Cancer Genome Analysis

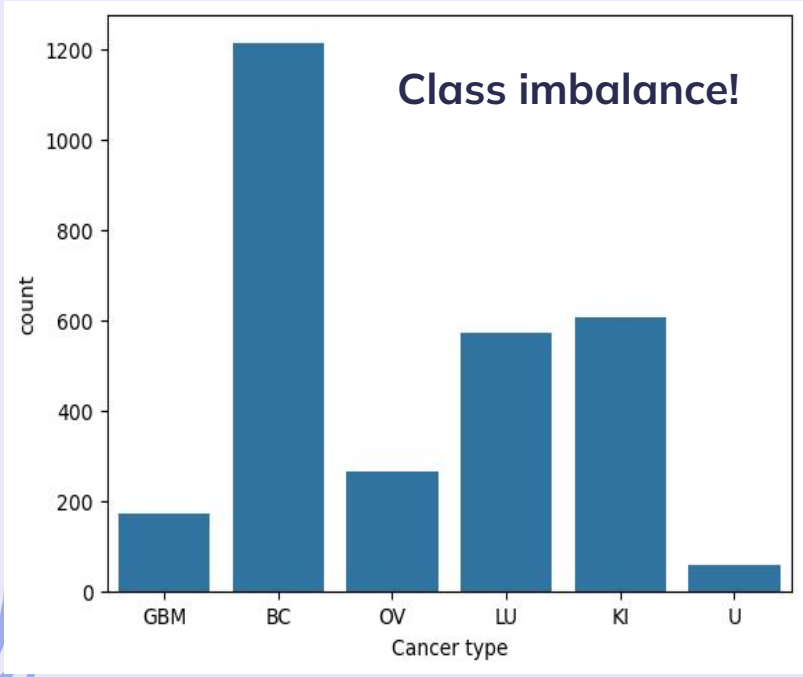
Elínborg Ásbergsdóttir
İpek Korkmaz
Luca Modica
Patrícia Marques

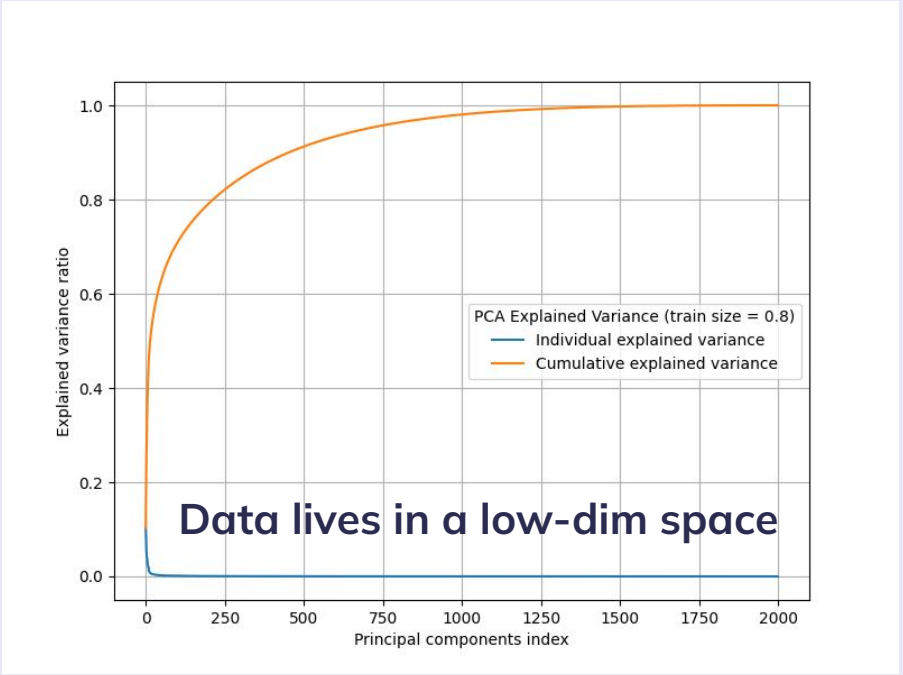
Group 30
Room HC1

18.04.2024

First glance at the dataset

Meaningless features





Methodologies

Dimension Reduction

PCA

Combining features while reducing dimension.

Feature Selection

ANOVA Criteria

Evaluating the impact of each gene based on its variance across different groups.

