Project 2: Random Forest, Bagging and Gradient Boosting

Group 4:

Daniel González Muela

Francisco Boudagh

Purusothaman Seenivasen

Sky Sunsaksawat

MVE441 Statistical Learning for Big Data

2nd May 2024.

CHALMERS UNIVERSITY OF TECHNOLOGY

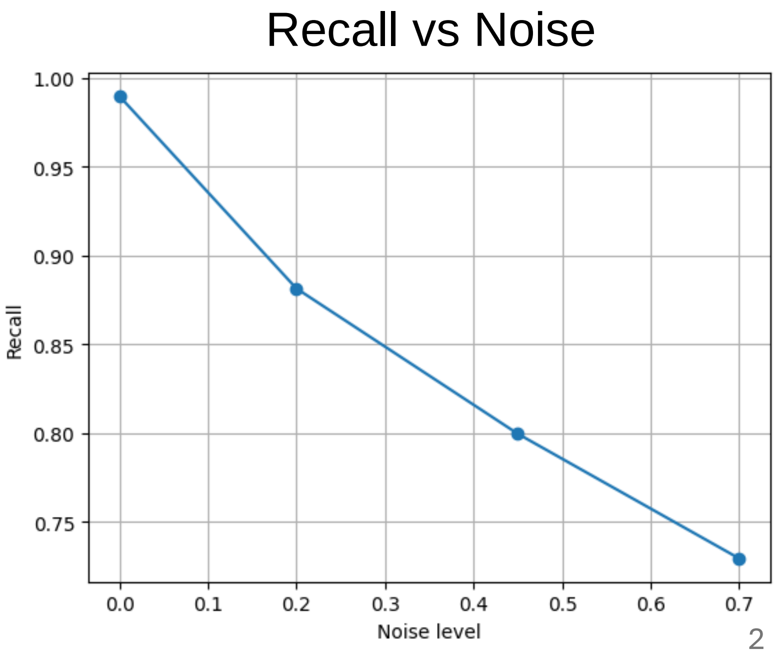
Cancer Data

Adding Noise:

- 2000 Features x 2887 Samples: 6 classes
- For each class, get mean and variance of each of the 2000 features
- Generate artificial sample of a class and label it as a different class
- Our datasets: 0, 20%, 45% and 70% noise

Bagging:

- Create subsets of the original dataset and train some weak learners (Random Forests). Aggregate the outputs
- Best parameters: Use Grid Search → n_estimators as large as possible
- Recall decreases linearly with noise

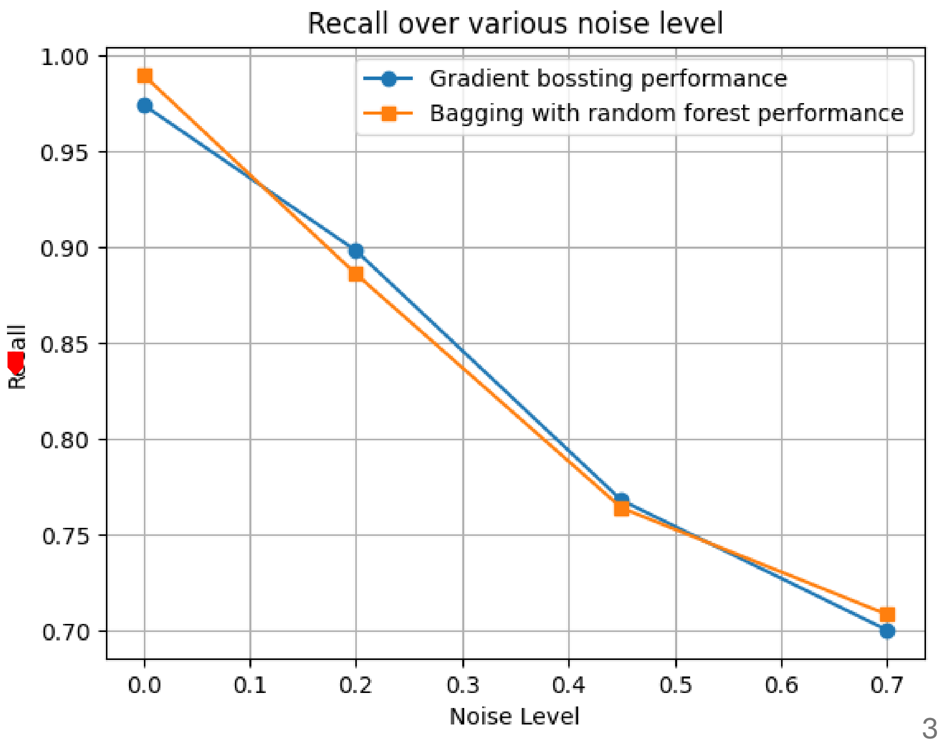


Cancer Data- Boosting

Best parameters found through grid search:
n_estimators: 15 (always getting the maximum number)
Depth: 3

