Distribution and Replication for Feature Selection

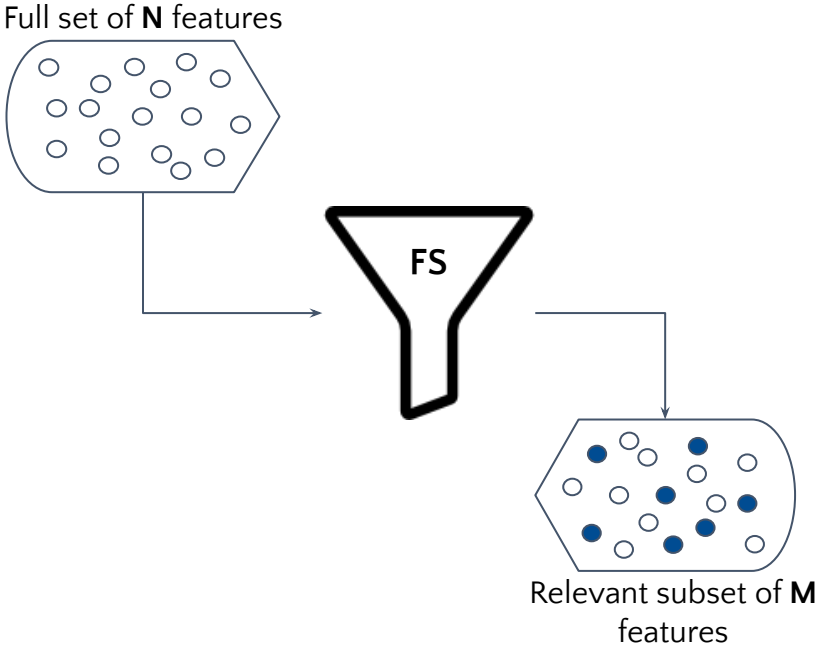
(ESR 2.2)

Uchechukwu Njoku

eBISS 2023
Barcelona, Spain
7th July, 2023

Supervisors: Alberto Abelló (UPC), Besim Bilalli (UPC), Gianluca Bontempi (ULB)

Feature selection (FS)



Why?

- Big data
- Redundancy
- Irrelevance
- Noise
- Data understanding
- Cost

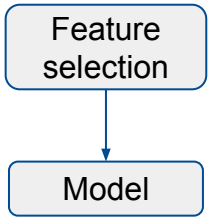
➔ M < N

Exploring the search space

- Starting point
- Search strategy
  - Exhaustive search
  - Sequential search
  - Population-based search
- Feature subsets evaluation
- Halting criterion

FS Types

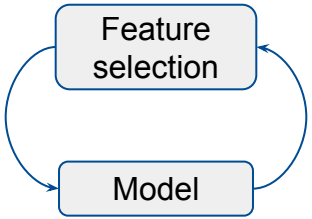
Filter methods



- Fast execution time
- Good generalization
- Robust to overfitting
- Possible redundancy
- Model independent
- Non 'optimal' selection

E.g: Mutual information

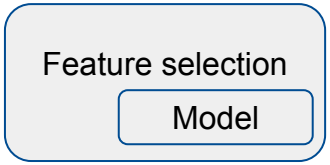
Wrapper methods



- Model dependent
- High accuracy
- Captures dependencies
- Poor generalization
- Risk of overfitting
- Computationally intense

E.g: Sequential Forward Selection

Embedded methods



- Model dependent
- Moderate execution time
- Captures dependencies
- Poor generalization

E.g: Tree based algorithms

Objectives

1. Study existing feature selection methods to:
   a. compare the several search algorithms of wrapper feature selection and existing tools
   b. evaluate their scalability, stability, and impact on performance (e.g., accuracy)
2. Propose a novel approach for **multi-criteria** wrapper feature selection
   a. find near pareto optimal set of solutions with population-based search
   b. present explanations for final selection using a dashboard user interface
3. Optimize wrapper feature selection methods by adopting frameworks for distribution, parallel computing, load partitioning, and communication methods for **scalable** feature selection

Work done

Objective 1: Study existing feature selection methods to:

- Extensive evaluation of the predictive performance and stability of **existing wrapper and filter methods** in Python libraries
- Empirical comparison of multi-objective and mono-objective wrapper FS by considering two well-known metrics, accuracy and AUC
- Analysis of the scalability and memory footprint of wrapper methods