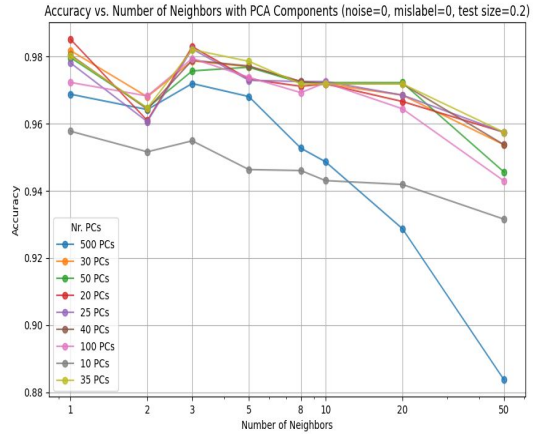
Classifier

KNN Model

Grid search with cross-validation to optimize parameters based on accuracy.

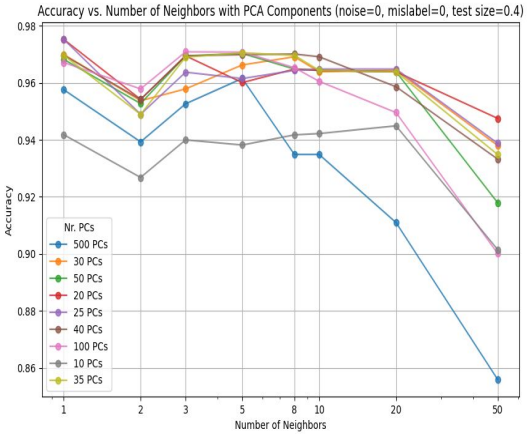
PC selection

Train size 0.8



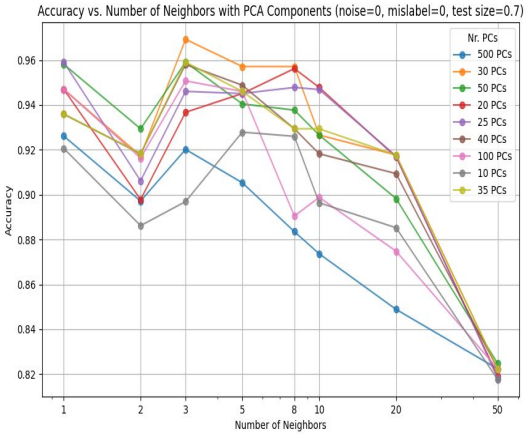
Similar accuracy for 20-100 PCs

Train size 0.6



Still good results for **enough neighbors**

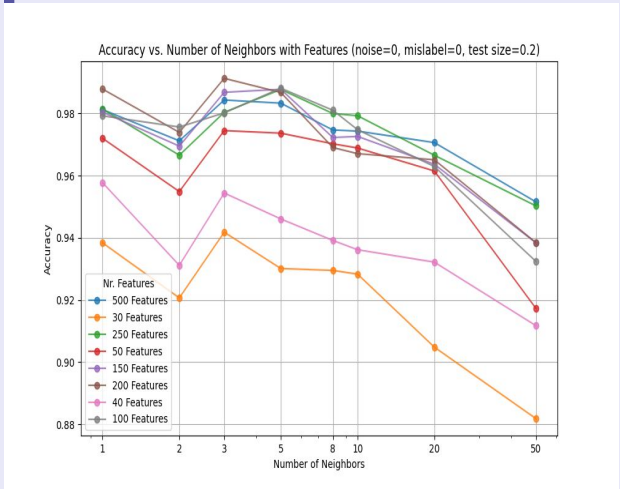
Train size 0.3



