Maintenance benchmark on real world data 2

Esteban Marquer

6 november

Benchmark experiment setup

Time complexity

Results on each dataset

What next

Baselines Green: maintenance baseline avail. Red: not found

Dataset Name	Cases	Attributes	Classes	Accuracy (%)
Balance	625	4	3	85.12
Breast Cancer Diagnostic	569	30	2	96.90
Breast Cancer Prognostic	198	33	2	71.58
Breathalyser	127	5	2	71.60
Credit Approval	690	15	2	86.92
Dermatology	366	34	6	97.75
Glass Identification	214	9	7	69.05
Haberman's Survival	306	3	2	69.51
Heart Disease Cleveland	303	14	5	53.22
Hepatitis	155	19	2	80.63
Ionosphere	351	33	2	86.71
Iris	150	4	3	97.00
Lenses	24	4	3	72.50
Liver Disorders	345	6	2	64.20
Lung Cancer	32	56	3	48.00
Pima Indians Diabetes	768	8	2	70.78
Post-Operative Patient	90	8	3	64.71
Spam1	1000	699	2	93.35
Spam2	1000	699	2	94.3
Spam3	1000	699	2	98.25
Spam4	1000	699	2	97.05
Spam5	1000	699	2	94.8
Teaching Assistant Evaluation	151	5	3	55.33
Wine	178	13	3	96.67
Zoo	101	16	7	91.50
Average over twenty-five datasets	-	-	-	80.30

Current datasets

verbose name	found	has maintenance baseline
Lenses	True	**True**
Credit Approval	True	**True**
Zoo	True	False
Wine	True	False
Teaching Assistant Evaluation	True	False
Post-Operative Patient	True	False
Pima indians Diabetes	True	False
Lung Cancer	True	False
Liver Disorders	True	False
Iris	True	False
Ionosphere	True	False
Hepatitis	True	False
Heart Disease Cleveland	True	False
Haberman's Survival	True	False
Glass Identification	True	False
Dermatology	True	False
Breast Cancer Pronostic	True	False
Breast Cancer Diagnostic	True	False
Balance	True	False

Process

train/dev/test split: 60%/20%/20%, as in the thesis

train: cases in the CB at the start
dev: cases used for maintenance decision
test: cases used to mesure performance

Similarity

Numeric attribute: 1− normalized absolute distance

$$sim(x, y) = 1 - \frac{x - y}{max(X) - min(X)}$$

with x, y the value of the att. for 2 cases,
X the values of the att. for all cases in the CB

Symbolic attribute:

$$sim(x, y) = 1 \text{ if } x = y \text{ else } 0$$

with x, y the value of the att. for 2 cases

Overall similarity: weighted similarity (Karabulut *et al.*, 2019)[1]

[1] *"Weighted Similarity Measure for k-Nearest Neighbors Algorithm"* B. Karabulut, G. Arslan, H. M. Ünver, 2019, Celal Bayar University Journal of Science

Weight computation

$C_i(a)$: set of values for attribute a belonging to class i

$$C_i(a) = \{X[k][a] : X[k] \in X \text{ and } y[k] = i\}$$

$A_i(a)$: set of cases with attribute a within values of class i

$$A_i(a) = \{X[k] \in X : min(C_i(a)) \leq X[k][a] \leq max(C_i(a))\}$$

$$A_i(a) = \{X[k] \in X : X[k][a] \in C_i(a)\} \text{ for nominal att. (defined by me)}$$

$B_i(a)$: set of cases with attribute a within values of class i but not any other class

$$B_i(a) = A_i(a) - \cup_{i \neq j, j \in classes} A_j(a)$$

w_a: weight for attribute a, average "ability to discriminate"

$$w_a = |\cup_{i \in classes} B_i(a)|/n, \; n : len(X)$$

$w_a* = w_a/(\sum_{a'} w_{a'})$: normalized w_a

(in paper it was $w_a = (\cup_{i \in classes} |B_i(a)|)/n$, but it makes no sense)

Models

- MeATCube
- 1-NN, 5-NN, 10-NN, all-NN
  outcome obtained by voting, vote weight inversely proportional to similarity of the case with the target

Processing

Processing (MeATCube):

1. find **weights** on the whole dataset for each feature in the similarity
2. (repeat) compress MeATCube using hinge competence
   - compute MeATCube prediction performance
   - compute 1-NN, 5-NN, 10-NN and all-NN performance

Theoretical time complexity

Time complexity of MeATCube prediction: $\Theta_{pred} = \Theta(|\mathcal{R}|\,|CB|^2)$
With $\mathcal{R}$ the set of possible outcomes

Uselessly detailed value: $||\mathcal{R}|\,|CB|(3|CB| + 2)$

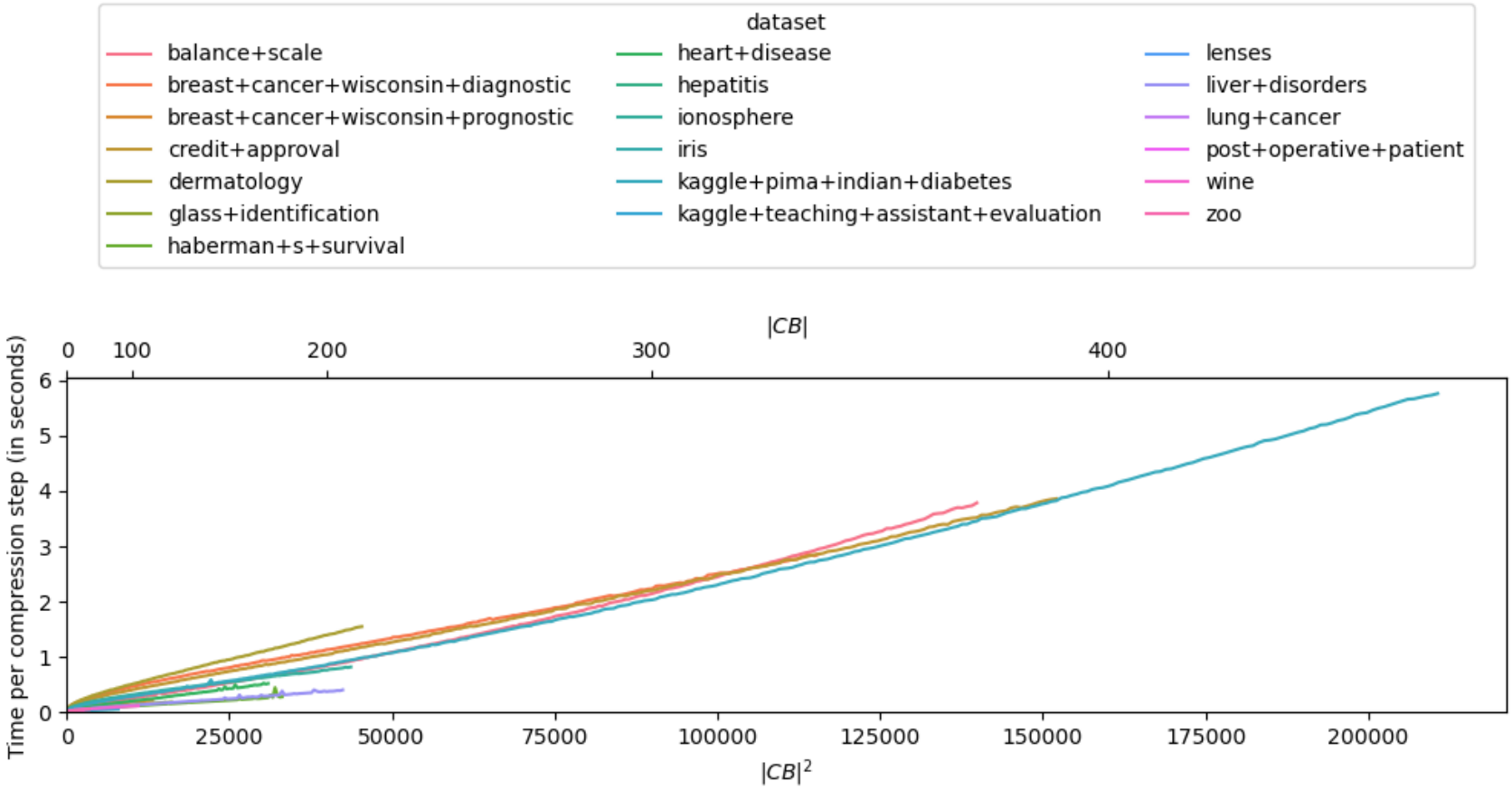
Time complexity of competence of 1 case with MeATCube: $\Theta_{\text{case comp.}} = \Theta(|CB_{ref}|\Theta_{pred})$ With CB_{ref} the case base on which we compute competence

For all cases at once: $\Theta_{\text{cases comp.}} = \Theta(|CB|\,|CB_{ref}|\Theta_{pred})$

Time complexity of 1 MeATCube compression iteration:

$\Theta_{\text{cases comp.}} = \Theta(|CB|^3\,|CB_{ref}|\,|\mathcal{R}|)$

Of which at least $O(|CB|)$ is strongly CPU bound (cannot be fully GPU accelerated)


dataset
balance+scale
breast+cancer+wisconsin+diagnostic
breast+cancer+wisconsin+prognostic
credit+approval
dermatology
glass+identification
haberman+s+survival
heart+disease
hepatitis
ionosphere
iris
kaggle+pima+indian+diabetes
kaggle+teaching+assistant+evaluation
lenses
liver+disorders
lung+cancer
post+operative+patient
wine
zoo
|CB|
0
100
200
300
400
Time per compression step (in seconds)
0
1
2
3
4
5
6
0
25000
50000
75000
100000
125000
150000
175000
200000
$|CB|^2$

