Створення дослідницького прототипу для аналізу можливостей Explainable AI

Виконав: Чалюк А. О.
Науковий керівник: Олецький О. В.

Актуальність

- Зростаюча роль моделей ШІ у сферах з високими ставками.
- Потреба у зрозумілих для людини поясненнях для побудови довіри.
- Регуляторні та етичні чинники.

Мета дослідження

Систематична оцінка методів XAI, їх аналіз та порівняння.

Розглянуті методи XAI

- Лінійні та коефіцієнтні шляхи (Lasso, Ridge, GLM, GAM)
- Деревоподібні, моделі засновані на правила та сурогати (Decision Tree, RuleFit)
- На основі відхилень (LIME, Anchors, PermImp, PDP/ICE/ALE)
- На основі градієнтів (Saliency, IG, Grad-CAM)
- Контрфактичні (DiCE, VAE)
- Зменшення розмірності (PCA, t-SNE)
- На основі концепцій (CAV, TCAV)

Розглянуті набори даних

XOR

Синтетичний набір даних з шумовими значеннями (8), регресія та кластеризація

MNIST

Рукописні цифри, класифікація та кластеризація зору

Brain Tumor

Виявлення пухлин, містить 4 класи, медична візуалізація

Моделі та їхня точність

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.965	0.9444	0.9903	0.9668
MLPClassifier	1.000	1.000	1.000	1.000
XGBoost	1.000	1.000	1.000	1.000
Deep MLP	1.000	1.000	1.000	1.000
BernoulliNB	0.500	0.5152	0.4951	0.5049
1D-CNN	0.975	0.9537	1.000	0.9763
LSTM	1.000	1.000	1.000	1.000
TabTransformer	1.000	1.000	1.000	1.000

XOR

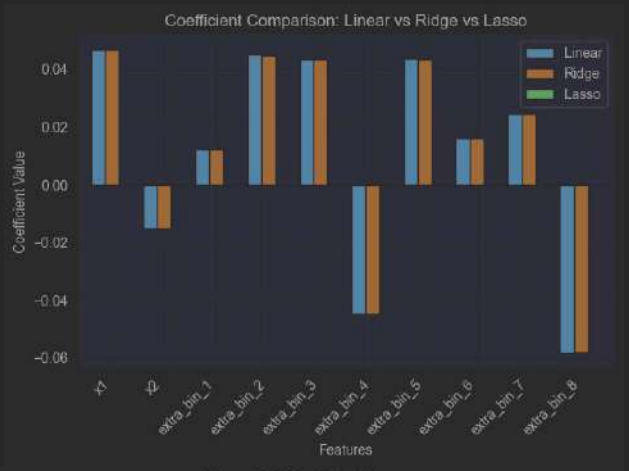
Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.8743	0.8725	0.8725	0.8725
MLPClassifier	0.9745	0.9743	0.9744	0.9743
Deep MLP (3 * 64)	0.9724	0.9723	0.9720	0.9721
XGBoost	0.9766	0.9766	0.9764	0.9764
BernoulliNB	0.8343	0.8349	0.8314	0.8321
GaussianNB	0.5539	0.6804	0.5460	0.5084
2D-CNN	0.9850	0.9852	0.9850	0.9850
LSTM	0.9844	0.9845	0.9842	0.9843