Rigid model **decreases accuracy** rapidly

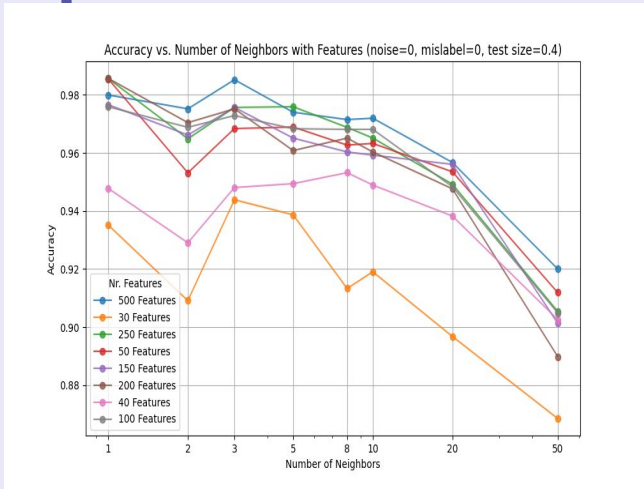
Feature selection

Train size 0.8



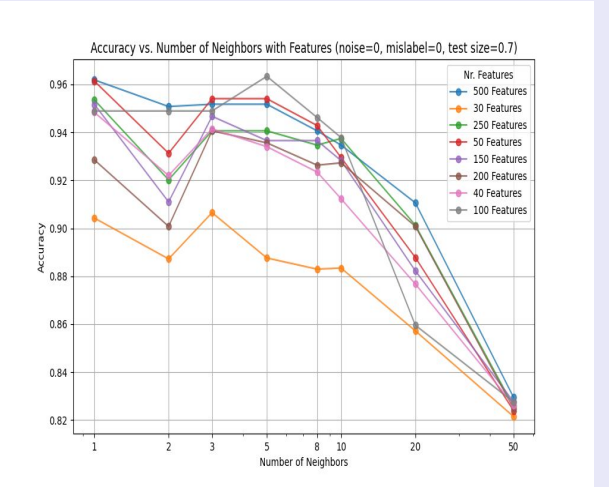
At least **50 features** are **needed**

Train size 0.6



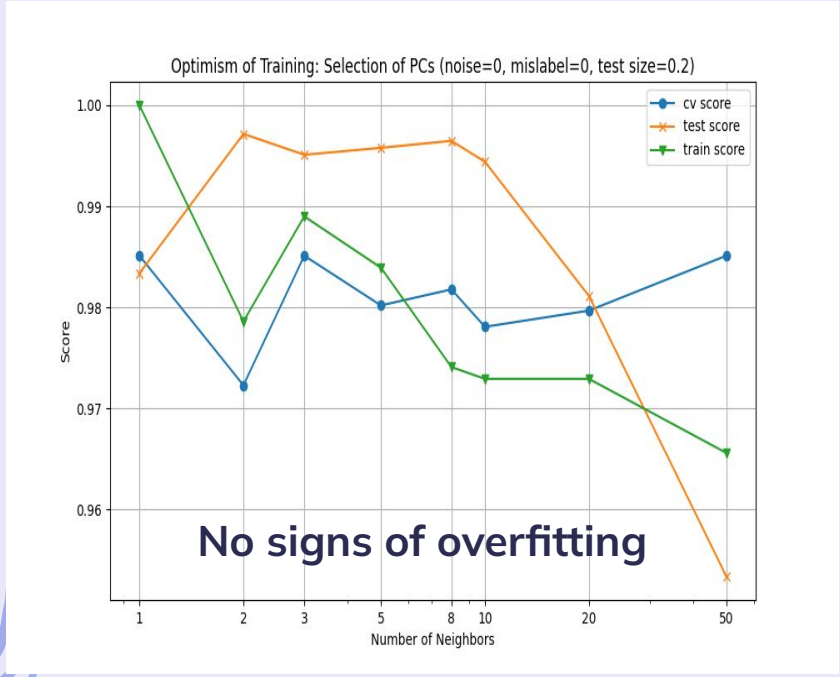
Rigid classifier **performance decreases**

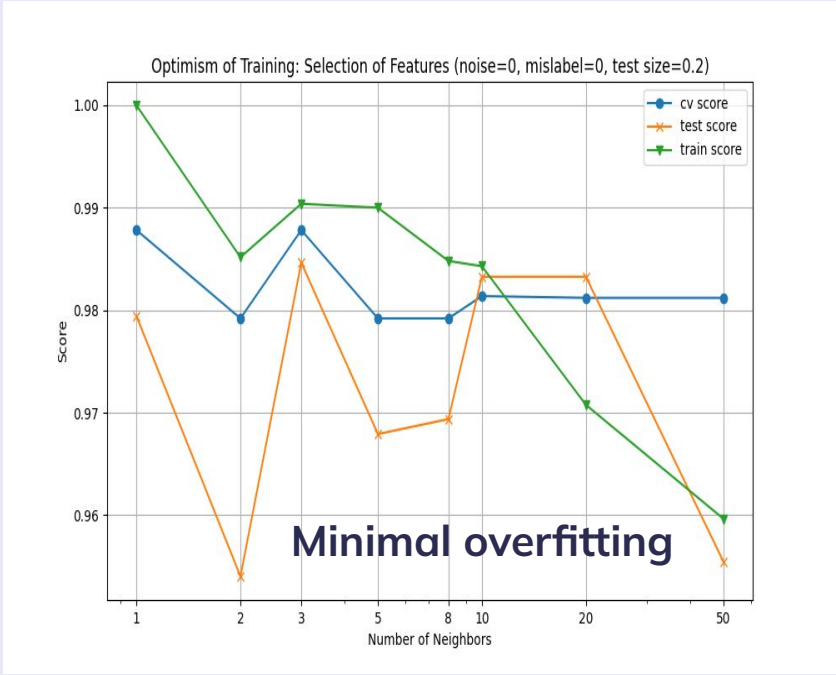
Train size 0.3



CV accuracy still **reasonable**

Optimism of Training (80% training data)