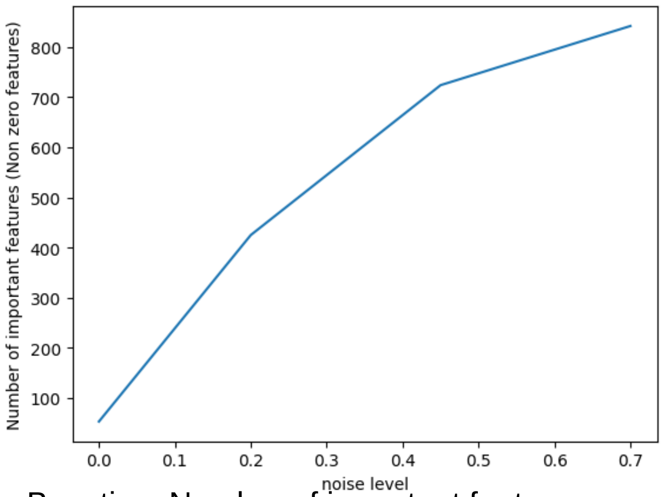
Recall performance of both Gradient boosting and bagging looks similar

Boosting:
- Decision trees are used as weak learners
- Successive trees are added to correct the errors of the combined preceding models.
- Loss function Guides the creation of new trees to correct errors from the previous iterations by minimizing the loss, using the gradient descent approach
- Recall decreases linearly with noise



Feature Importance for Cancer Data-Bagging and Boosting

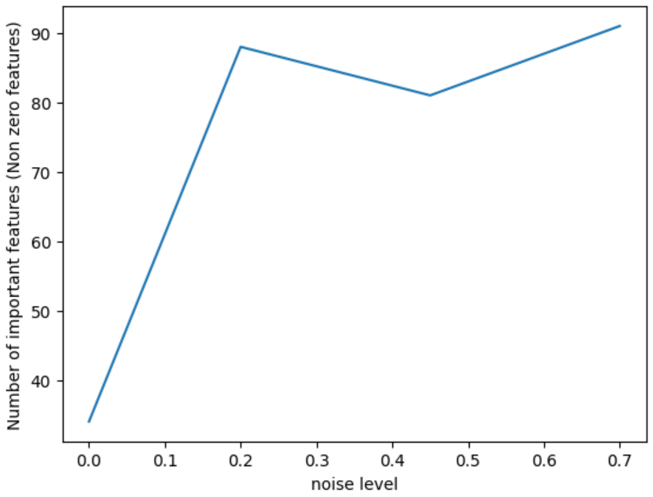
Bagging: Number of important features vs noise



Key points:

- Noise 0 features -> real important features
- Top 10 features: noisy models agree on Feature 3 and 324
- However, no consistency with the others
- Number of important features linearly increase with noise