Outcome: publications

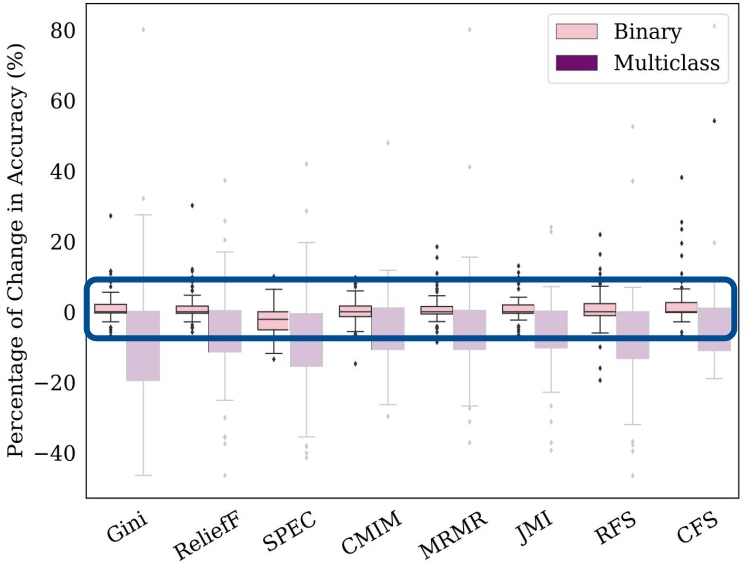
- Njoku, Uchechukwu Fortune, et al. "Impact of filter feature selection on classification: an empirical study." Proceedings of the 24rd International Workshop on DOLAP 2022
- Njoku, Uchechukwu Fortune, et al. "Wrapper methods for multi-objective feature selection." 26th International Conference on EDBT 2023

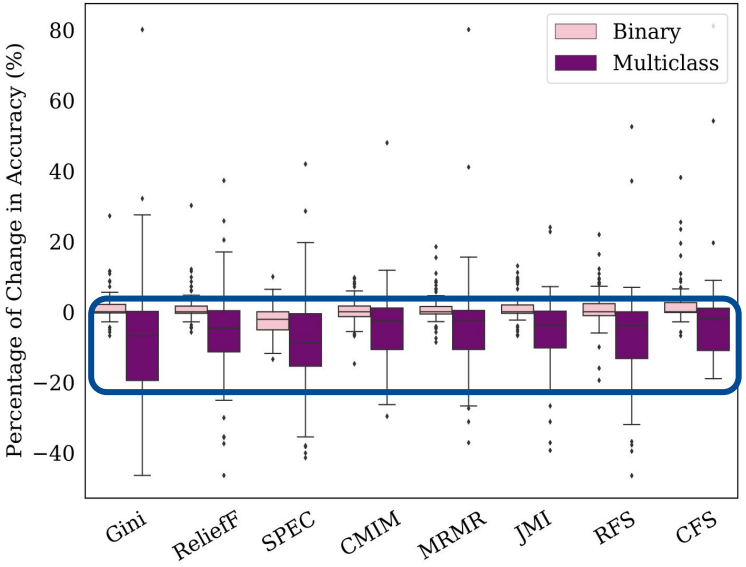
Set-up

- 32 datasets (binary and multi-class)
- Four classification algorithms: KNN, NB, DT, SVM
- Filter methods: Gini, MRMR, ReliefF, JMI, CMIM, SPEC, CFS
- Four wrapper methods: SFS_{KNN}, SFS_{NB}, SFS_{DT}, SFS_{SVM}
- Metrics: Accuracy and AUC

Results

The improvement in accuracy after feature selection is different for binary and multiclass classifications for **filter methods**





➔ Recommendation for use

Results

- Wrapper methods show superior results in predictive performance

Method	NB	KNN	DT	SVM	Average Rank
SFS	10	9	4	9	1
JMI	4	4	8	6	2
Gini	5	2	6	5	3
CMIM	3	2	4	5	4
MRMR	3	3	4	3	5
Relief	2	1	4	3	6
SPEC	0	2	1	2	7