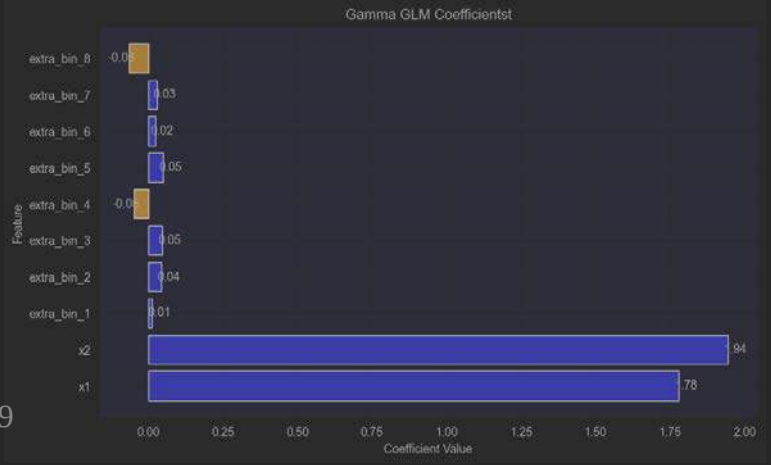
MNIST

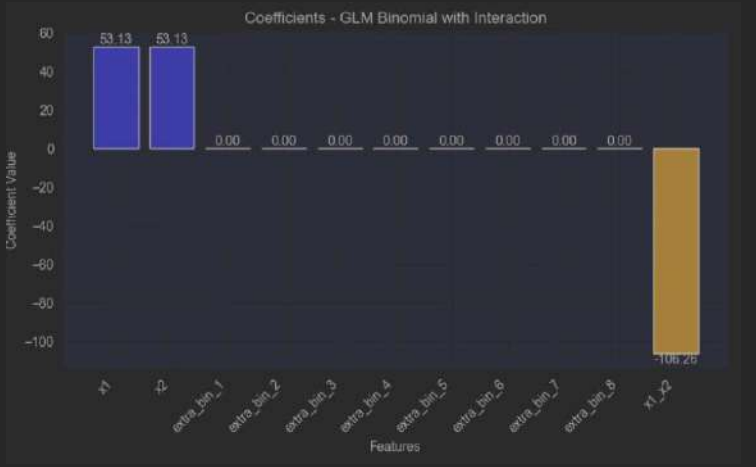
Лінійні методи

Dataset	Model	MSE	R^2
XOR	Linear	0.25	−0.01
	Ridge (α=1.0)	0.25	−0.01
	Lasso (α=0.1)	0.25	~0
MNIST	Linear	$9.71 \cdot 10^{1}$□	$-1.16 \cdot 10^{1}$□
	Ridge (α=1.0)	3.15	3.15
	Lasso (α=0.1)	4.59	0.45
Brain Tumor	Linear	0.61	0.53
	Ridge (α=1.0)	0.64	0.51
	Lasso (α=0.1)	1.14	0.13
	Lasso (α=0.001)	0.40	0.70