Boosting: Number of important features vs noise

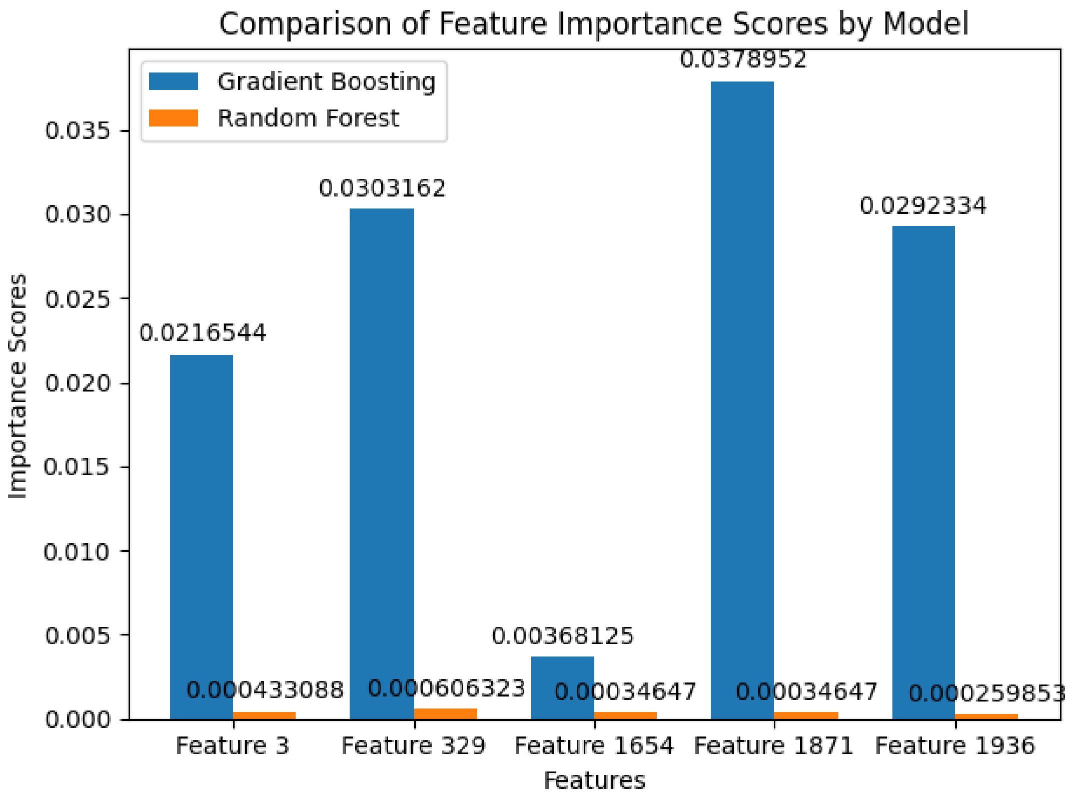


Key points:

- Top 10 features: noisy models agree on Feature 3, 1744, 29,1936, 1657, 657,1871, 1654
- Number of important features linearly increase with noise however number of feature drops for 45% noise.

Common important features

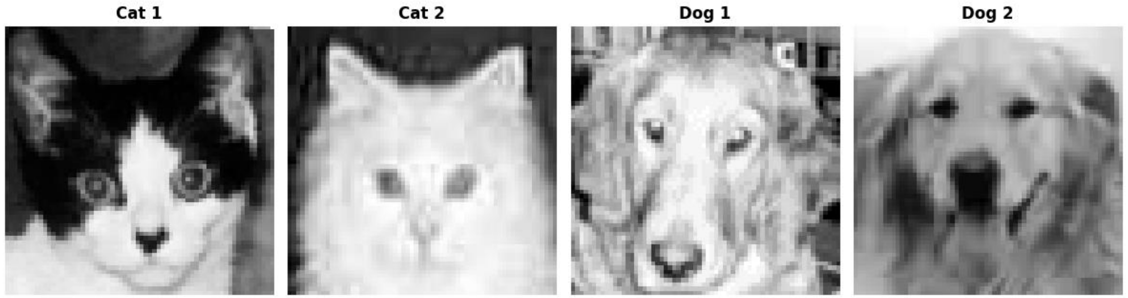
- The common feature that were considered important by both Bagging and boosting
  - Feature 3
  - Feature 329
  - Feature 1654
  - Feature 1871
  - Feature 1936



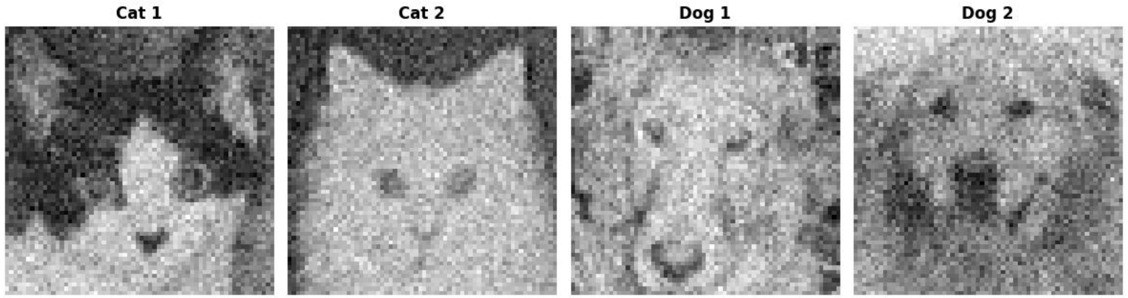
- Imbalance between important scores is because many features are important as random forest bagging