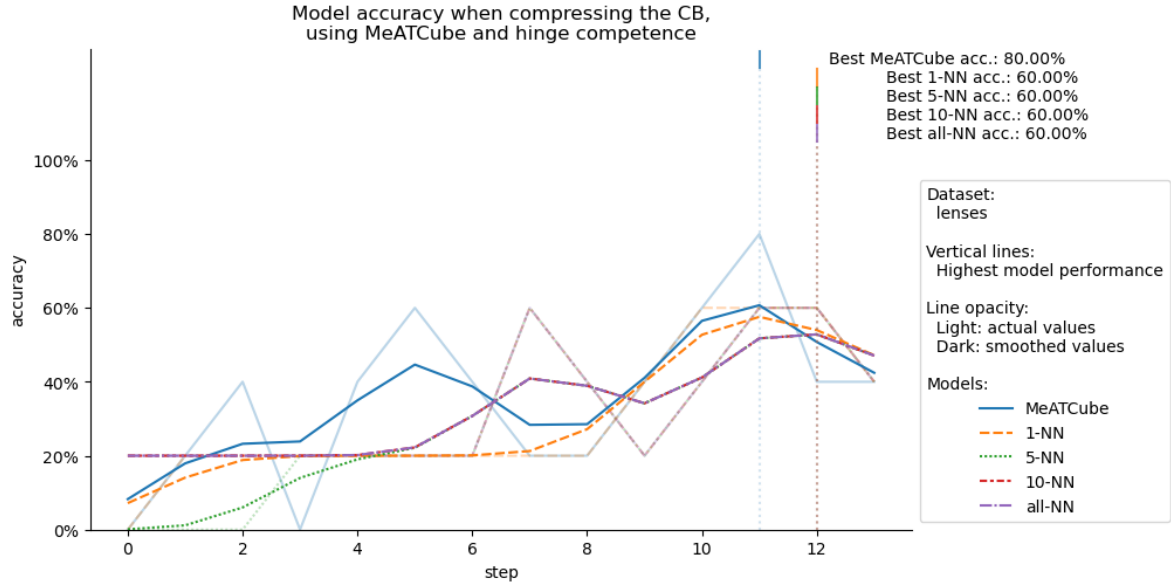
Metrics

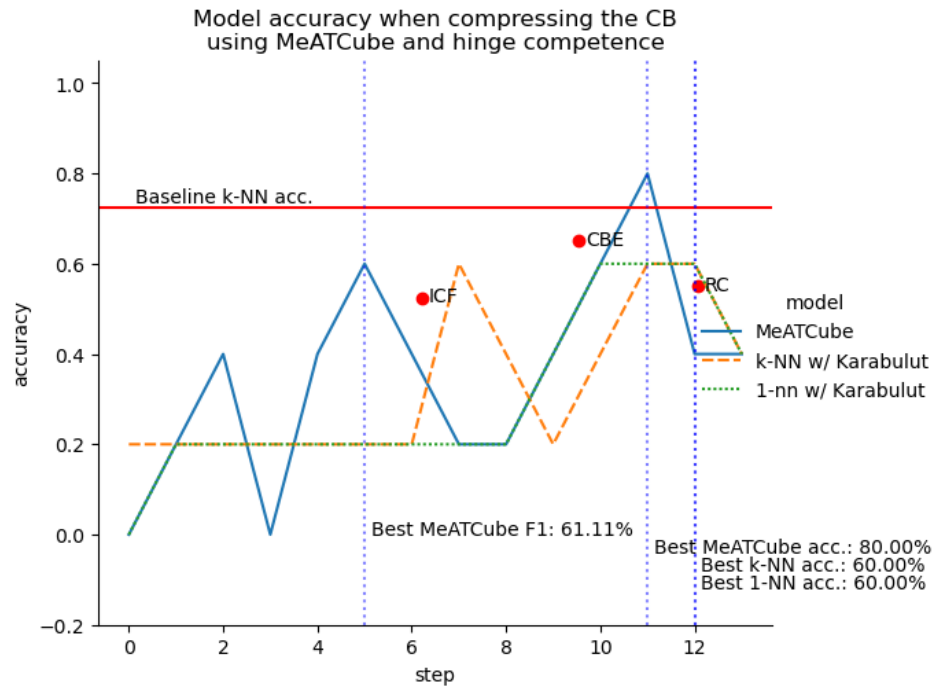
Compression step: $step_i : |CB_i| = |CB_0| - i$

Position and value for the maximum accuracy for all models
if multiple with same score, take the one with the smallest CB

Result: Lenses

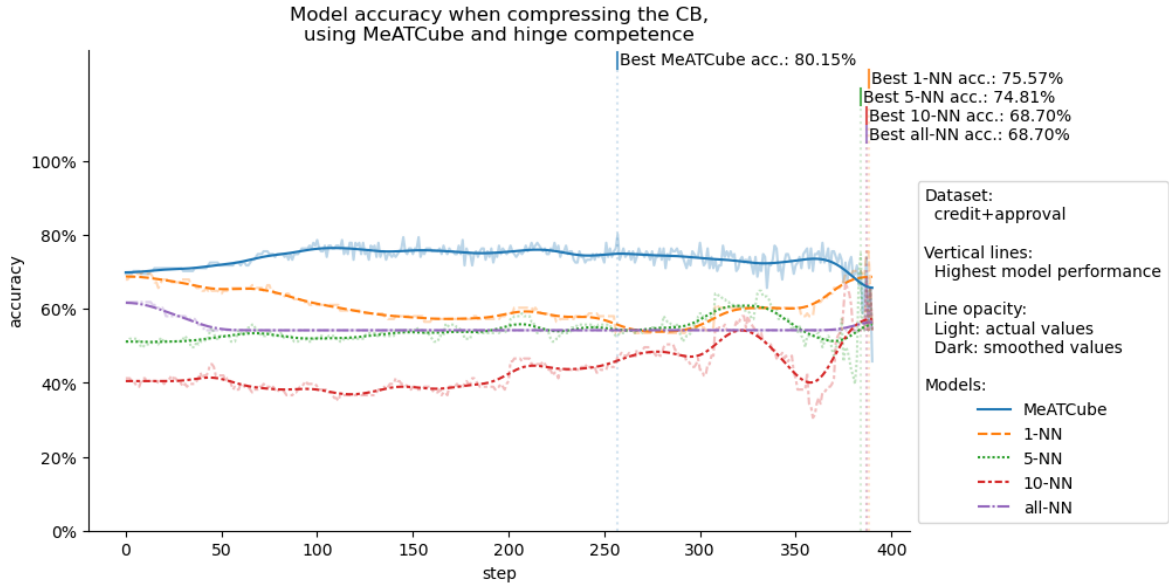
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

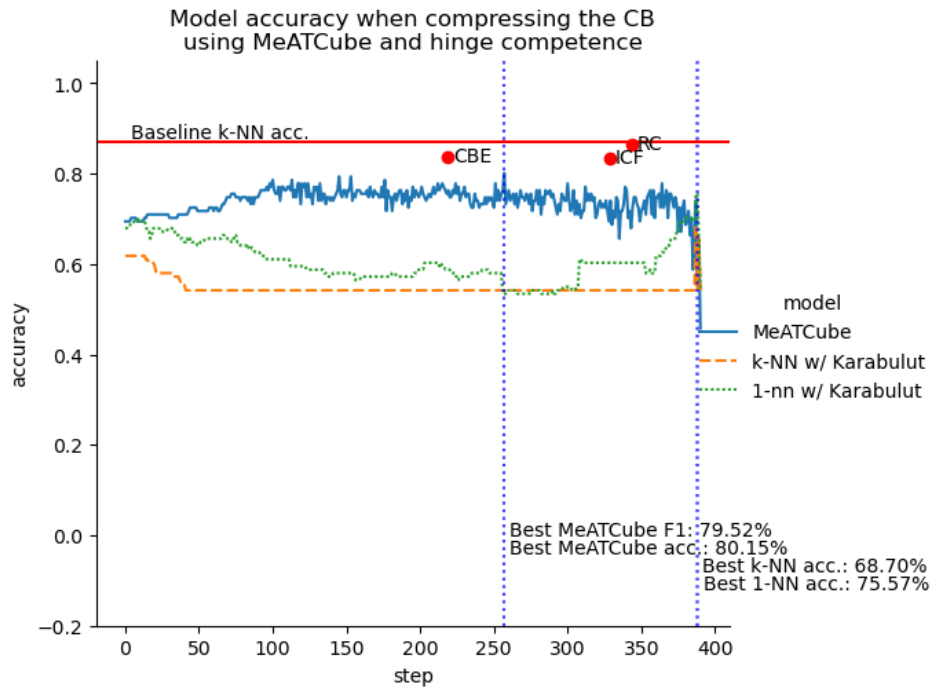




Result: Credit Aproval

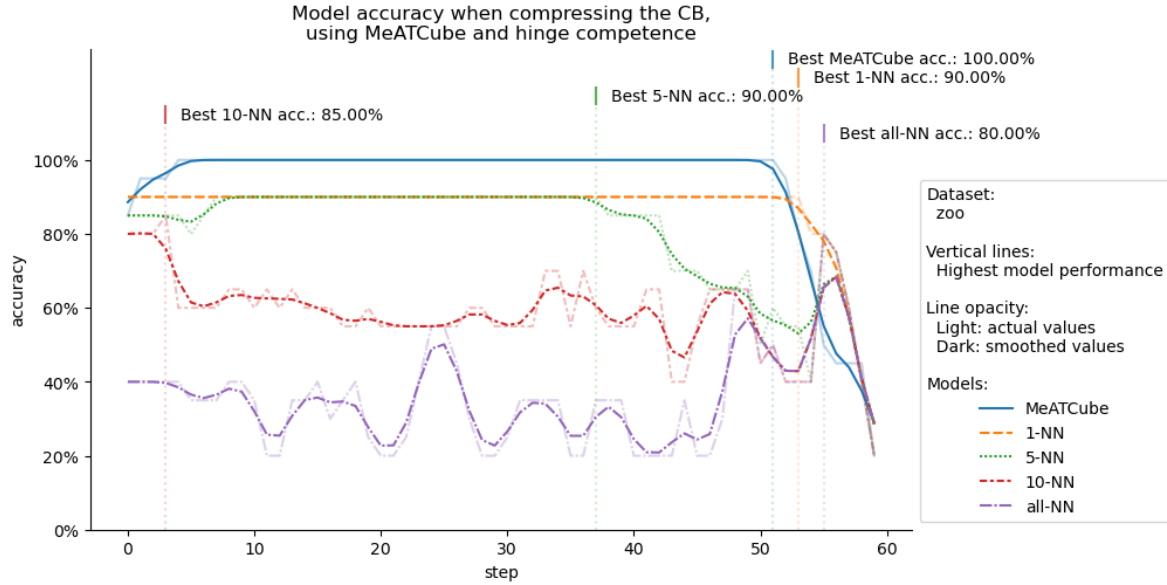
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