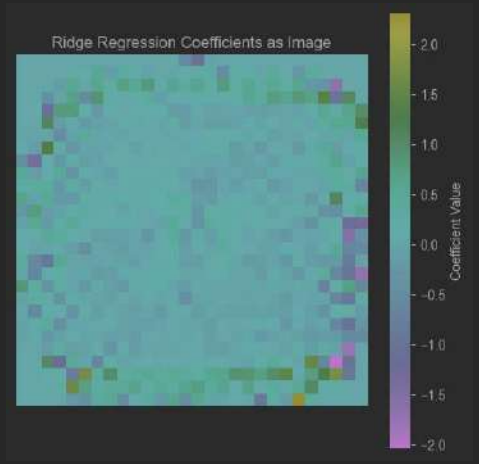
Коефіцієнти XOR

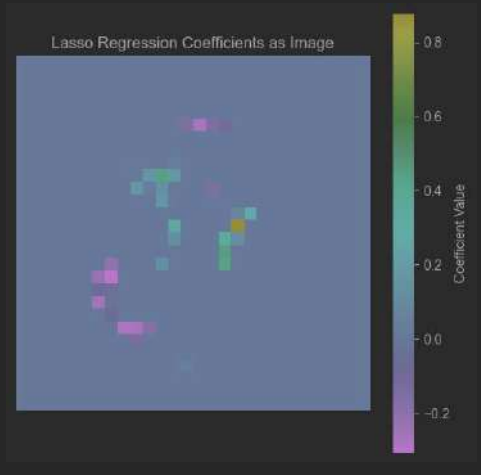


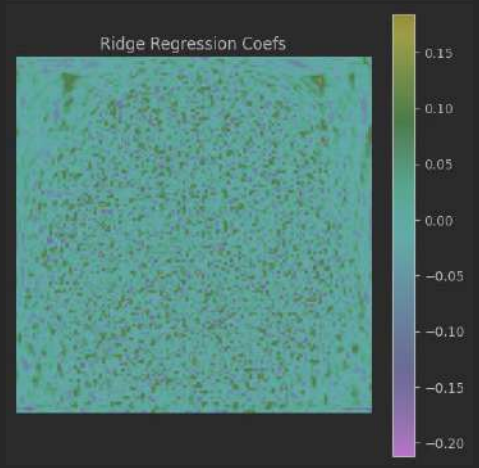


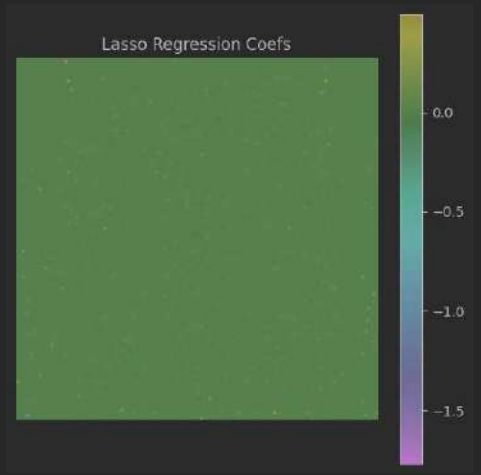


Коефіцієнти MNIST та Brain Tumor

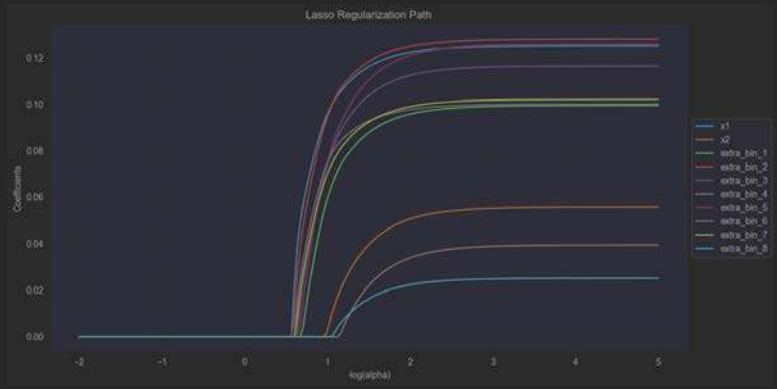




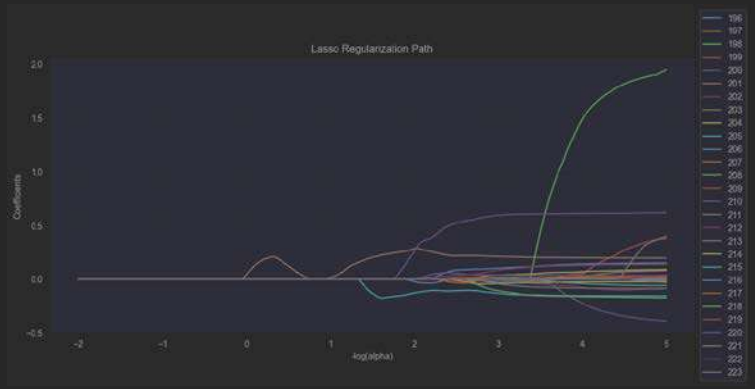




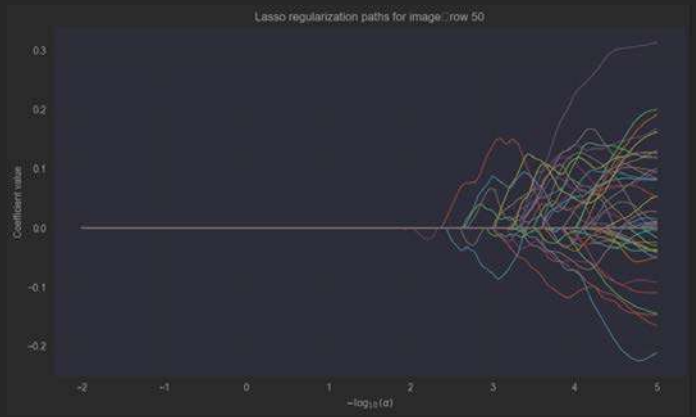
Lasso Regularization Path



XOR



MNIST



Brain Tumor

Linear Surrogate

Model	Runtime (s)	Comp	R^2	RMSE	Cls. Fidelity
Decision Tree	0.0008	9.9	0.1039	0.4714	0.6525
XGBoost	0.0224	10	0.1075	0.4609	0.6430
MLP	0.0011	10	0.1083	0.4649	0.6465