Cats and Dogs, Gaussian Noise

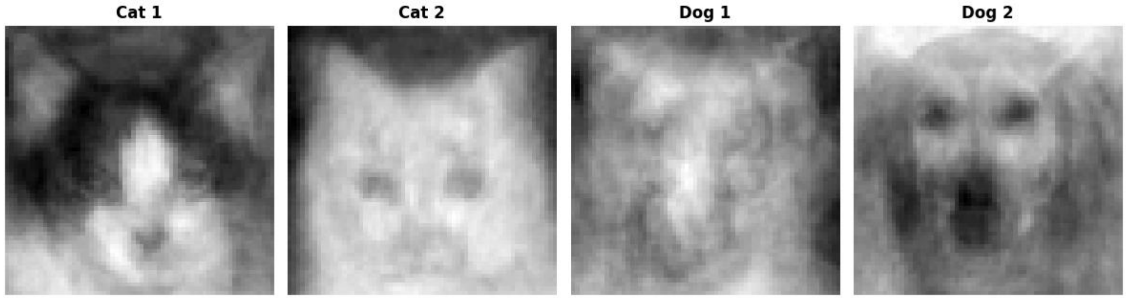
Original Data



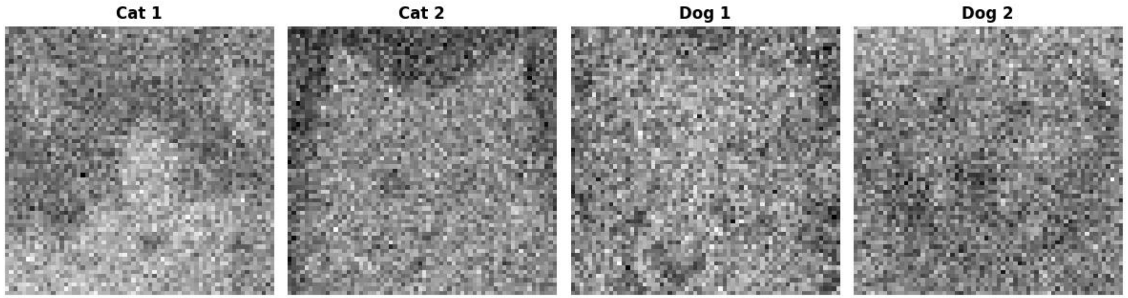
Noise Level: 30



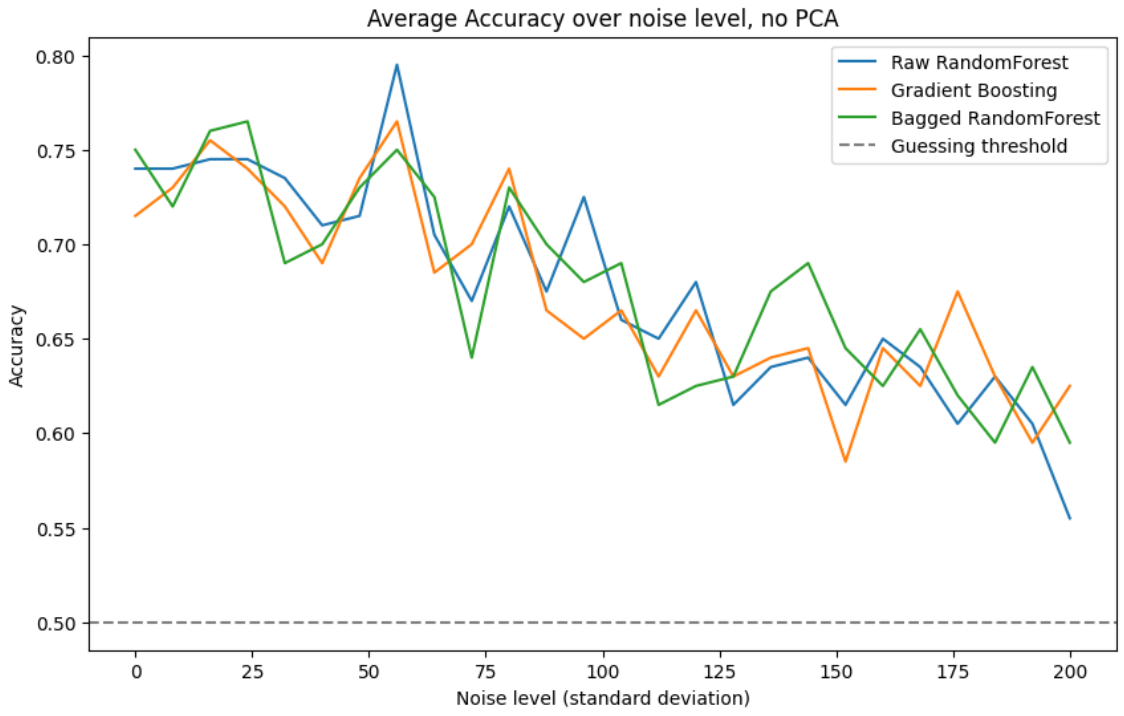
Noise Level: 100

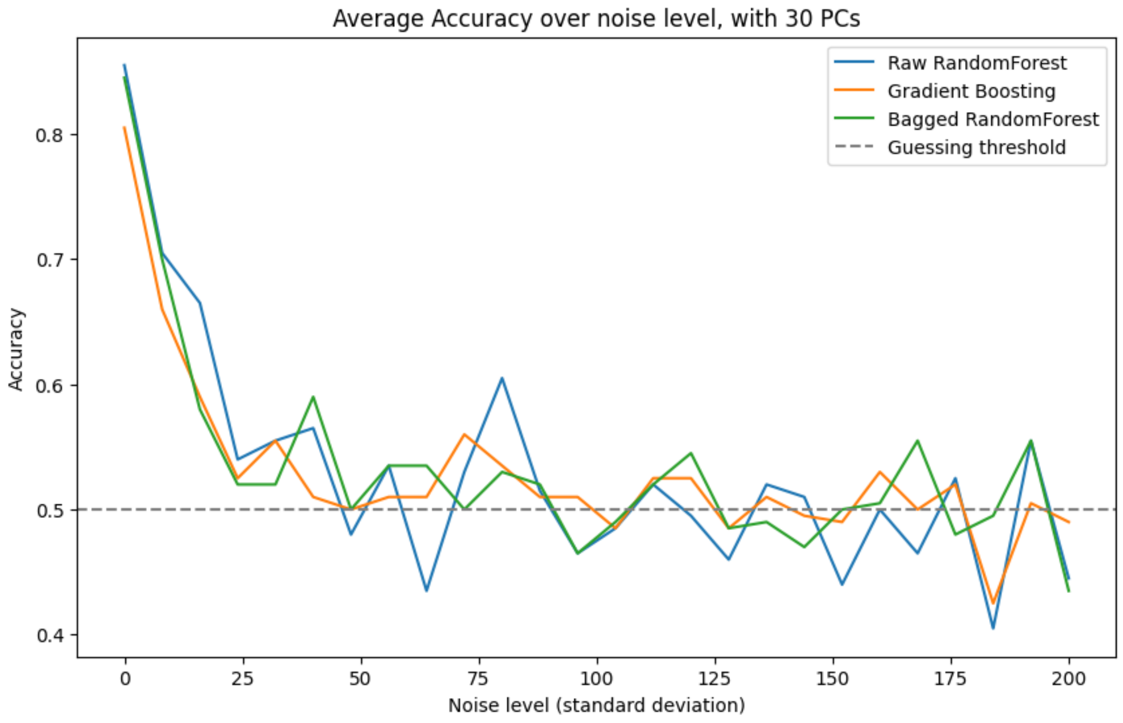