Accuracy change in binary problems

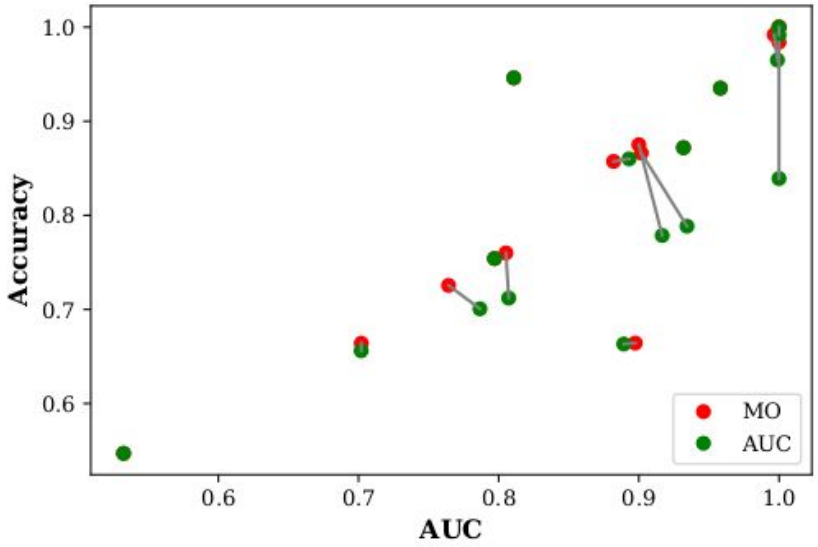
Method	NB	KNN	DT	SVM	Average Rank
SFS	5	13	7	8	1
Gini	7	4	3	1	2
MRMR	5	4	2	1	3
Relief	2	5	2	3	2
JMI	3	3	2	4	5
CMIM	3	2	2	2	6
SPEC	2	2	0	1	7

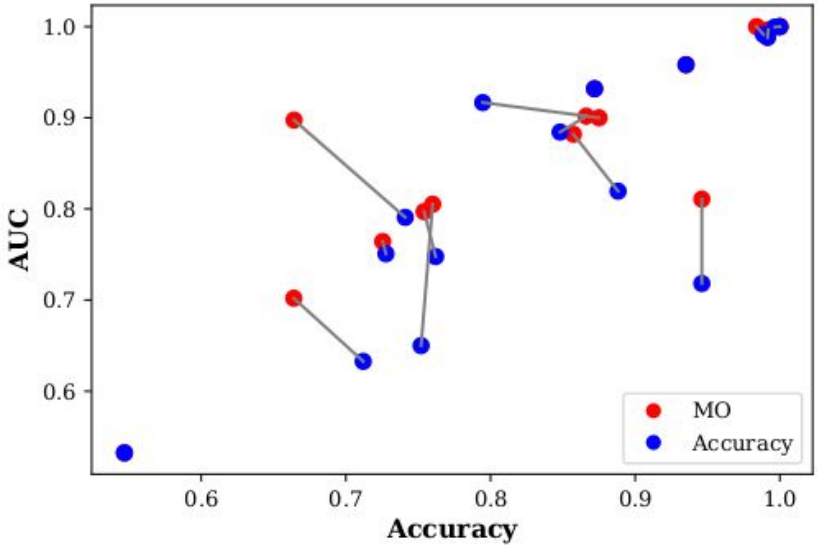
AUC change in binary problems

➔ Numbers of datasets out of 16

Results

- Wrapper method for multi-criteria feature selection using scalarization





MO-multi-objective, **ACC**-Accuracy, **AUC**- Area under the ROC Curve

Current work

Objective 2: Propose a novel approach for multi-criteria wrapper feature selection

Traditionally, multi-criteria feature selection is limited to **two objectives** - number of features and model performance

We propose the use of **more than two objectives** simultaneously for feature selection using **genetic algorithms** to find more **robust solutions** followed by the **explainability** of the trade-offs through an interactive visualization board

Multi-criteria feature selection

Criteria:

Internal evaluation criteria

- AUC
- Precision
- Accuracy
- Redundancy
- Number of features

External evaluation criteria

- Relevance
- Shapley function

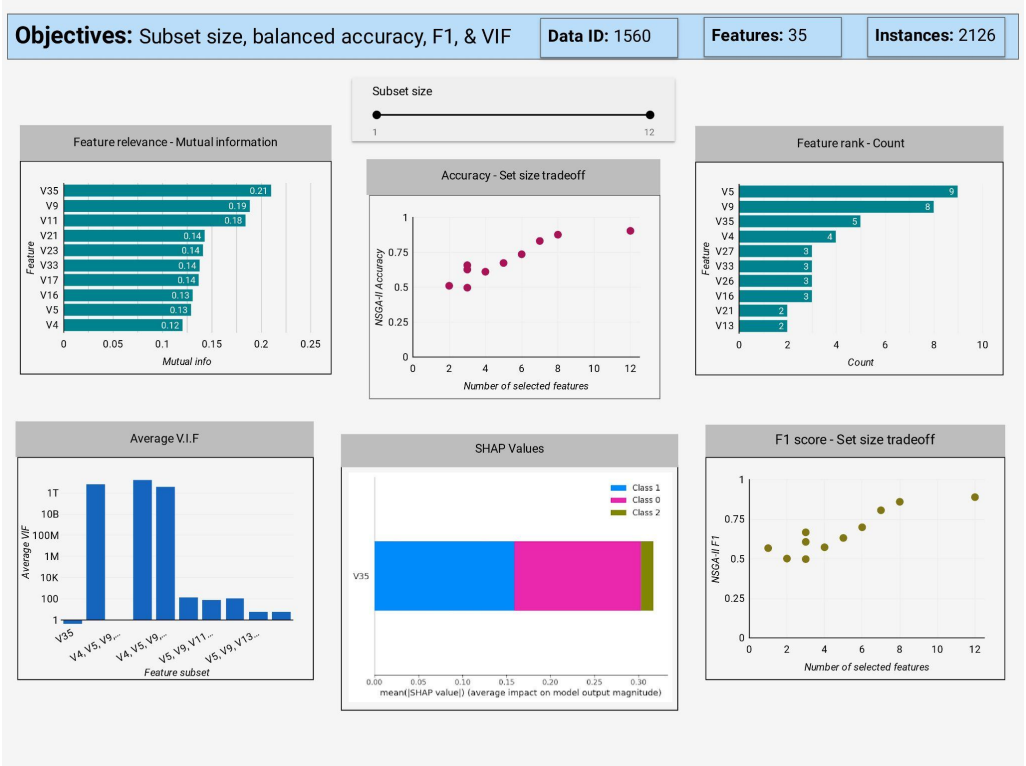
Outcome:

- Set of near-optimal feature subsets
- Interactive dashboard of results for explainability

Sample dashboard

Search method: NSGA-II
Data source: OpenML
Objectives:

- Subset size
- Balanced accuracy
- F1-score
- Variance inflation factor (VIF)



https://sites.google.com/view/explainablefs/home?authuser=1

Next steps

- Publication at Multi-Objective Decision Making Workshop
- More robust visualization
- More metrics
- Custom operations in NSGA-II for selection
- Optimization of the search process

Secondment at Orange BE

Duration: April 3rd– June 30th, 2023

Problem: From a common data source, three models were built to solve several problems. However, the preprocessing of the data before model building was lacking in evaluating the relevance of the features used.

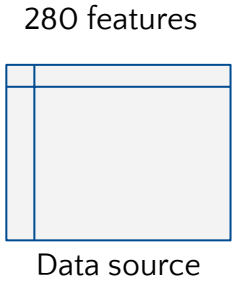
Cases:

- Use of redundant and irrelevant features
- Omission of relevant features

Solutions

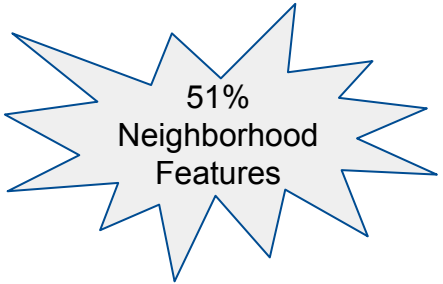
Approach

- Filter → correlation detection
- Embedded → XGBoost
- Feature Extraction (FE)→ 5-means clustering