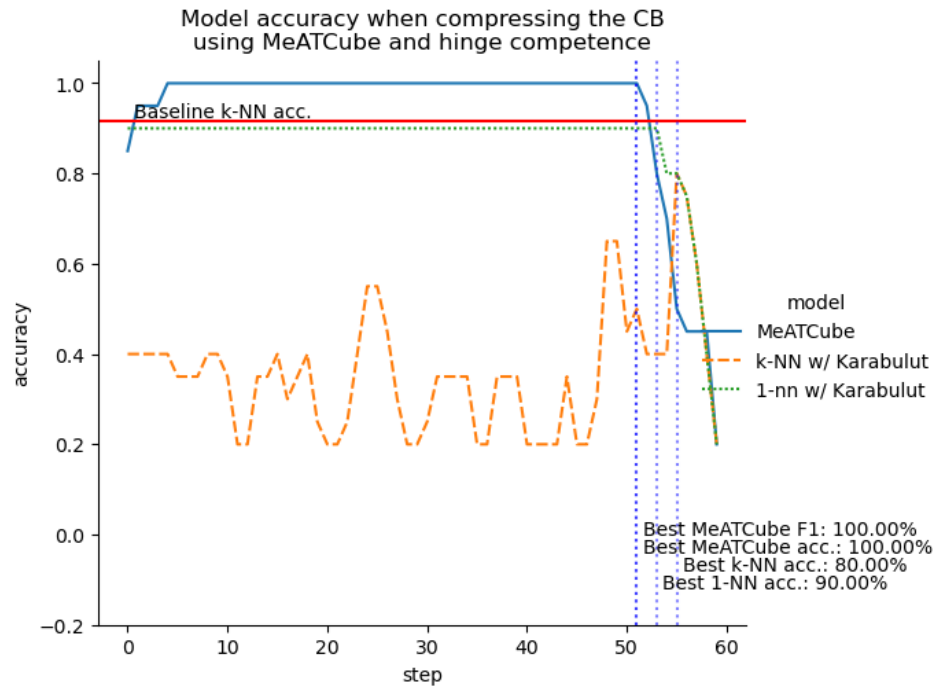




Result: Zoo

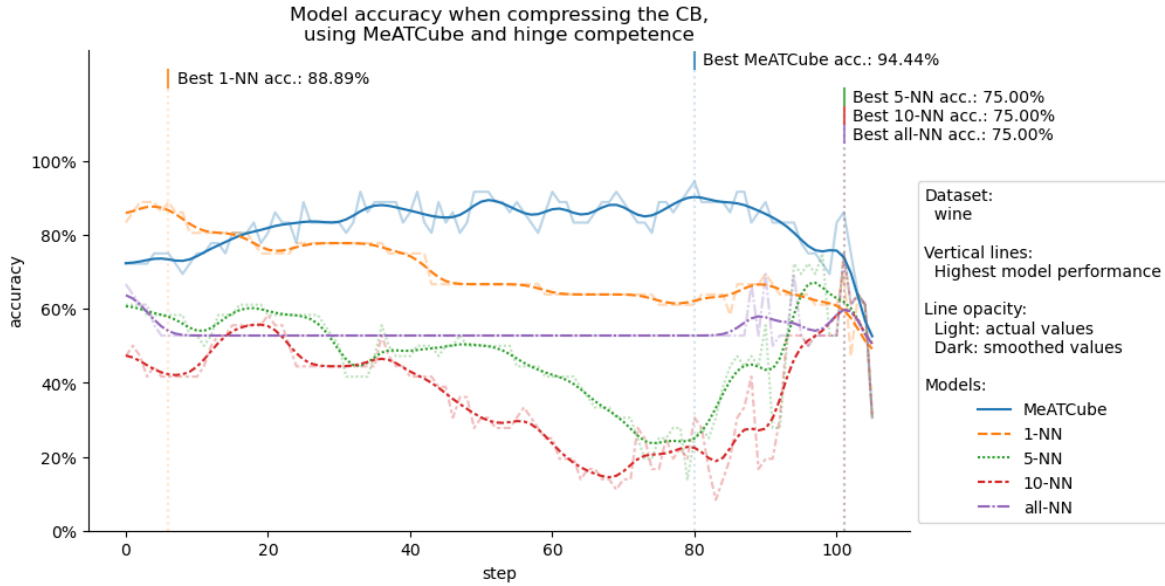
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

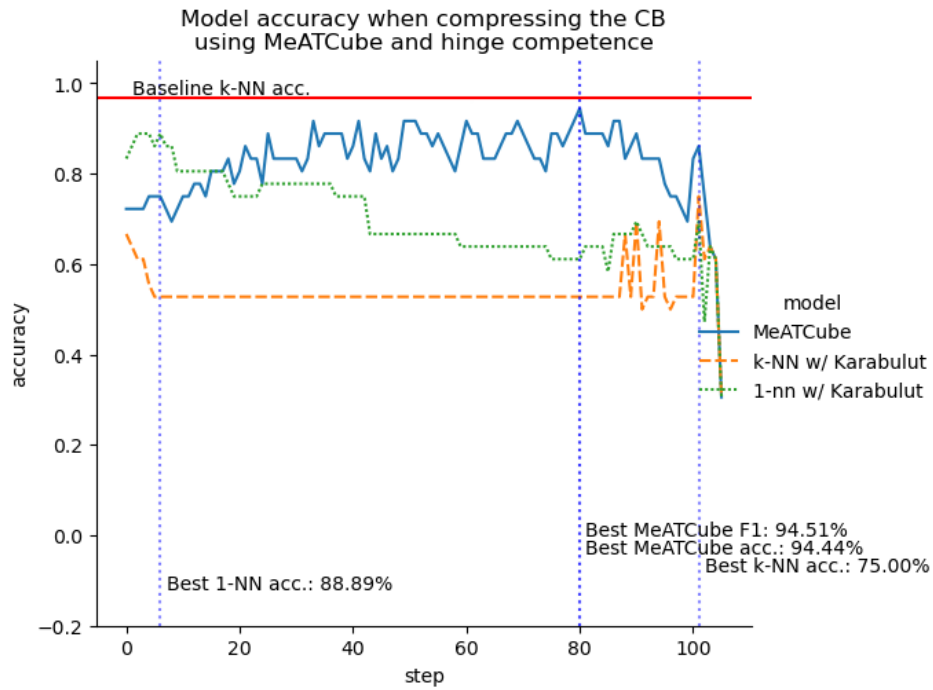




Result: Wine

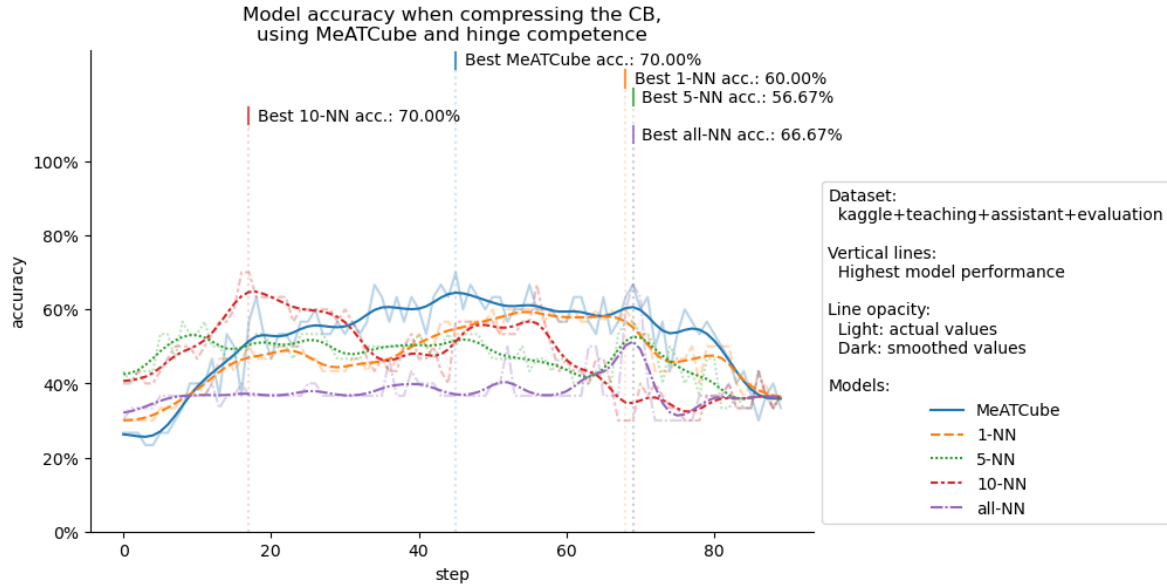
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

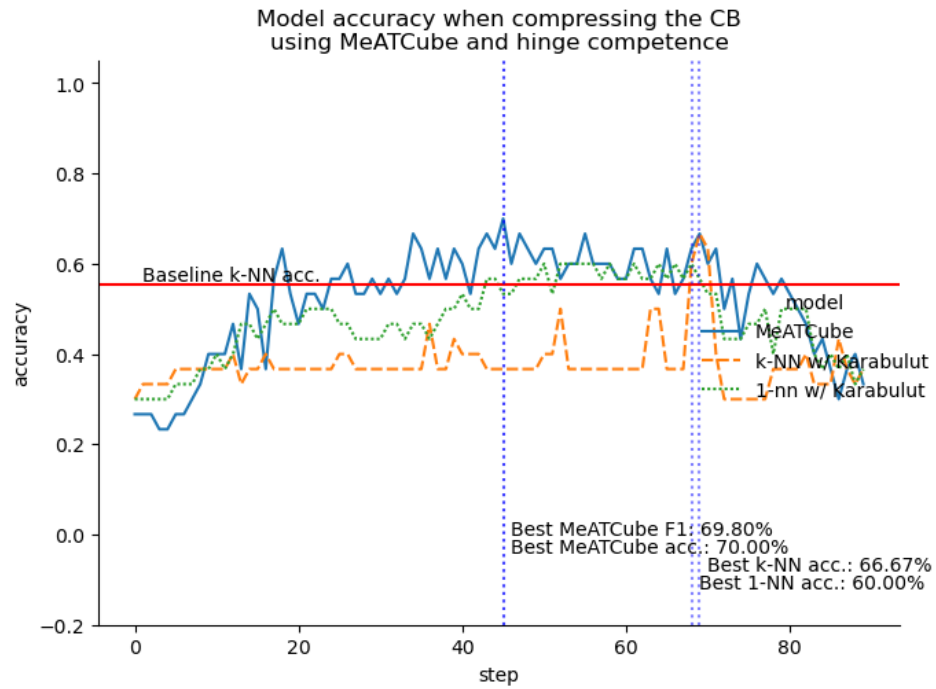




Result: Teach. Aassistant

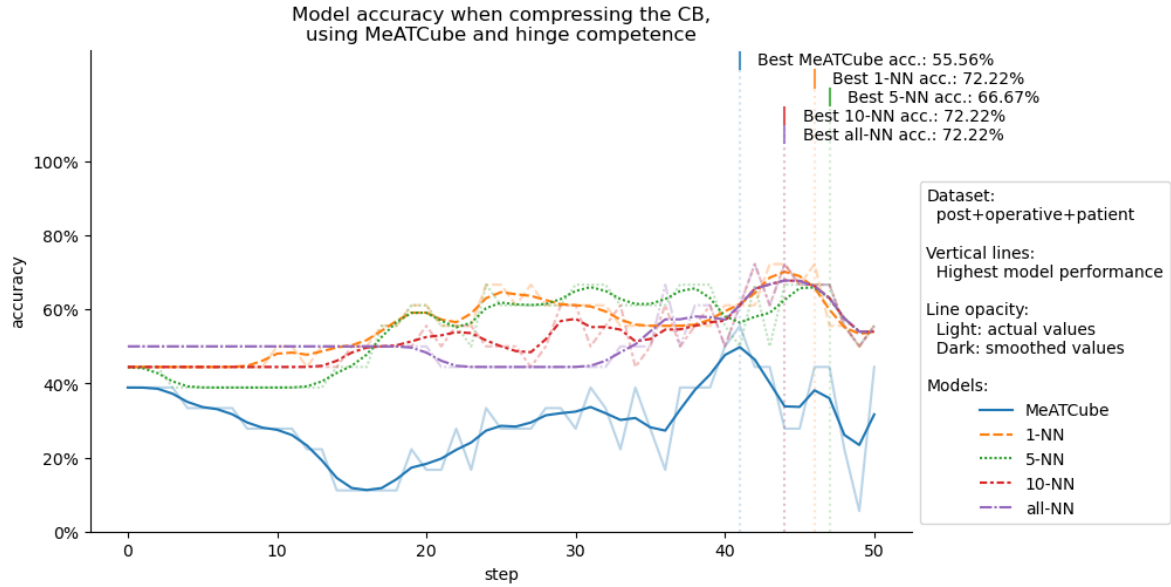
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