Top 30 PCs



Accuracy vs Noise Level

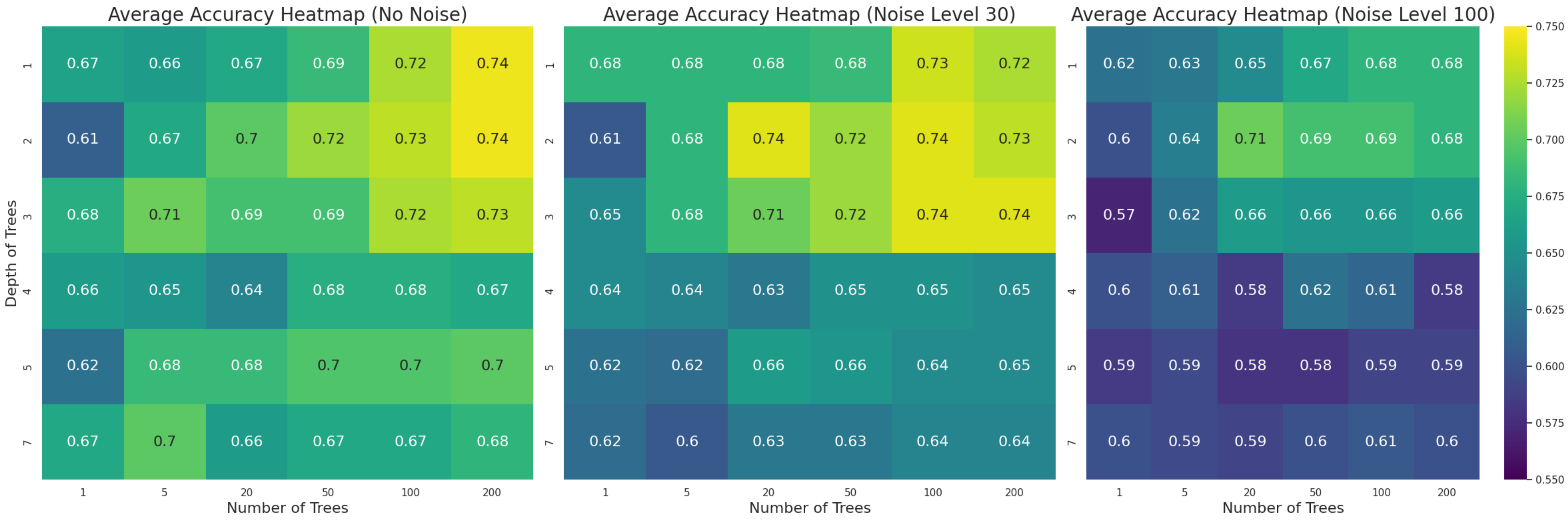




- Bagged Random Forest does not outperform Random Forest – interesting?
- All models similar in performance (accuracy)