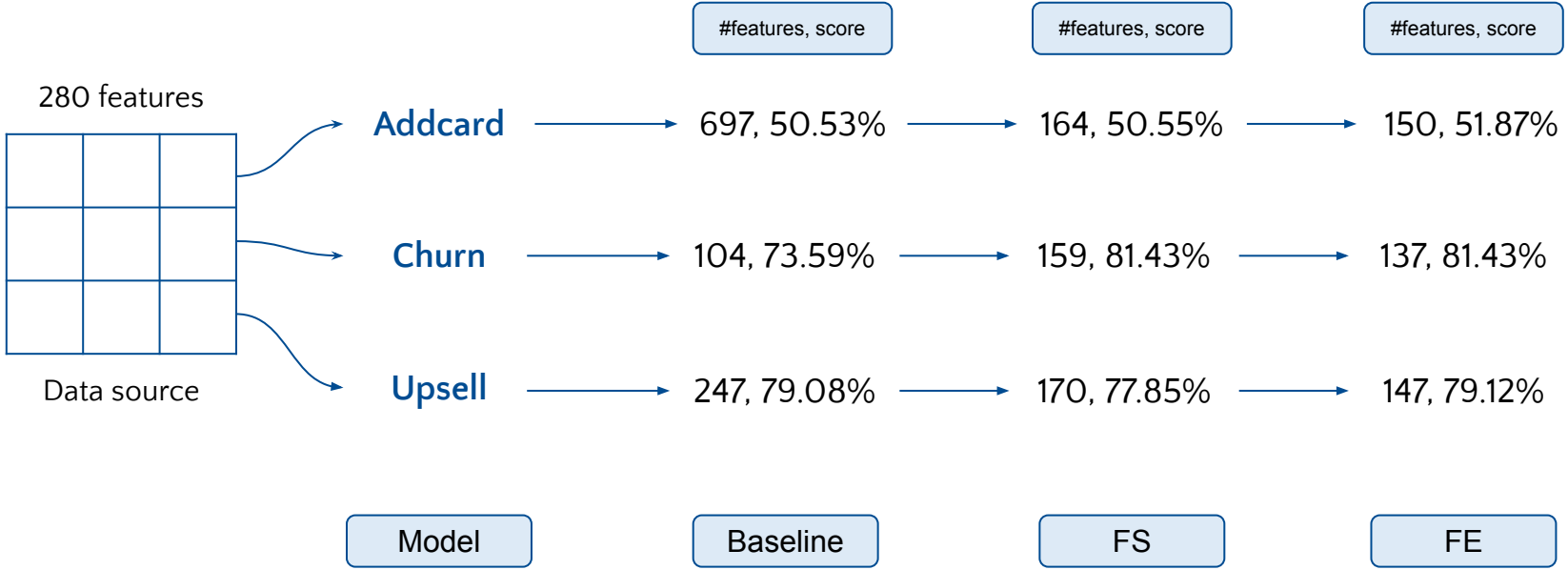


Outcome

- Over 20 pairs of redundant features
- 150 commonly relevant features
- 7 commonly irrelevant features
- Reduction from 280 to 256 features in data source