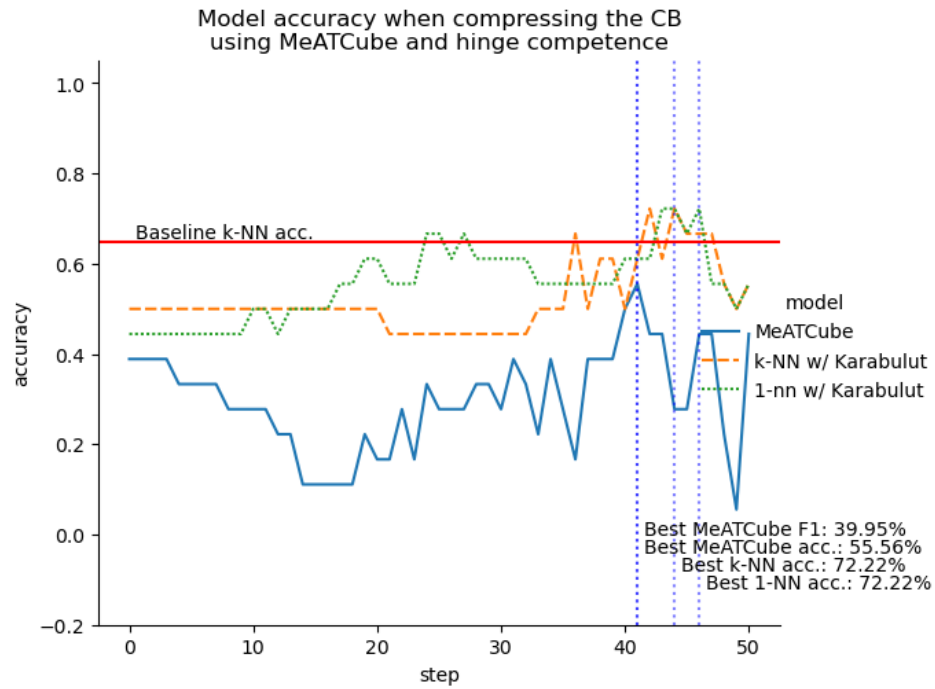




Result: Post-Op. Patient

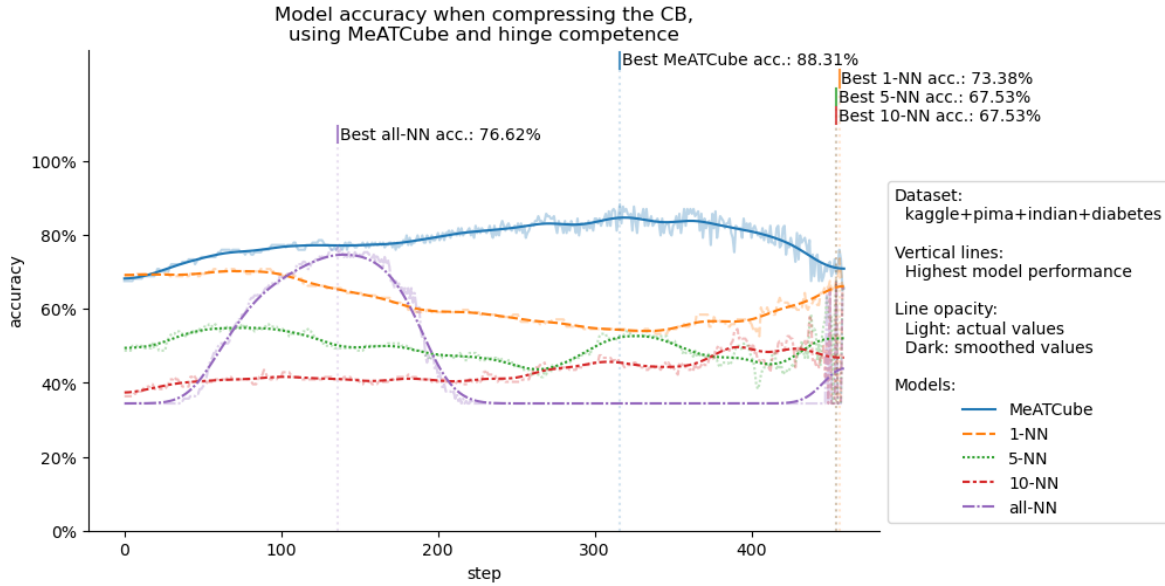
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

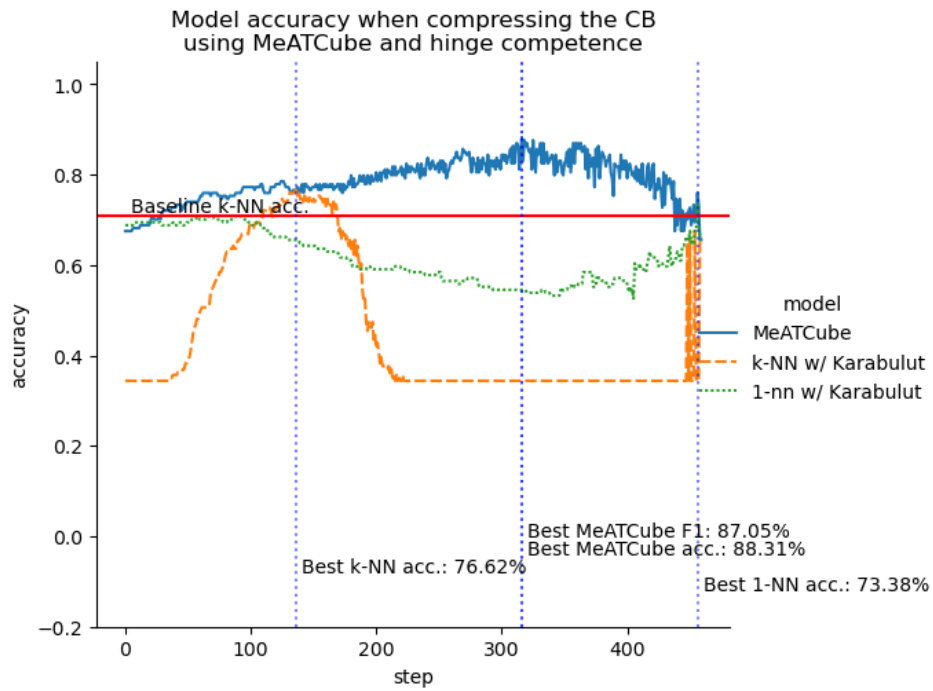




Result: Pima

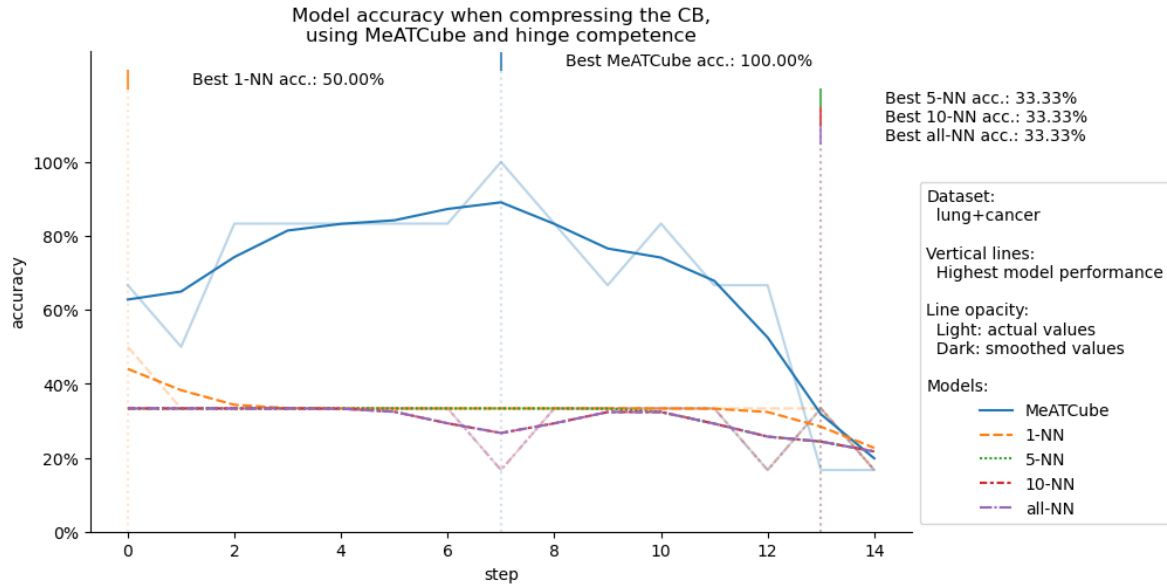
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

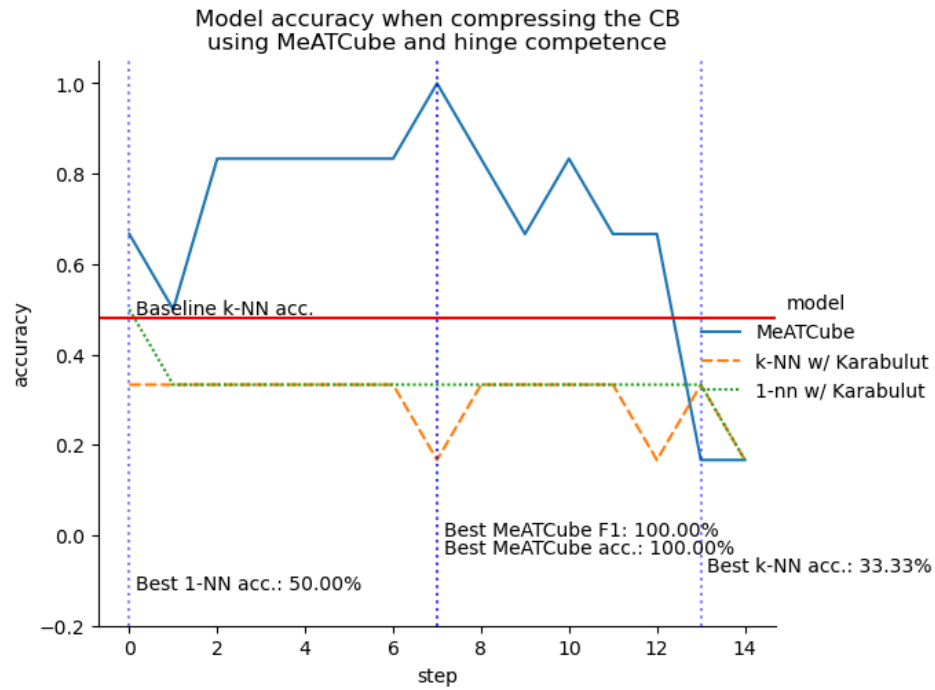




Result: Lung Cancer

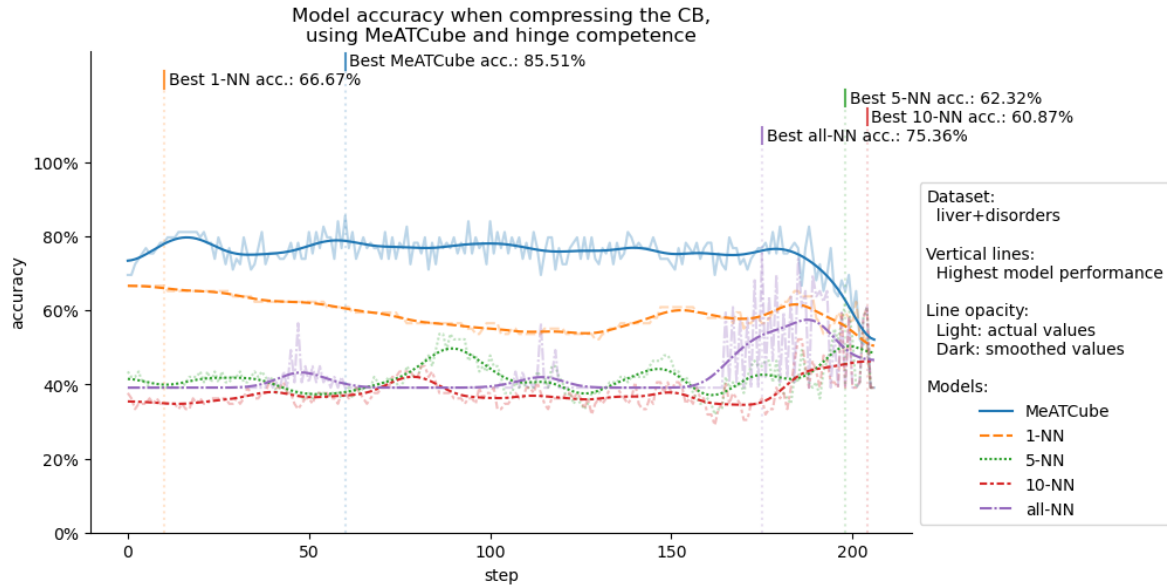
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