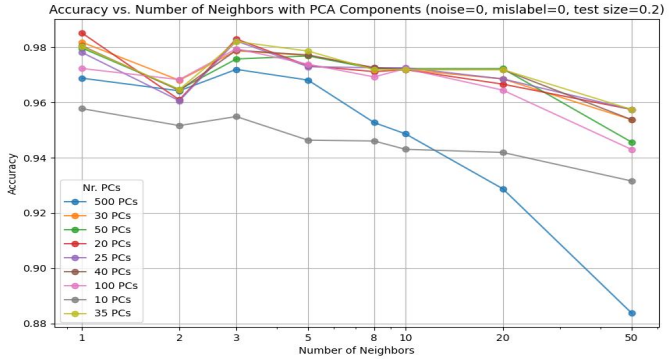


Variations

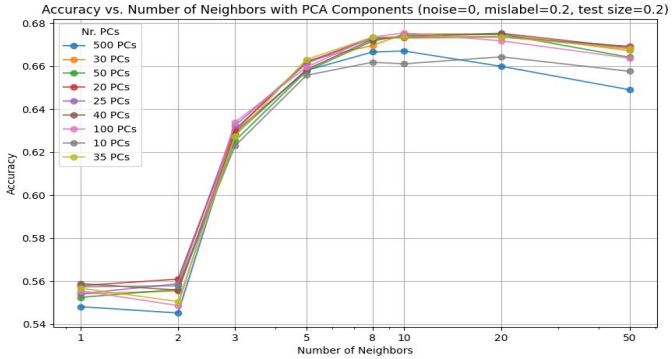
Mislabeling Levels

PC selection (80% training data)

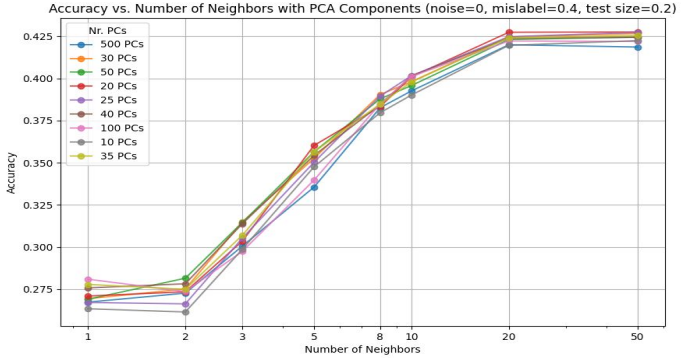
0% mislabeled



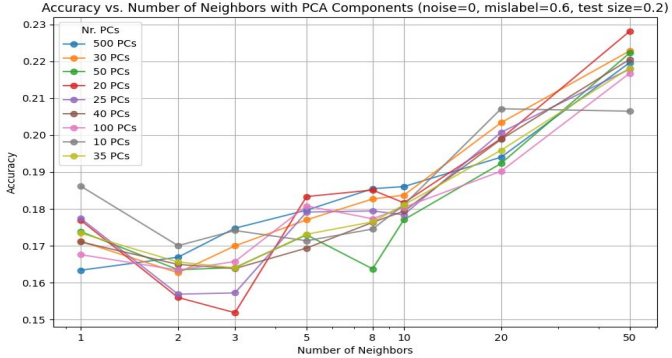
20% mislabeled



40% mislabeled

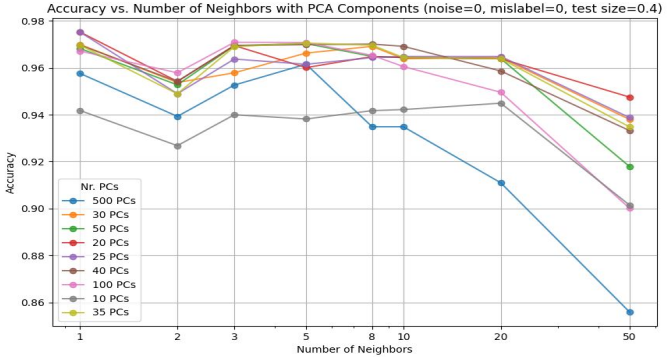


60% mislabeled

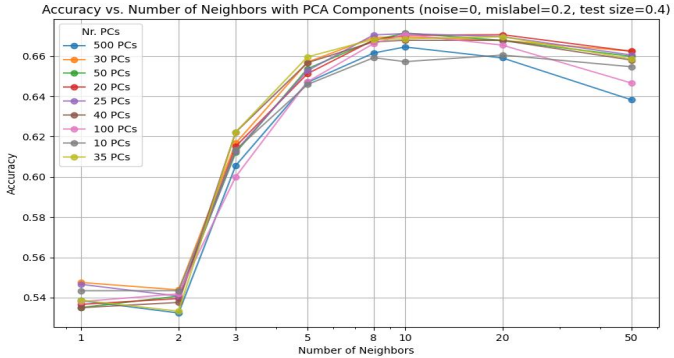


PC selection (60% training data)