Deep MLP	0.0109	10	0.1078	0.4695	0.6505
TabTransformer	0.0544	9.9	0.1132	0.4674	0.6450
Bernoulli NB	0.0030	10	0.9984	0.0024	0.9935
1D-CNN	0.0506	9.6	0.2494	0.1228	0.5950
LSTM	0.0570	9.9	0.1125	0.4605	0.6415

Дерева рішень

XOR

- Для необмежених за глибиною моделей точність завжди 100%.
- З точки зору ентропії розділення за основною і шумовою ознаками еквівалентні.

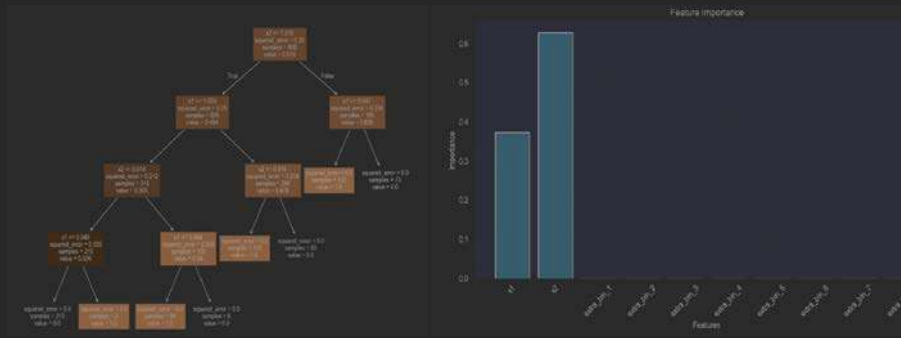
MNIST

- Структура дерева мало впливає на важливості пікселів.

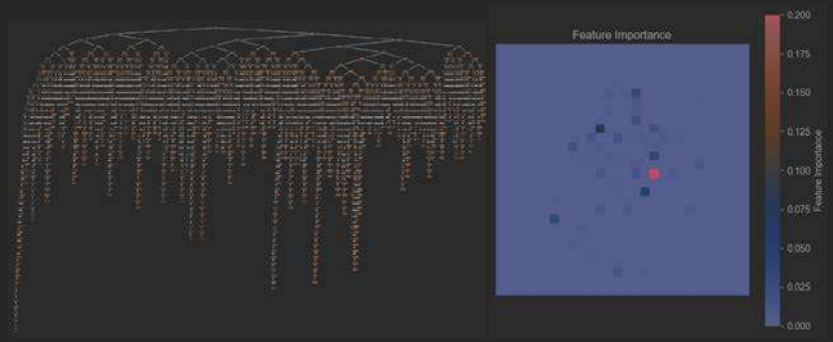
Brain Tumor

- Дерево є помітно меншим (порівняно з MNIST), що говорить про важливість лише кількох пікселів.