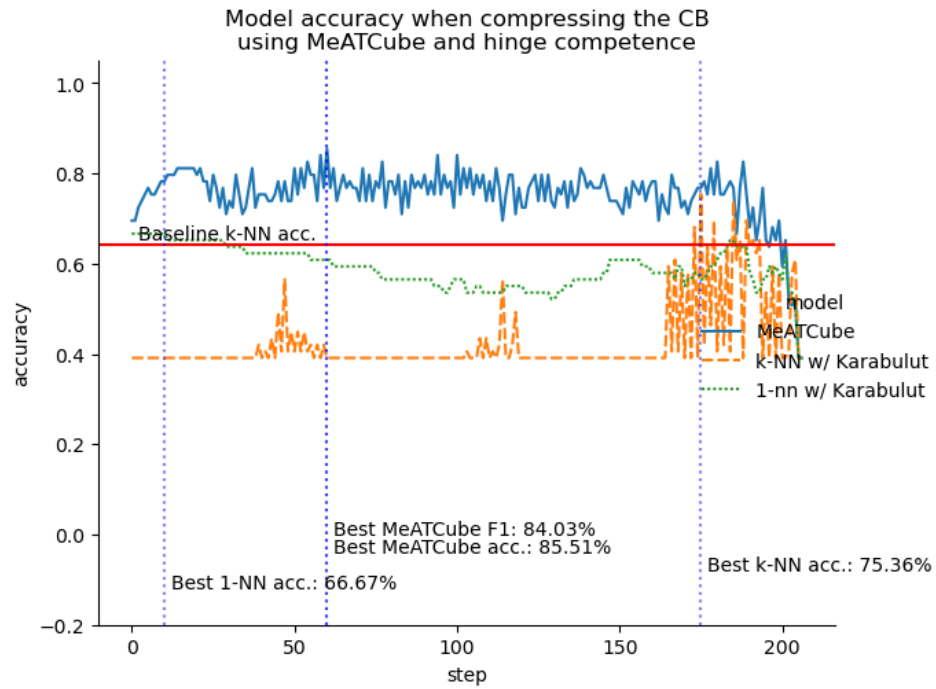




Result: Liver Disorders

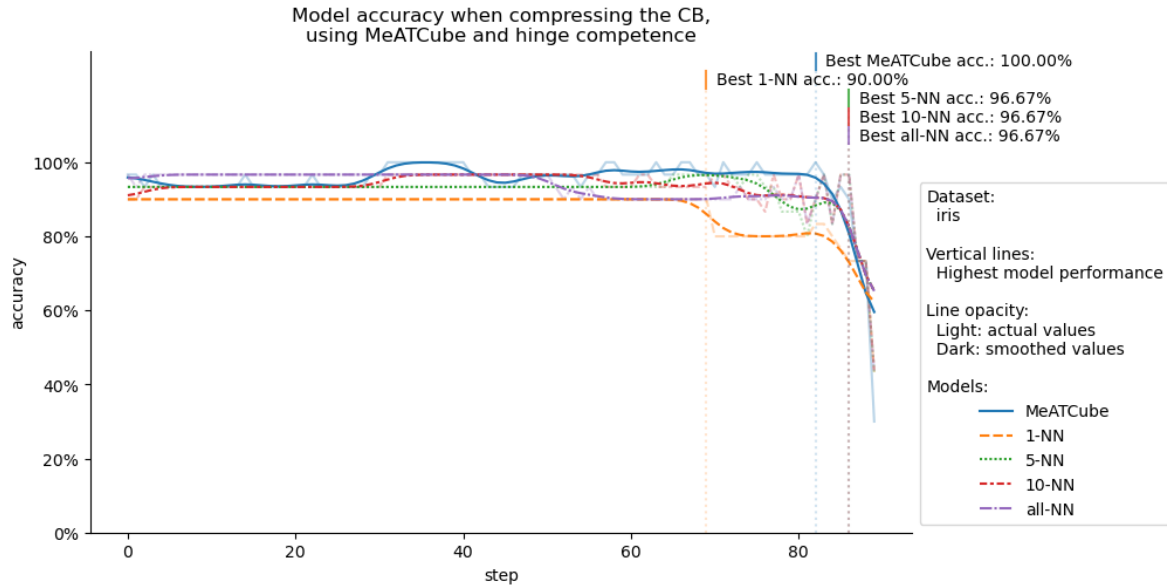
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

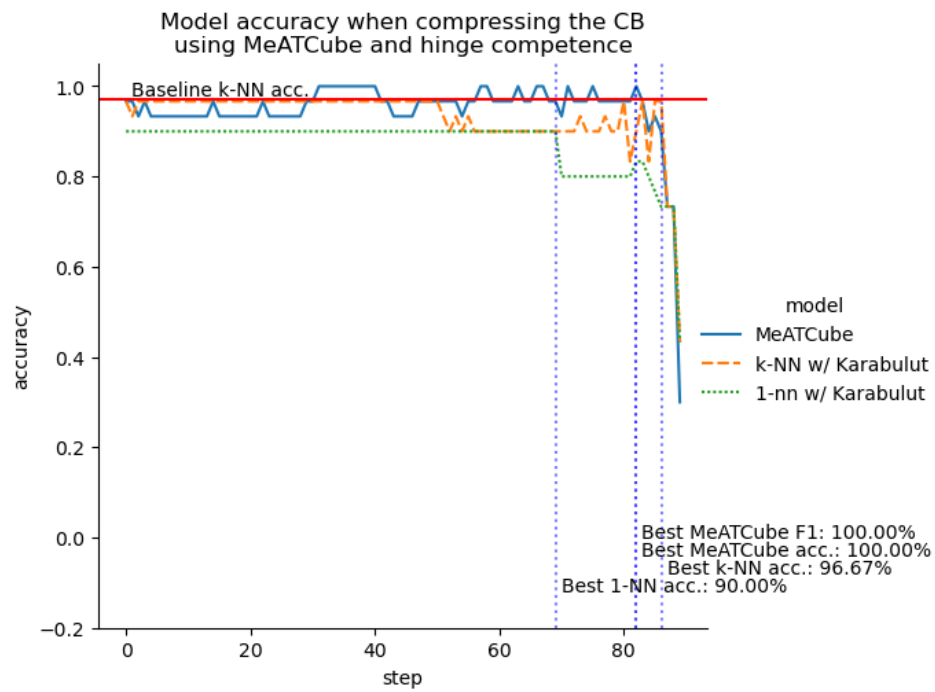




Result: Iris

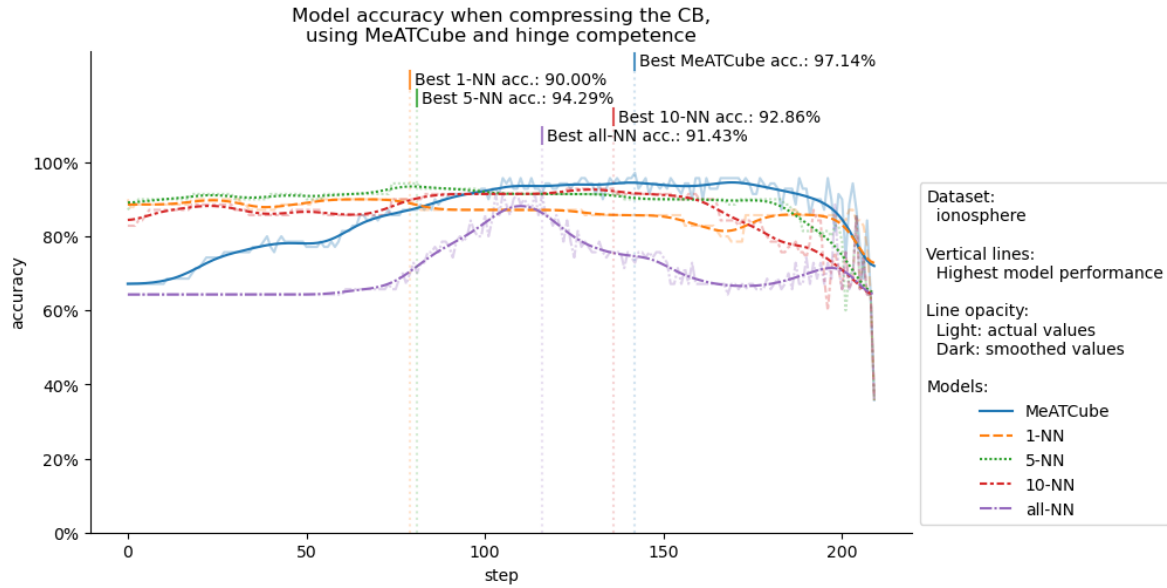
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