- More noise → more errors
- Performance with all PCs ≫ Performance with 30 PCs

Iterations = 5

Making Gradient Boosting Fail



iter = 5

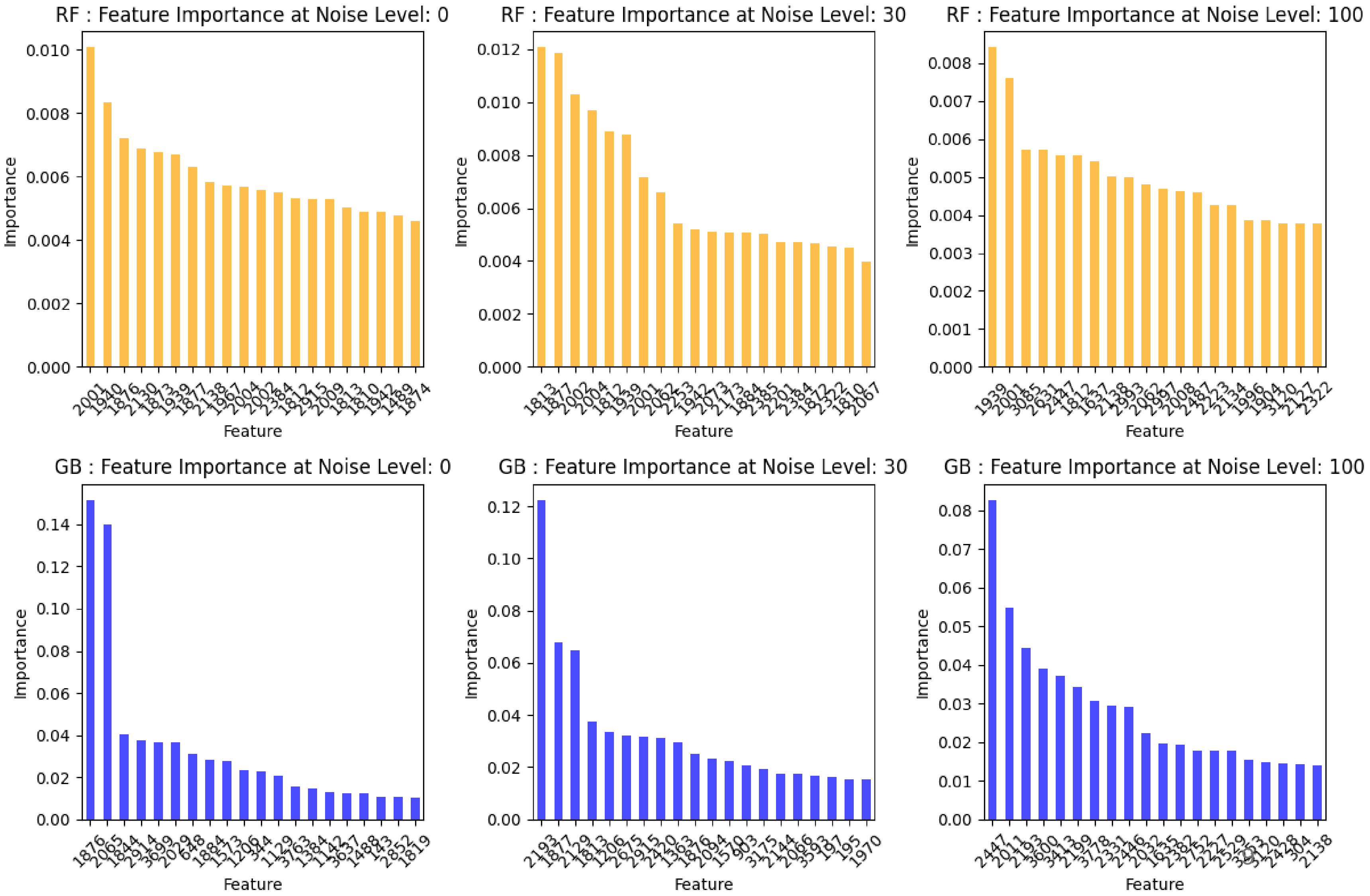
Feature Importance for Cats and Dogs

RandomForestClassifer:

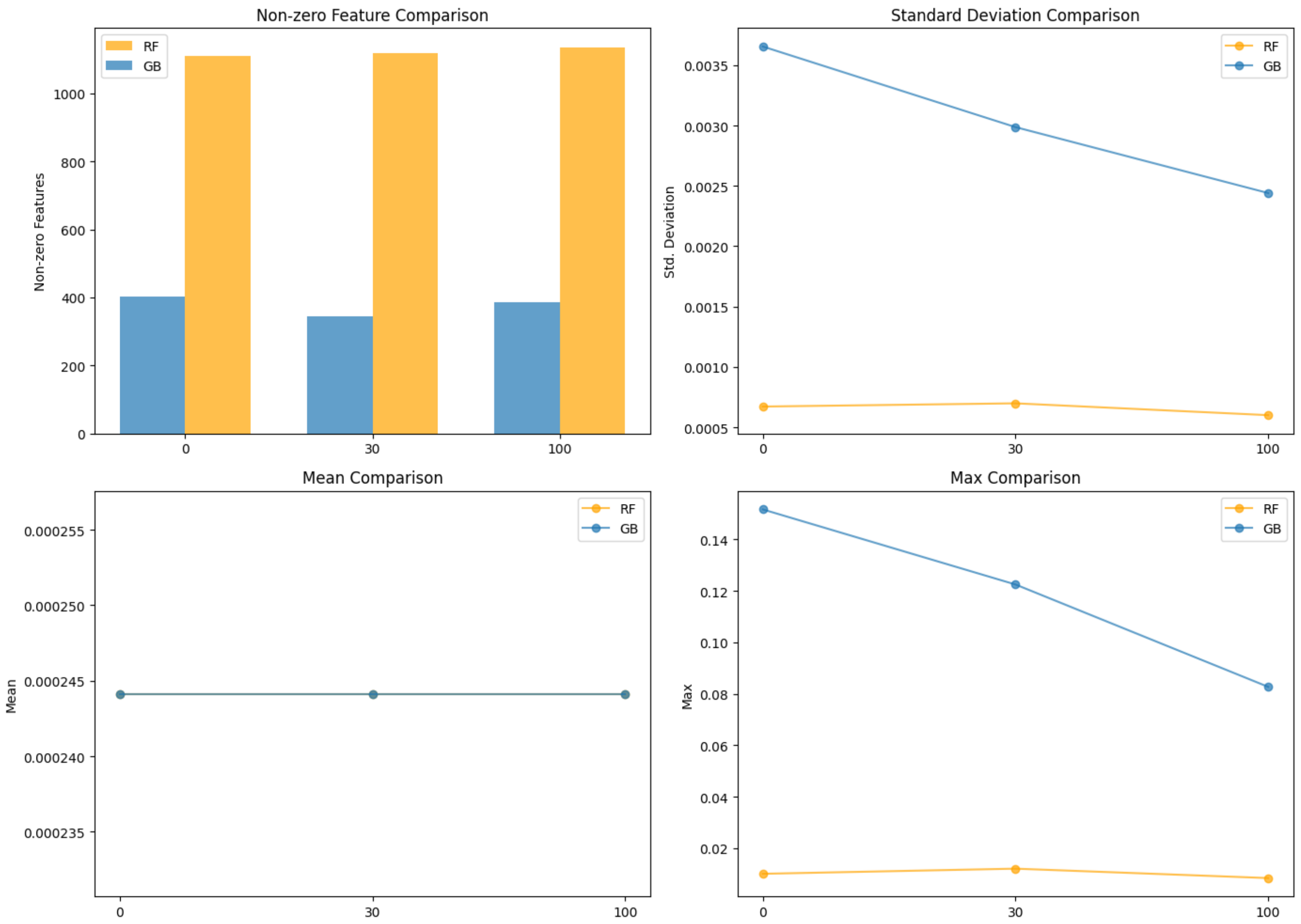
- The impurity-based feature importances.
- Impurity Reduction
- Accumulate Impurity Decrease
- Normalization
- Gini importance
- The higher, the more important the feature

GradientBoostingClassifer :

- Sequential Tree Building
- Impurity Reduction in Each Tree
- Accumulate Impurity Decrease
- Normalization



Feature Importance for Cats and Dogs



<u>Non-zero Feature</u>

- GB shows a decrease in the number of non-zero feature importances as noise increases