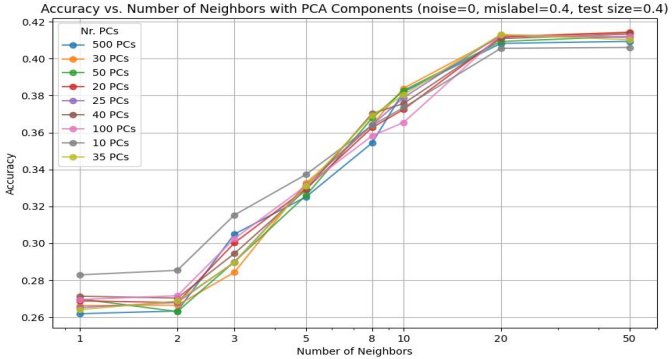
0% mislabeled



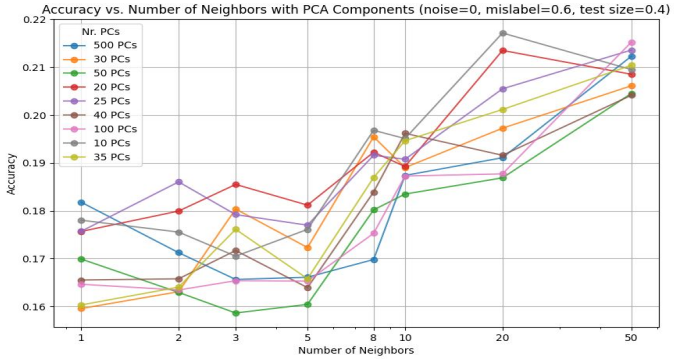
20% mislabeled



40% mislabeled

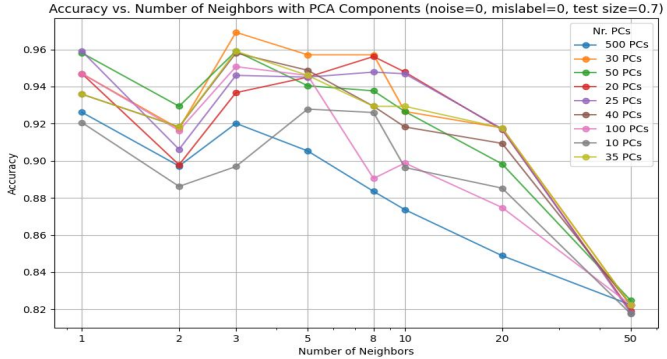


60% mislabeled

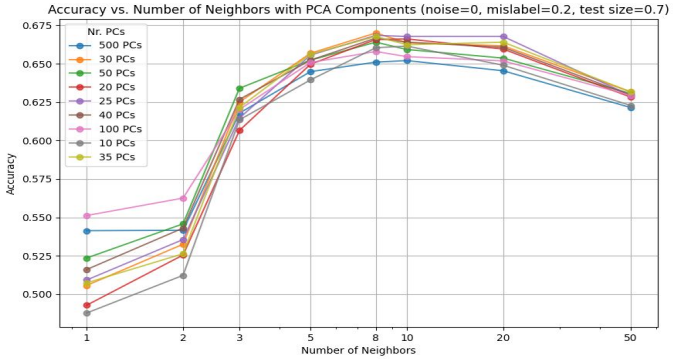


PC selection (30% training data)