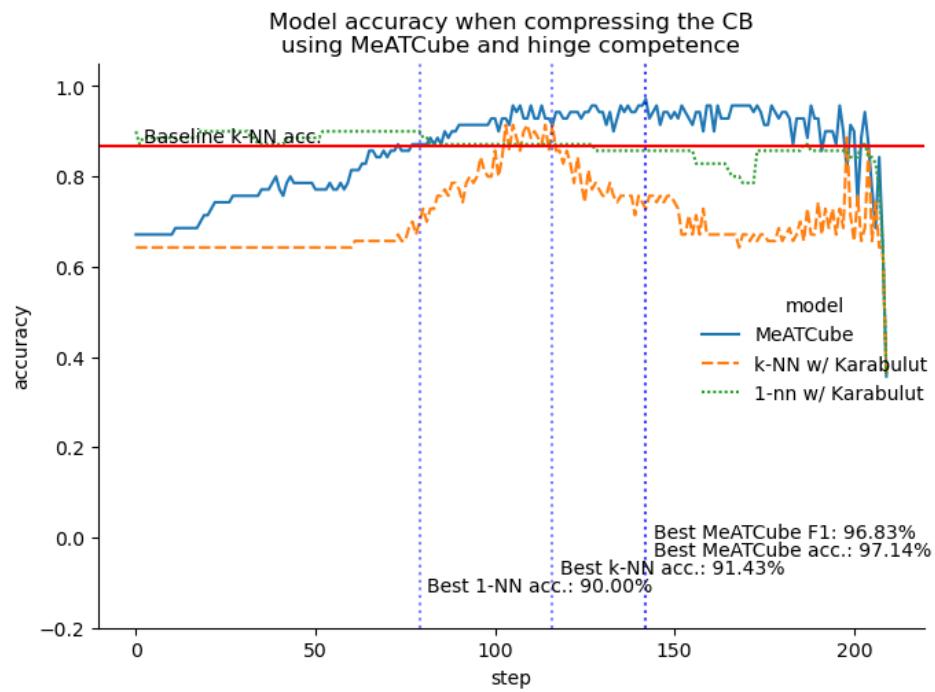




Result: Ionosphere

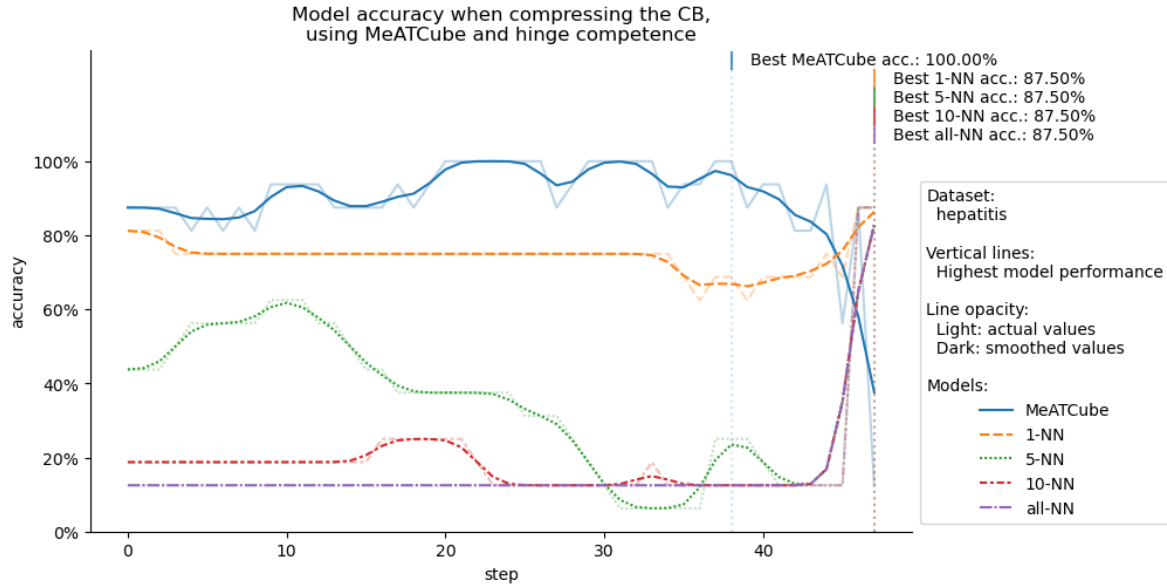
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

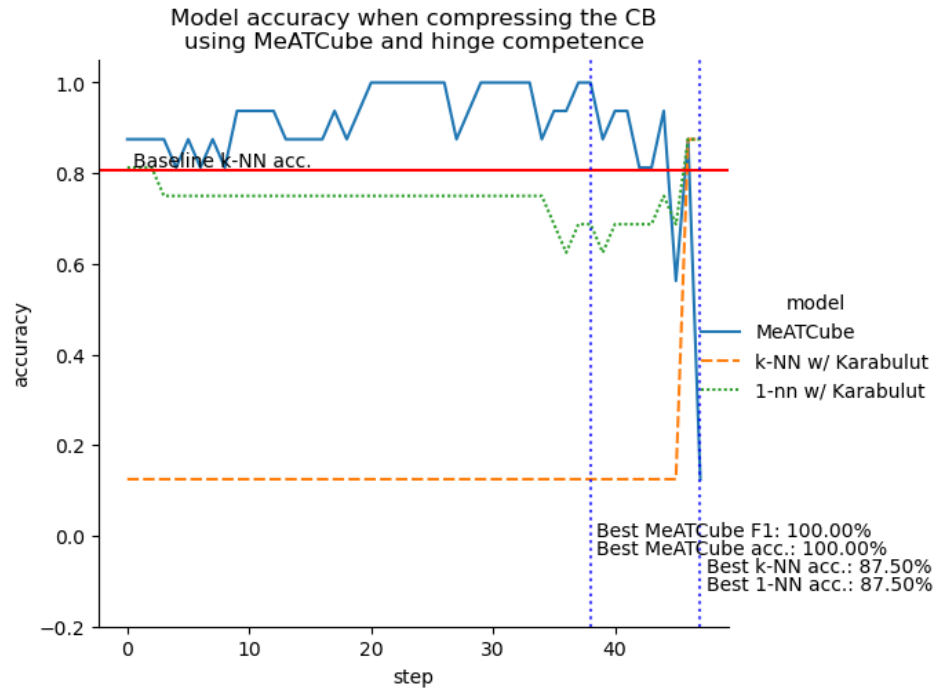




Result: Hepatitis

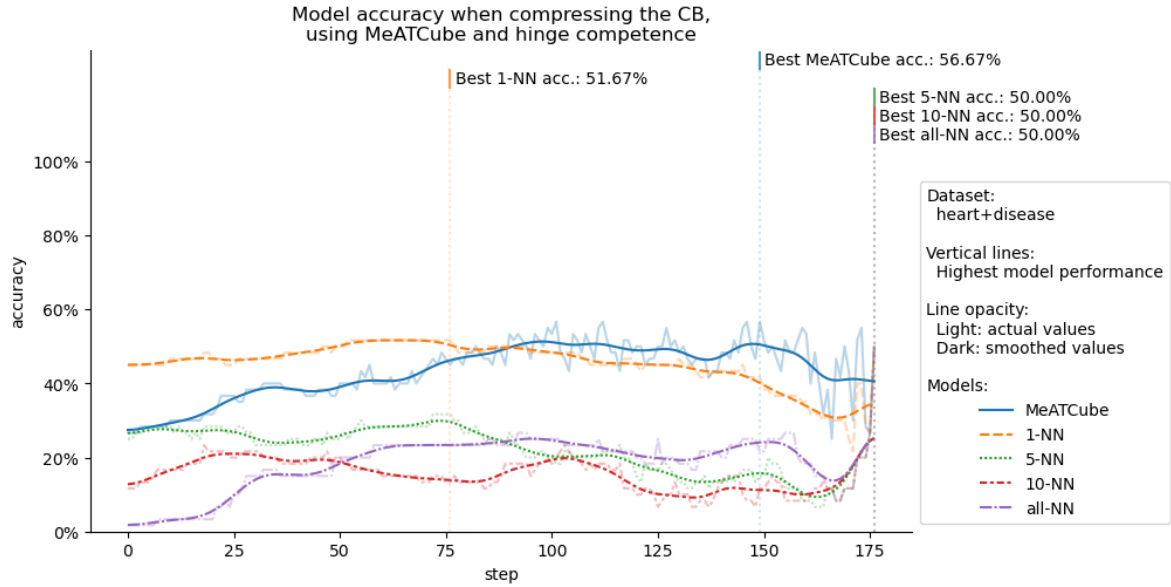
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

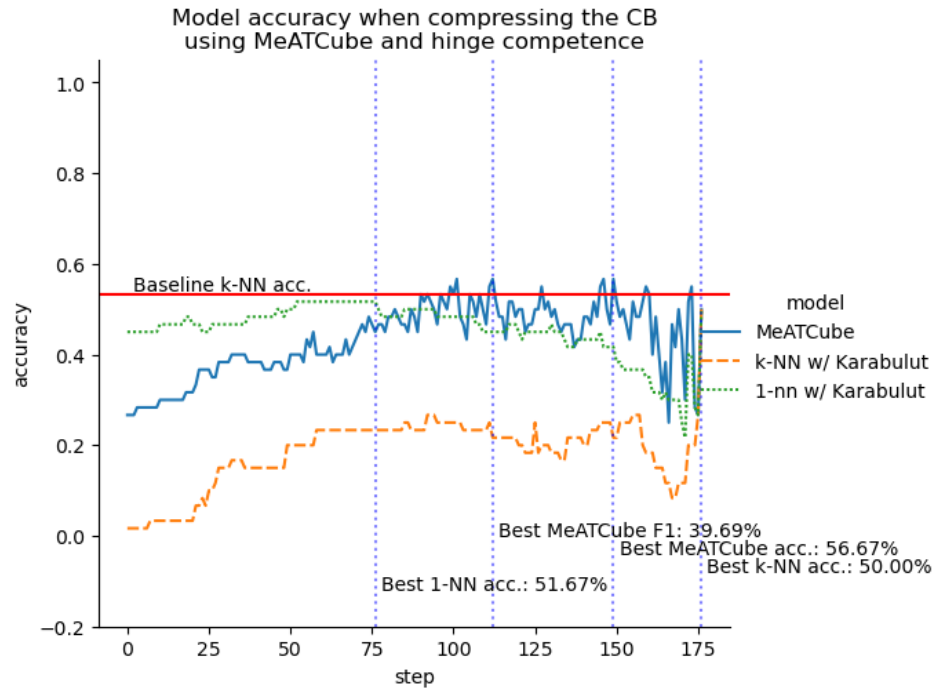




Result: Heart

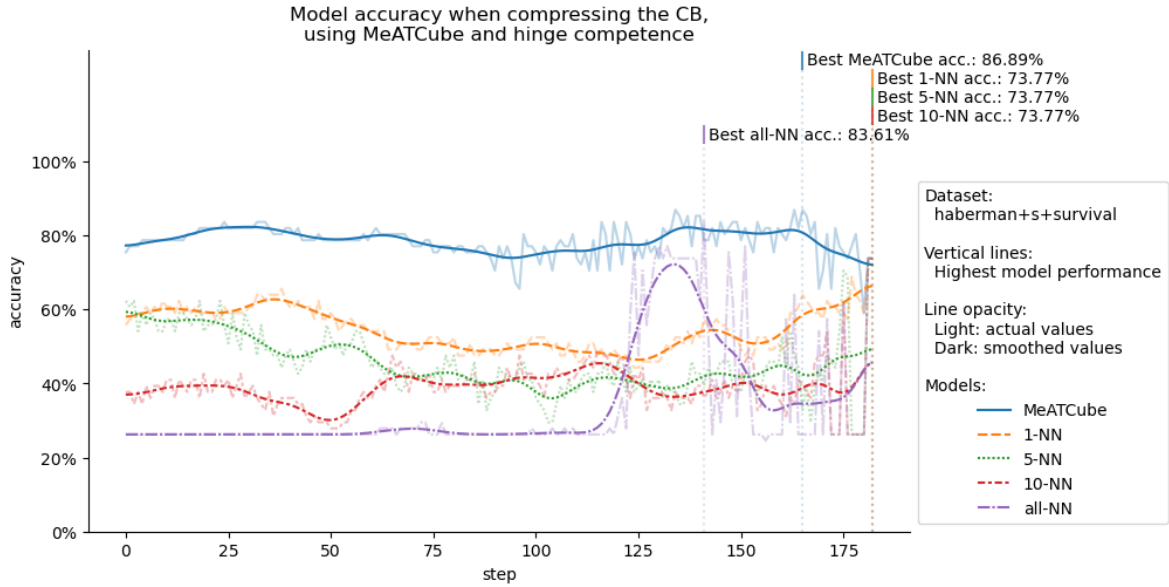
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