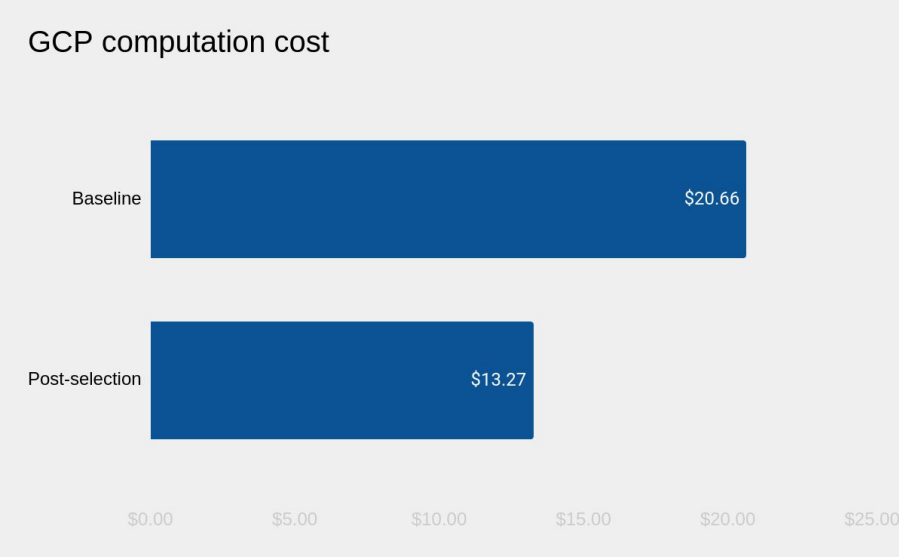


Results


#features, score
#features, score
#features, score
280 features
Addcard
697, 50.53%
164, 50.55%
150, 51.87%
Churn
104, 73.59%
159, 81.43%
137, 81.43%
Upsell
247, 79.08%
170, 77.85%
147, 79.12%
Data source
Model
Baseline
FS
FE

Cost implication

Case: Addcard model (end-to-end)

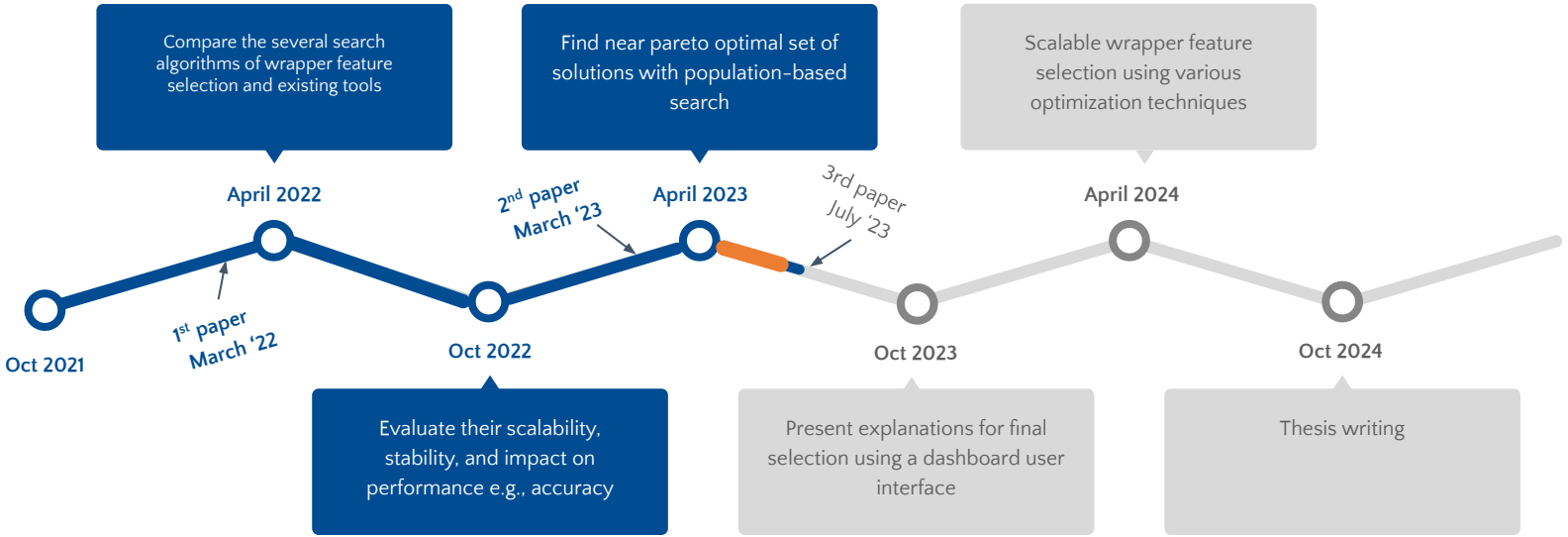


35% cheaper

Secondment outcome

- Better performance
  - Score
  - Time & cost
- Improved data understanding
- Improved explainability
- Extended collaboration

Timeline


Compare the several search algorithms of wrapper feature selection and existing tools
Find near pareto optimal set of solutions with population-based search
Scalable wrapper feature selection using various optimization techniques
April 2022
2nd paper March '23
April 2023
3rd paper July '23
April 2024
Oct 2021
1st paper March '22
Oct 2022
Oct 2023
Oct 2024
Evaluate their scalability, stability, and impact on performance e.g., accuracy
Present explanations for final selection using a dashboard user interface
Thesis writing

THANK YOU

Q & A