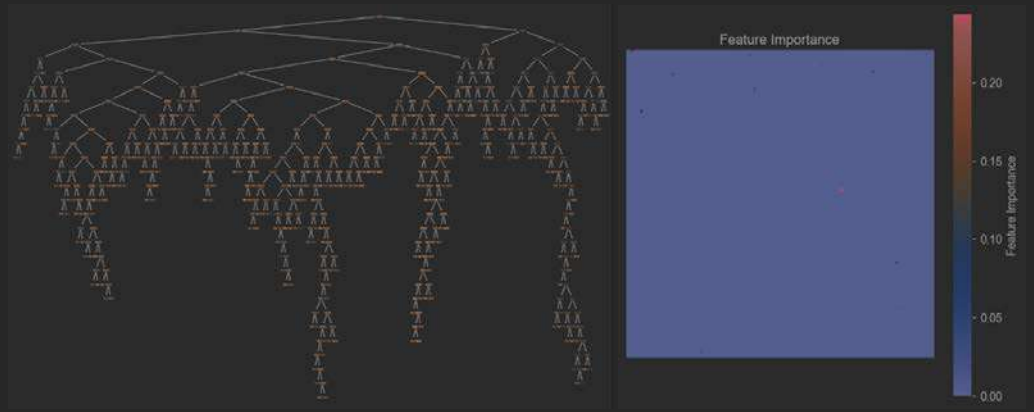
Дерева рішень



XOR (100%)



MNIST (88%)



Brain Tumor (89%)

Синтетичний набір даних

01 Навчаємо дерево рішень (глибина = d або необмежена).

02 Екстрагуємо правила.

03 Генеруємо синтетичний датасет за цими правилами.

04 Перевіряємо точність нейронної мережі на синтетичних даних і порівнюємо з оригіналом.

XOR

Noise Feats	Unrestricted rules	Synthetic samples	MPL accuracy on synthetic	Depth = 3 rules	Synthetic samples	MPL accuracy on synthetic
0	4	200	100%	4	200	100%
2	4	200	100%	4	200	100%
4	20	1000	100%	8	400	50%
6	28	1400	100%	8	400	77.75%
8	254	12700	91.93%	8	400	50%
10	284	14200	77.65%	8	400	55.75%
12	343	17150	70.12%	8	400	51.50%
14	280	14000	67.01%	8	400	53.25%
16	448	22400	63.32%	8	400	49.25%
18	576	28800	61.07%	8	400	50.50%