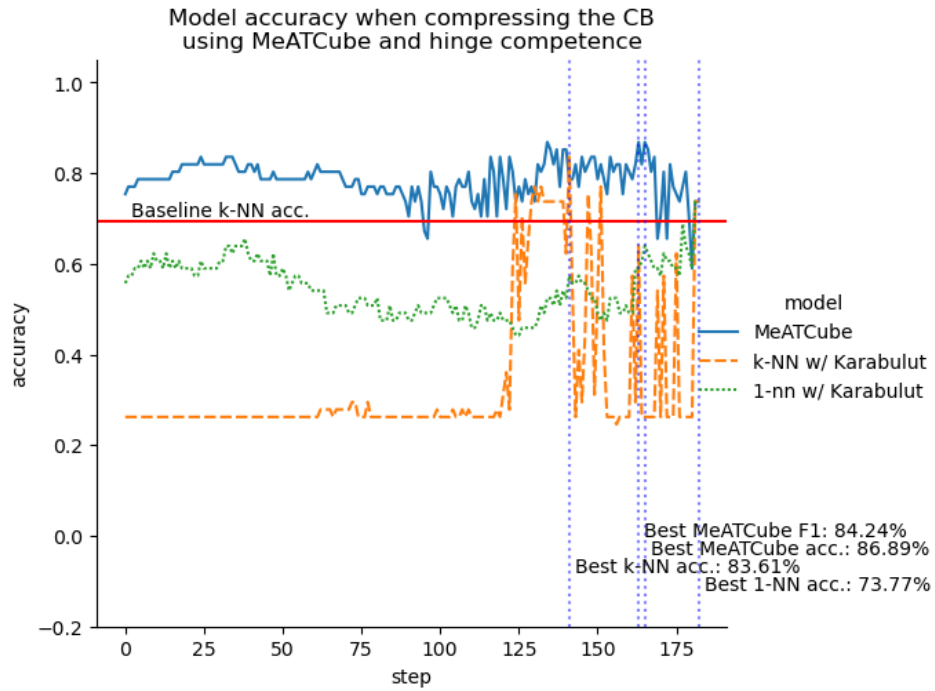




Result: Haberman's

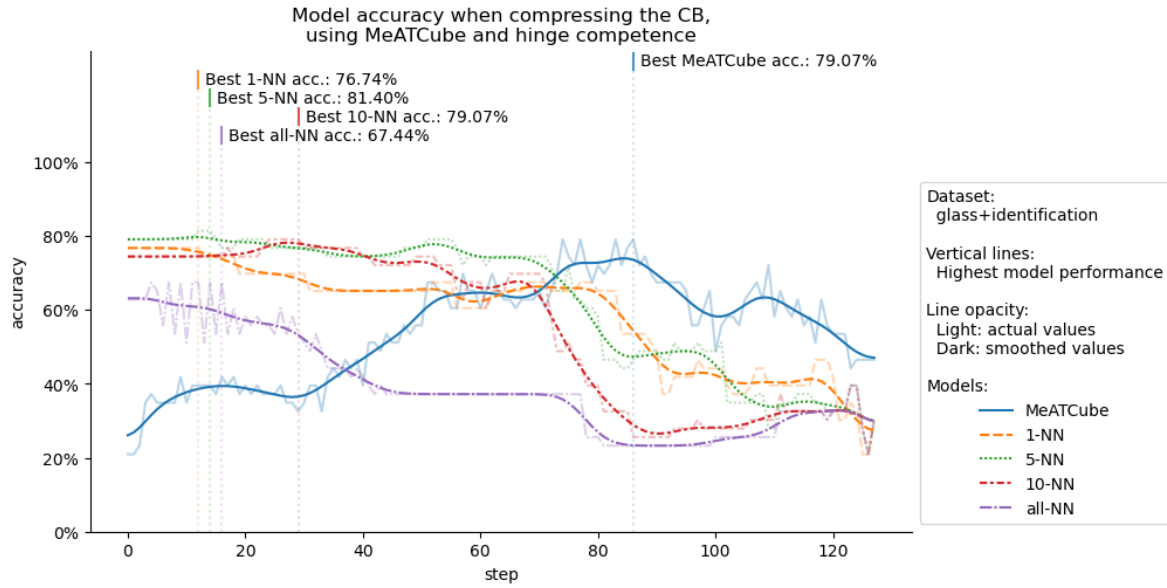
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

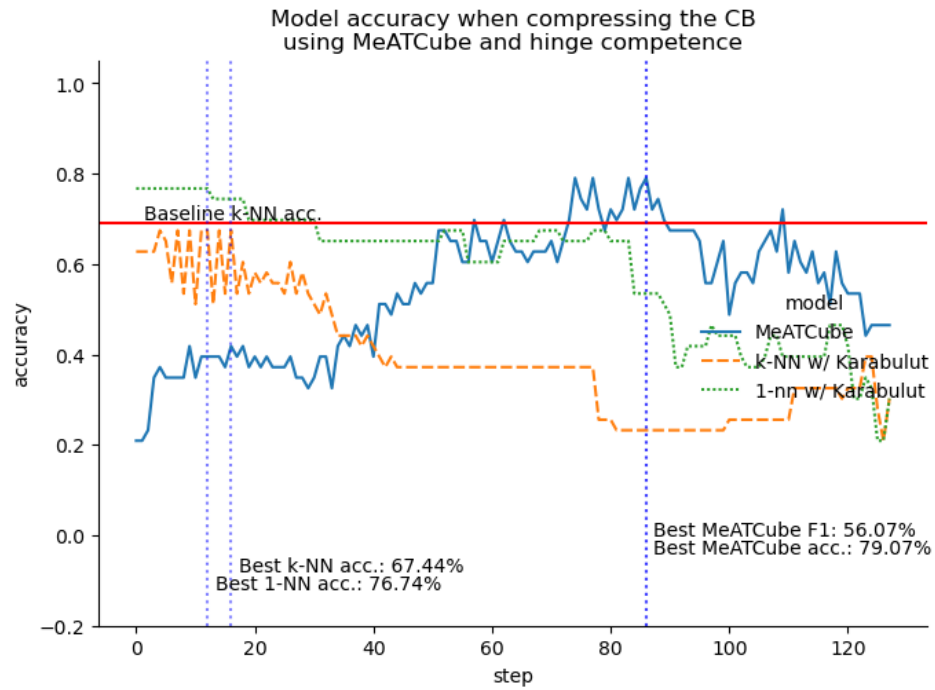




Result: Glass

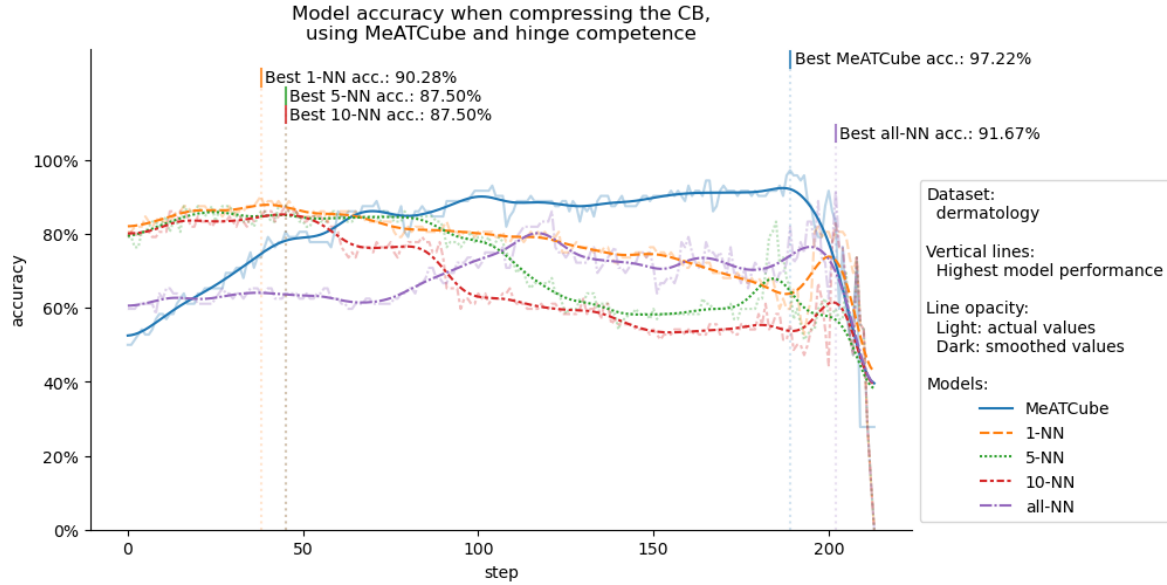
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

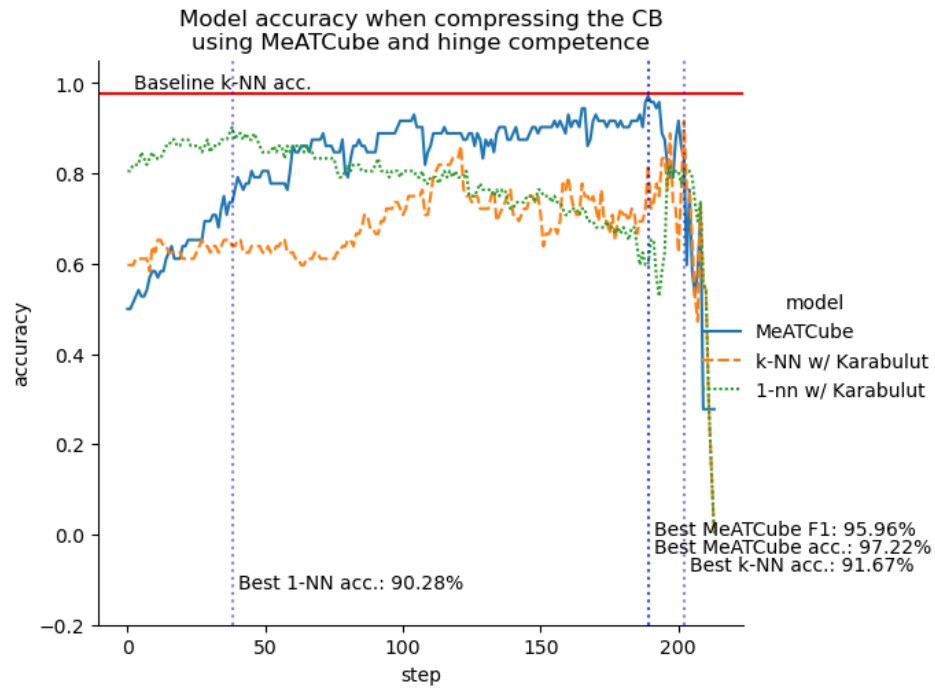




Result: Dermatology

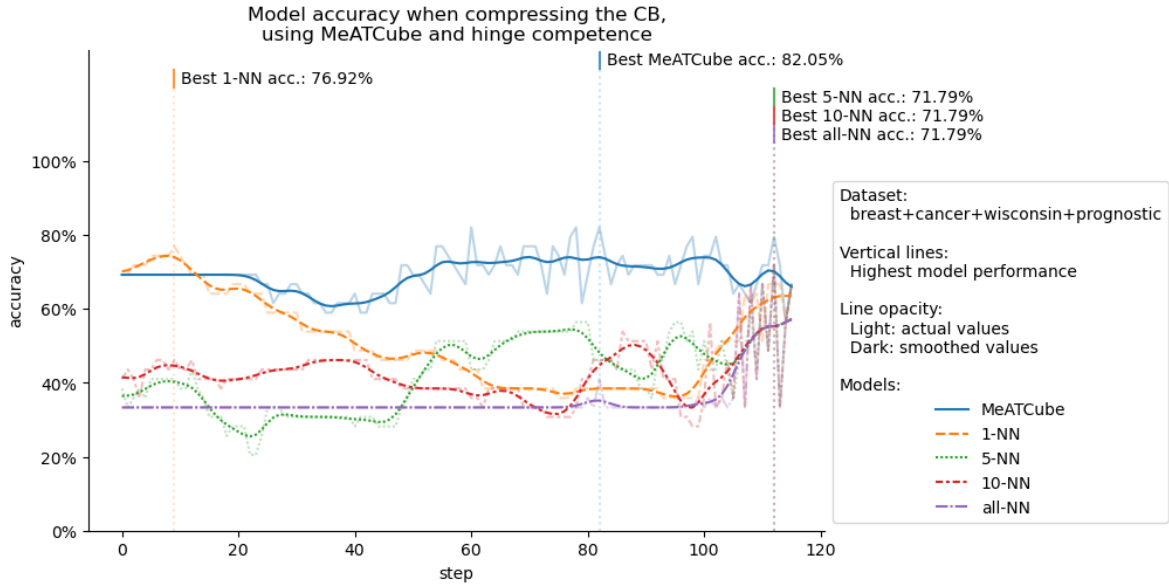
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