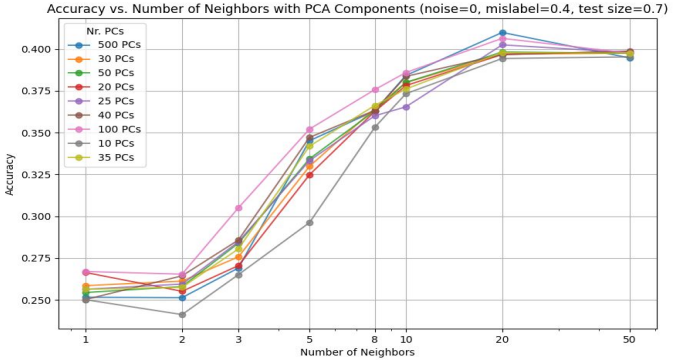
0% mislabeled



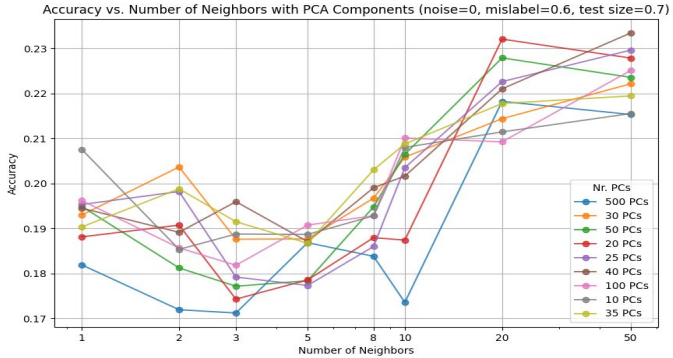
20% mislabeled



40% mislabeled

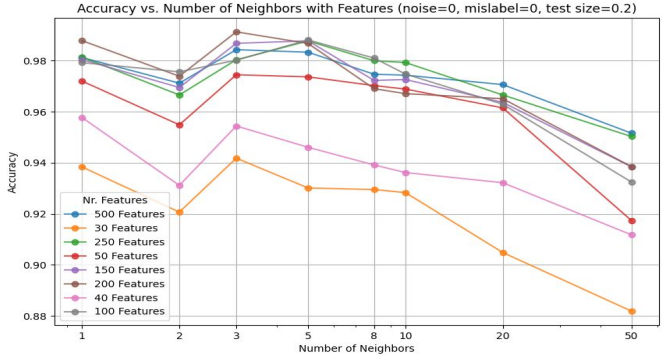


60% mislabeled

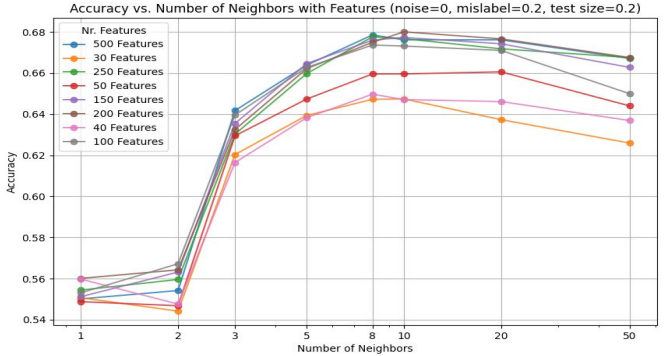


Feature selection (80% training data)

