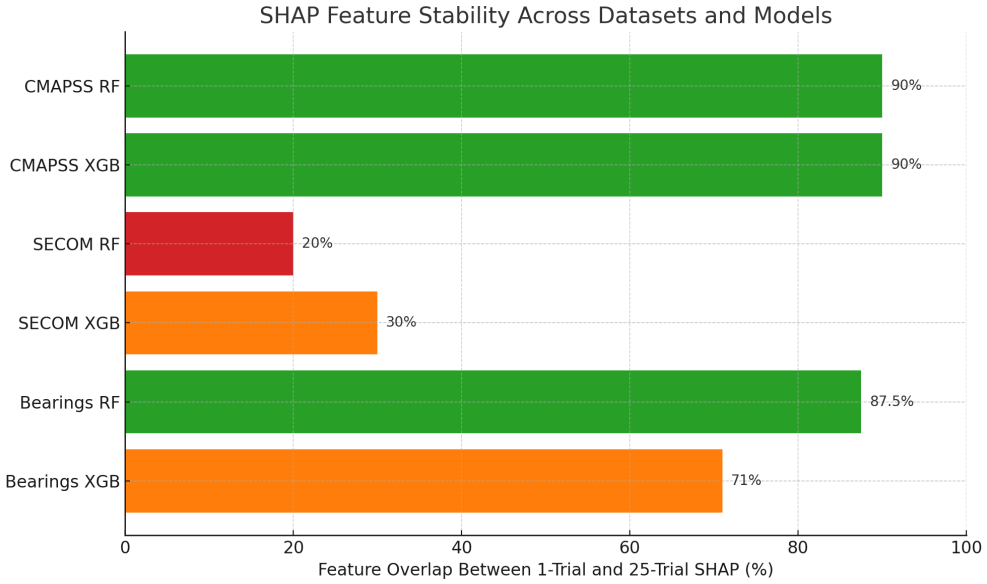
- **Random Forest**: Extremely unstable, with only 2 out of 10 features overlapping between 1-trial and 25-trial (20% overlap). Feature rankings reshuffle drastically.

- **XGBoost**: Slightly better but still unstable, with only 3 out of 10 features overlapping (30% overlap). Multi-trial averaging is clearly required.

NASA Bearings

- **Random Forest**: Mostly stable, with 7 out of 8 features matching between 1-trial and 25-trial (87.5% overlap). Timestamp_index disappears in the 25-trial average, confirming it was a spurious single-trial feature.

- **XGBoost**: Moderately stable, with 5 out of 7 features overlapping (~71%). Single-trial overemphasizes Timestamp_index and misses RMS and Peak_to-Peak, while 25-trial averaging recovers all of the physics-driven vibration features.



2. Feature Order Shifts (Importance Volatility)

- **CMAPSS**: The top 3–4 sensors stay consistent for both models, showing a stable signal.

- **SECOM**: Only Feature 59 consistently remains on top, while the rest reshuffle heavily. This indicates the dataset is noisy and unstable.

- **NASA Bearings**: Standard deviation remains the top feature, serving as a reliable health indicator. Timestamp_index drops after averaging, confirming it was a spurious single-trial signal. RMS and Peak-to-Peak only appear in 25-trial SHAP, showing that multi-trial averaging recovers the true vibration health features.

3. Quantitative Insights

- When **feature overlap exceeds 80%**, 1-trial SHAP is likely sufficient for operational decision-making. This applies to CMAPSS (RF and XGB) and NASA Bearings (RF).

- When **feature overlap is below 50%**, multi-trial averaging becomes mandatory. This applies to SECOM (RF and XGB) and, to a lesser extent, NASA Bearings (XGB).

- **Random Forest is consistently more stable than XGBoost**, likely due to its lower sensitivity to noise. This also aligns with its stronger predictive performance observed in prior experiments.

4. Operational and Business Implications

- **Stable datasets** like CMAPSS and NASA Bearings with RF models allow 1-trial SHAP to provide fast, cost-effective maintenance insights.

- **Noisy datasets** like SECOM or Bearings with XGBoost require multi-trial SHAP (25 or more runs) to prevent misleading interpretations.

- **Multi-trial averaging recovers domain-aligned features** such as standard deviation, RMS, and kurtosis, increasing trust for engineers and operations teams when using model-driven maintenance strategies.