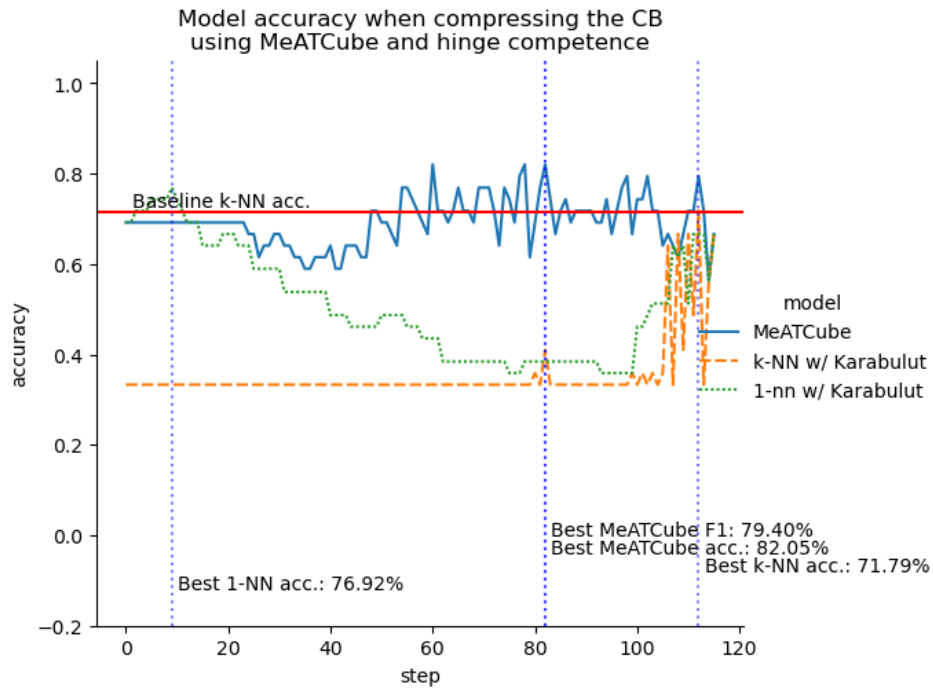




Result: Breast Cancer Pronostic

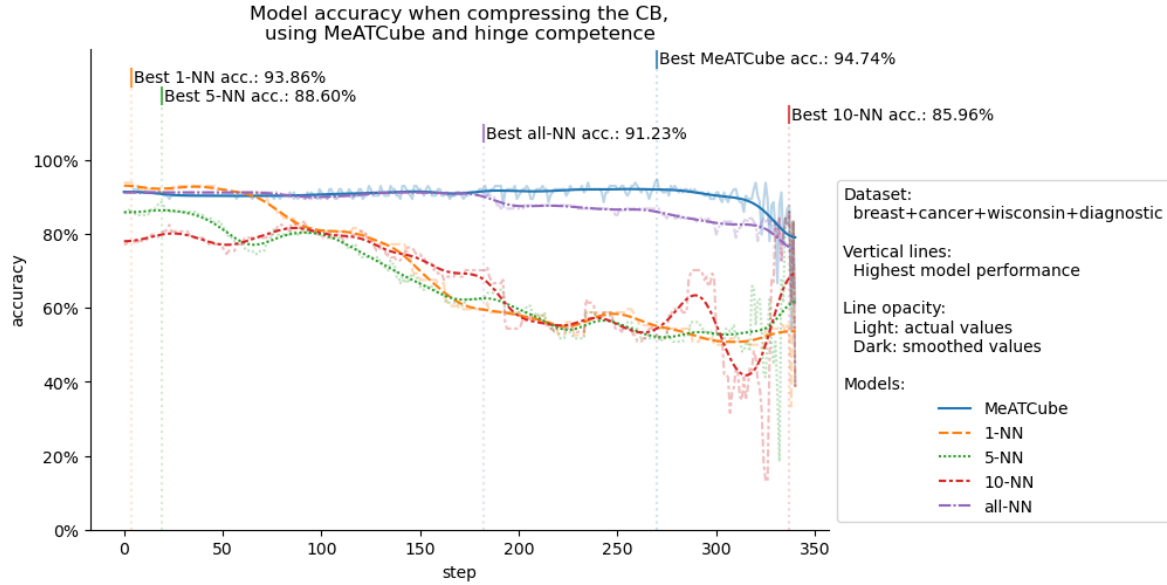
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

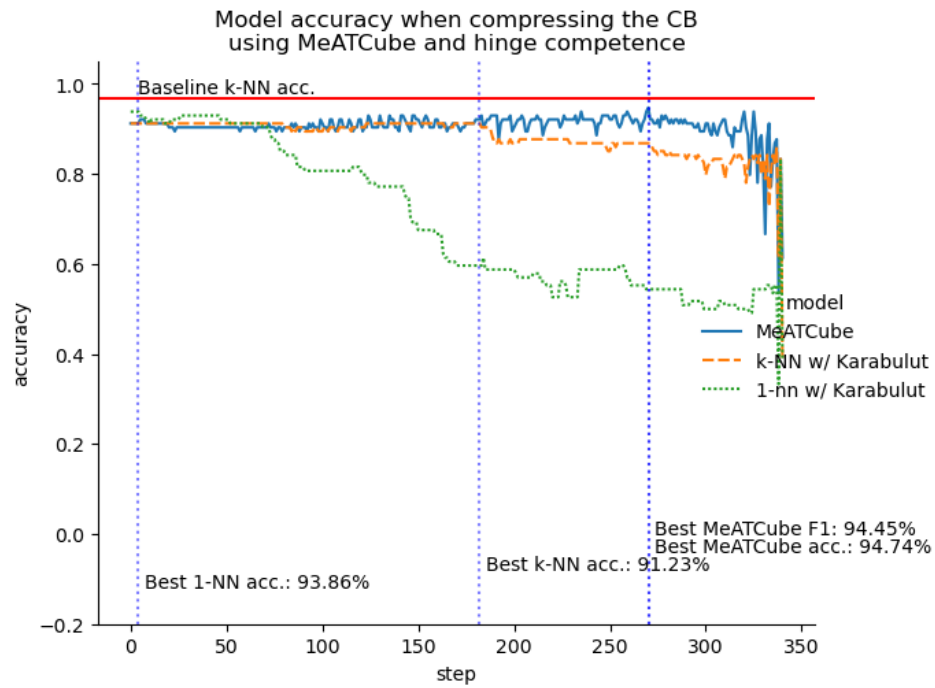




Result: Breast Cancer Diagnostic

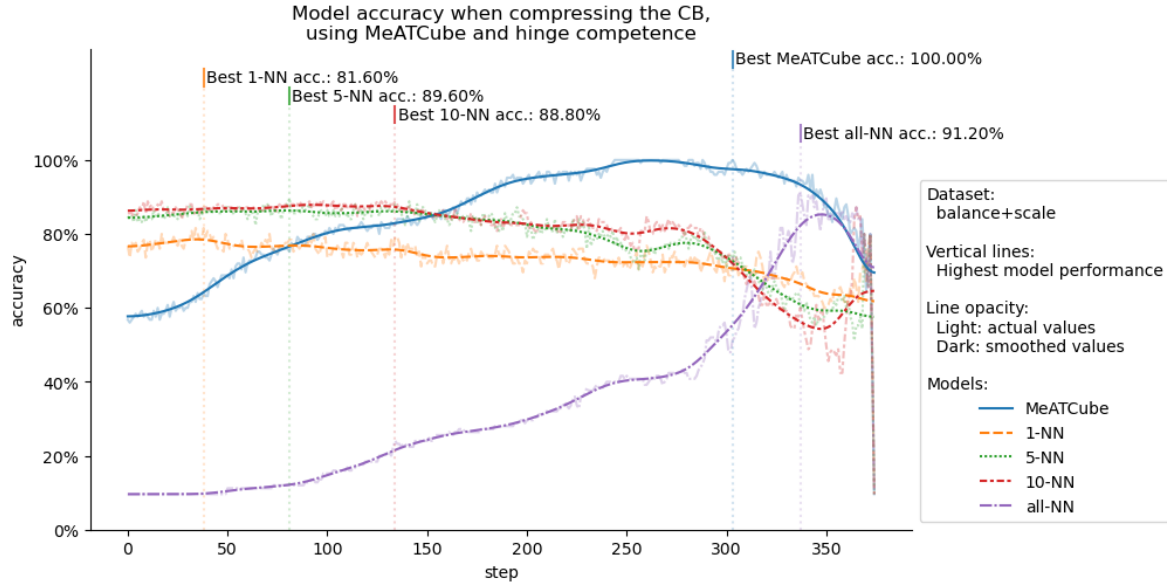
Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance

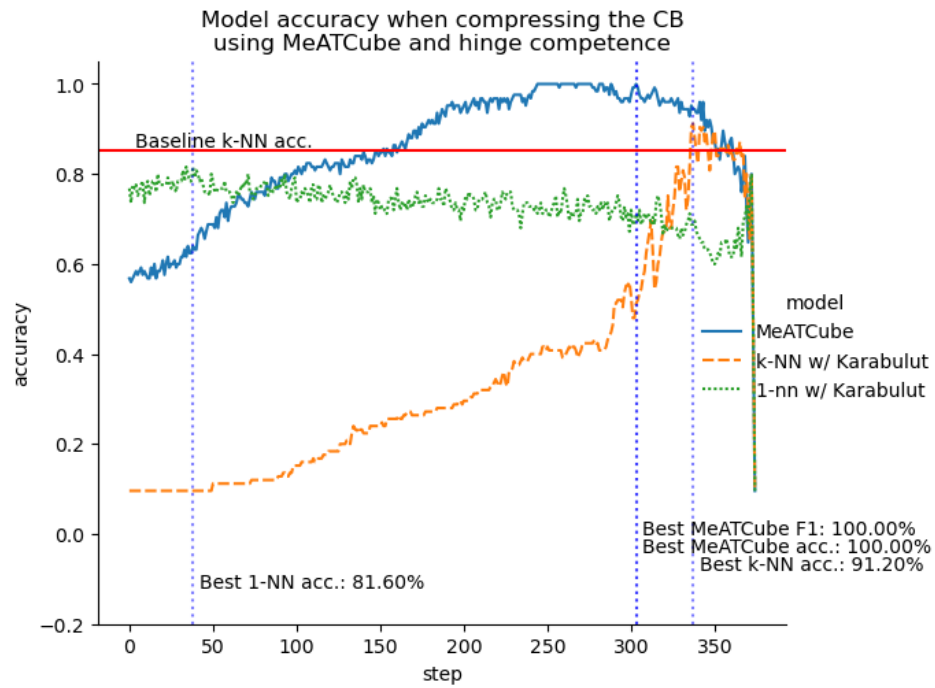




Result: Balance

Lenses
Credit Approval
Zoo
Wine
Teaching Assistant Evaluation
Post-Operative Patient
Pima indians Diabetes
Lung Cancer
Liver Disorders
Iris
Ionosphere
Hepatitis
Heart Disease Cleveland
Haberman's Survival
Glass Identification
Dermatology
Breast Cancer Pronostic
Breast Cancer Diagnostic
Balance





What next

Datasets to add?

Fixing the weights: how?

Apply the baselines instead of copying them
Apply more recent baselines

Cross-validation