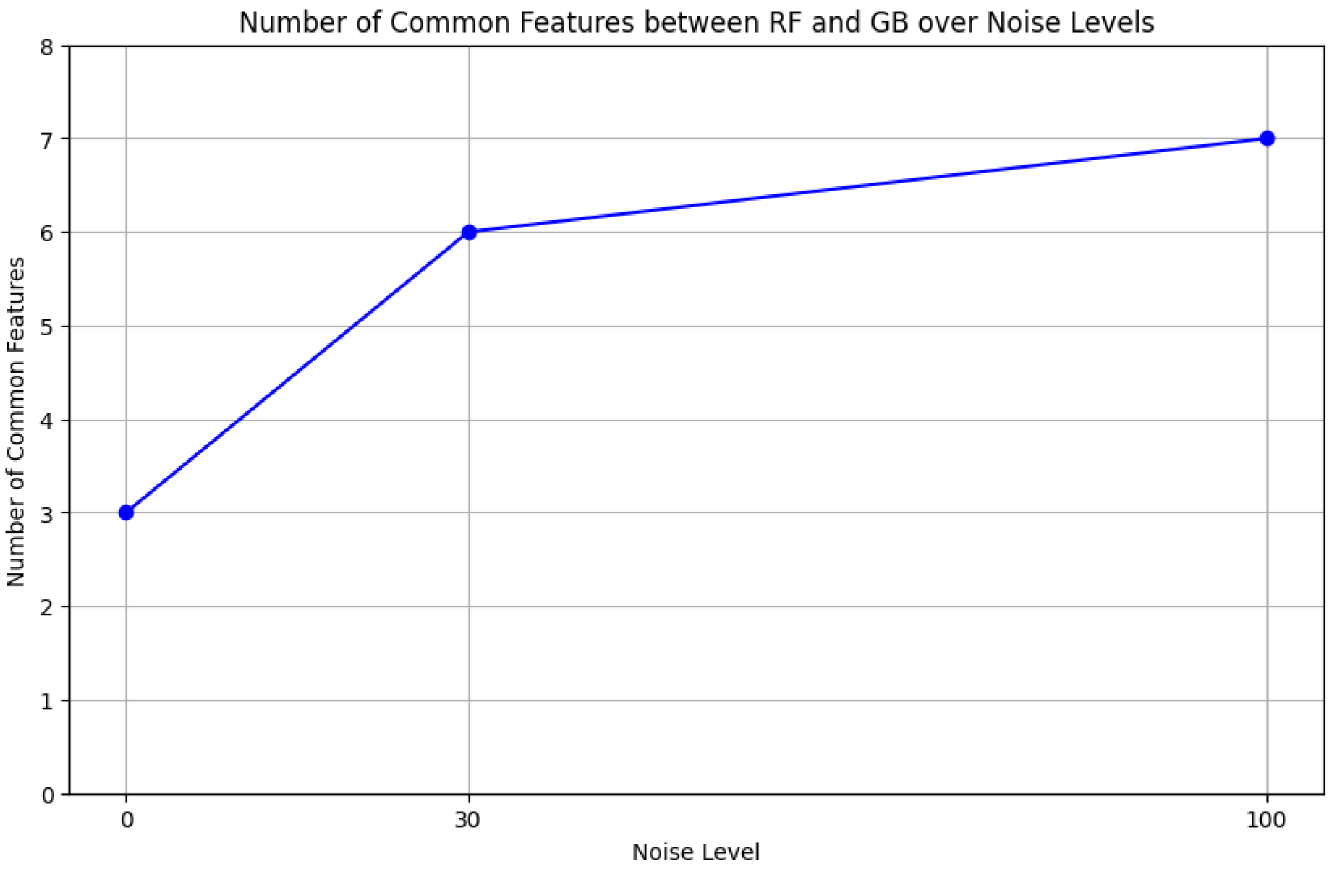
<u>Standard deviation</u>

- GB exhibits higher variability in feature importance scores
- For both models, the standard deviation of feature importances decreases as noise levels increase

<u>Max value</u>

- GB shows much higher maximum feature importance values compared to RF
- The maximum feature importance value for GB decreases with increased noise level

Feature Importance for Cats and Dogs



- Common feature of <u>Top 50 largest importance feature</u>

• <u>Noise level : 0</u>
Overlap between top features:
*1940, **1877**, **1876***

• <u>Noise level : 30</u>
Overlap between top features:
*1741, 2066, 2322, **1876**, 1813, **1877***

• <u>Noise level : 100</u>
Overlap between top features:
2529, 2447, 3120, 2961, 1939, 2134, 2138

- The number of common feature increase because model need more features to classify data