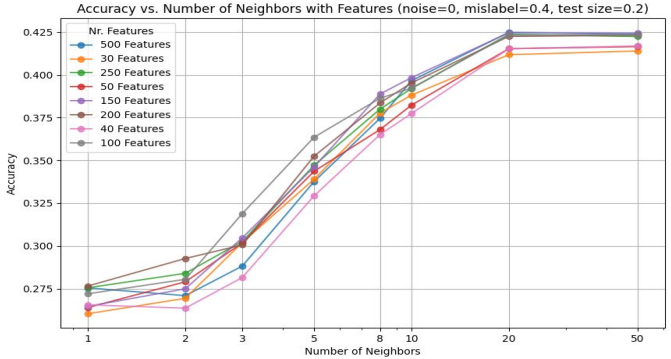
0% mislabeled



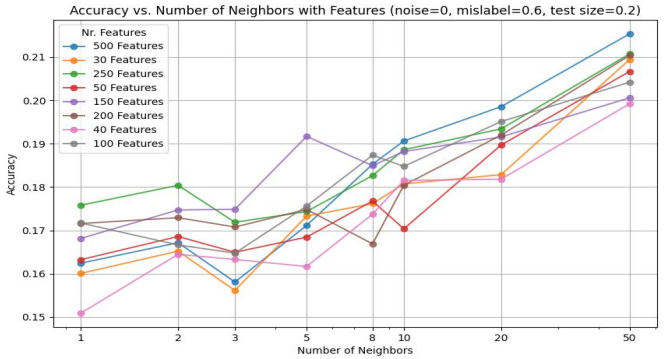
20% mislabeled



40% mislabeled

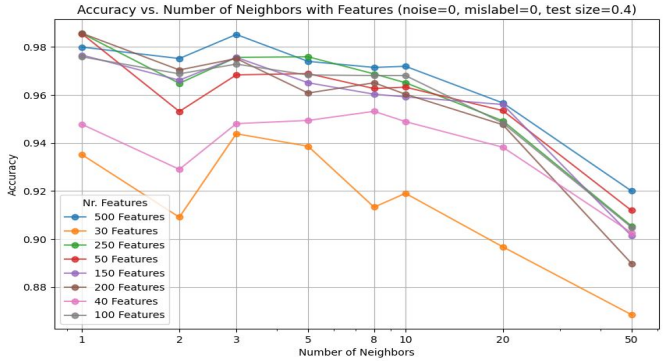


60% mislabeled

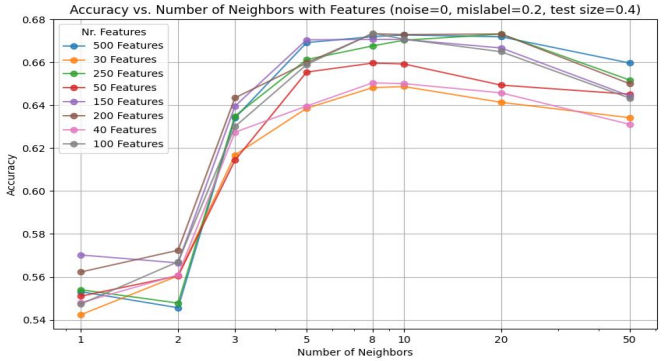


Feature selection (60% training data)

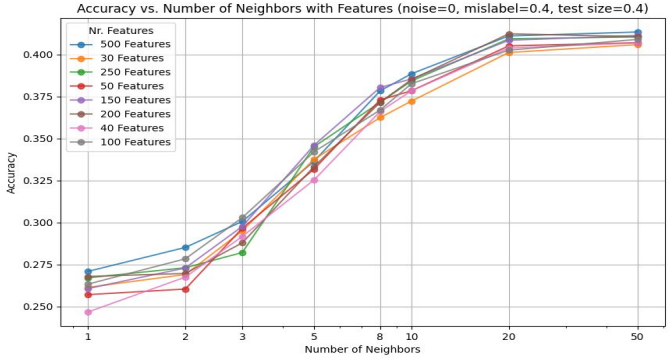
0% mislabeled



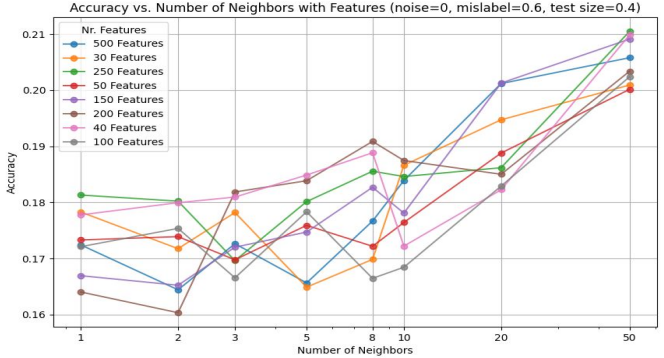
20% mislabeled



40% mislabeled

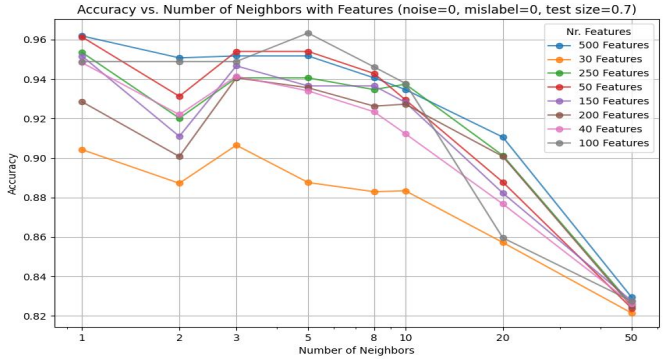


60% mislabeled

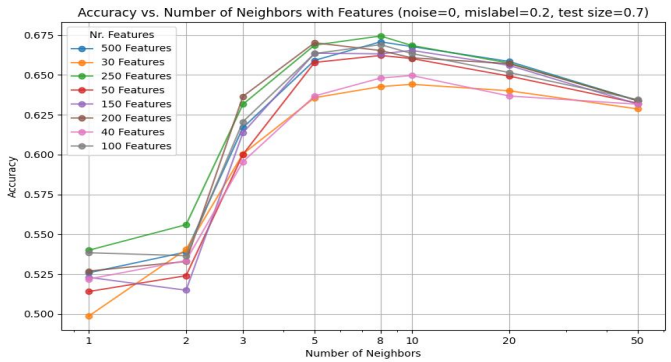


Feature selection (30% training data)

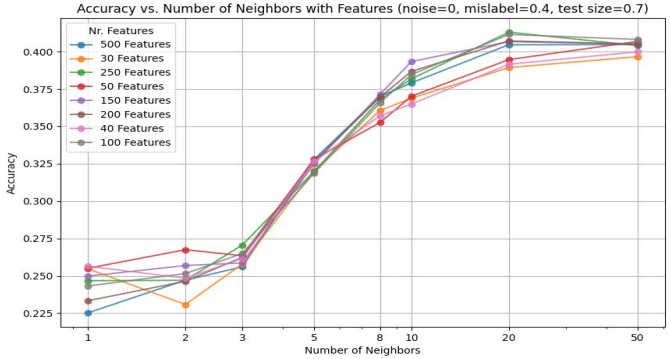
0% mislabeled



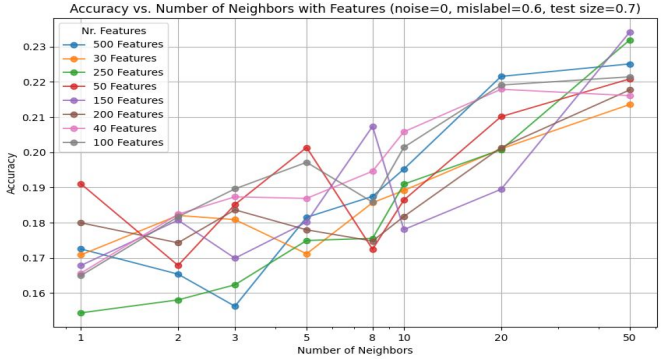
20% mislabeled



40% mislabeled

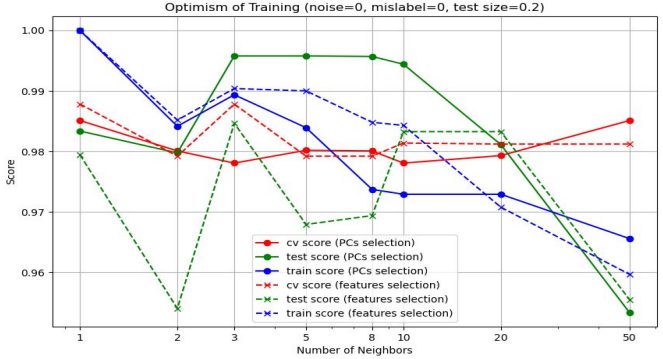


60% mislabeled

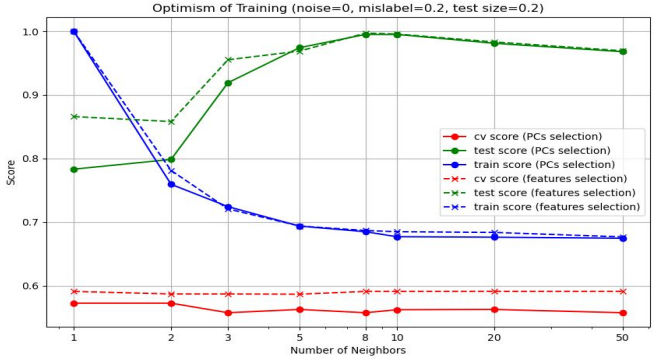


Optimism of training (80% training data)

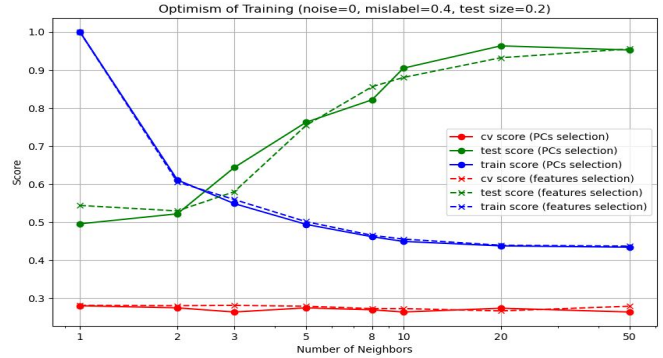
0% mislabeled



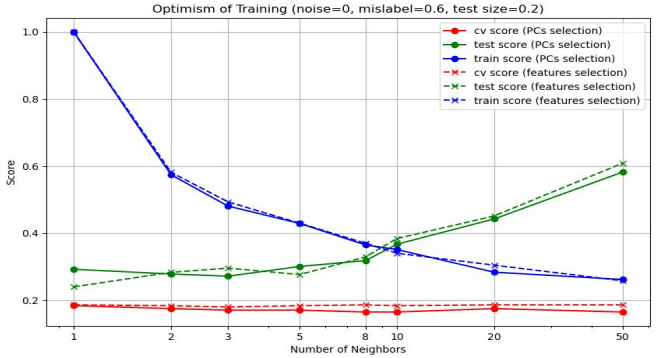
20% mislabeled



40% mislabeled



60% mislabeled



Take-home messages

Feature Selection and Principal Component Selection

From 2000 features to ~ 200 features or ~ 50 PCs

Train and Test Size

Reducing the train size negatively impacts on rigid models

Mislabeling Effect

Rigid models handle mislabeling better then flexible models

Thanks

Does anyone have any questions?

Elínborg Ásbergsdóttir
İpek Korkmaz
Luca Modica
Patrícia Marques