Точність на синтетичних даних

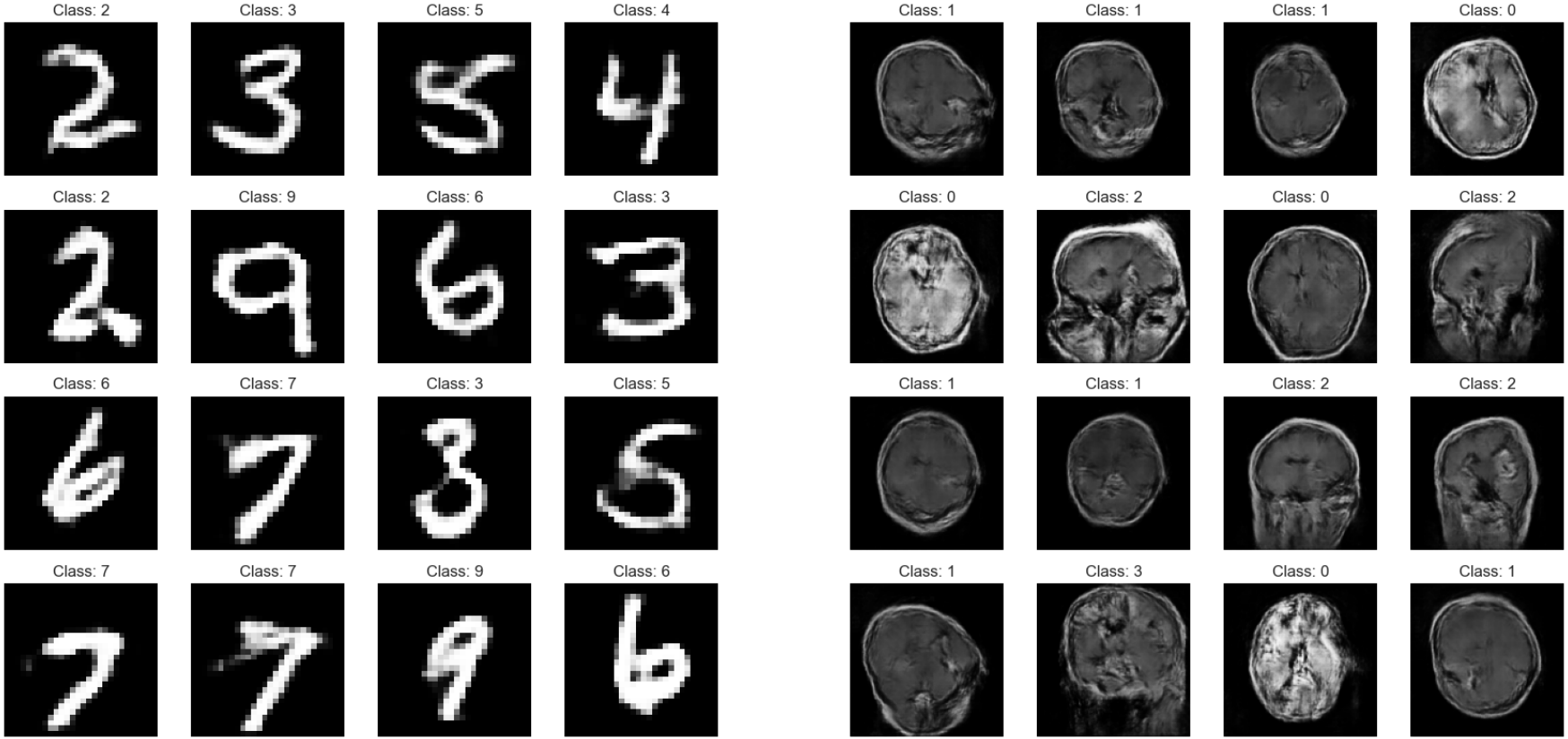
Точність MLP навченої на реальних даних і перевірених на синтетичному наборі:

- MNIST — 50.5%
- Brain Tumor — 57.7%

Зворотна задача. Навчання нейронної мережі на синтетичному наборі даних з подальшою перевіркою на оригінальному:

- MNIST — 53%
- Brain Tumor — 55%

Synthetic samples (AC-GAN)



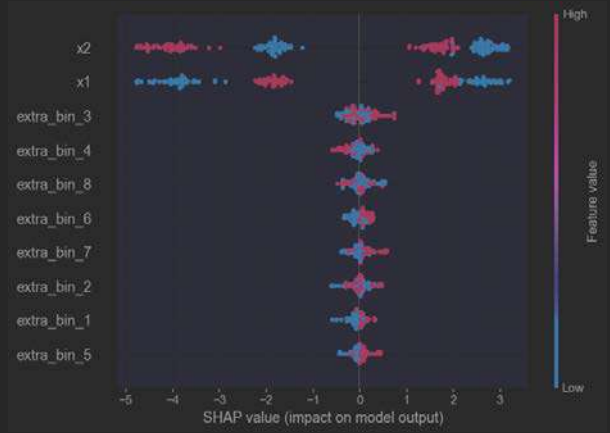
DT Surrogate

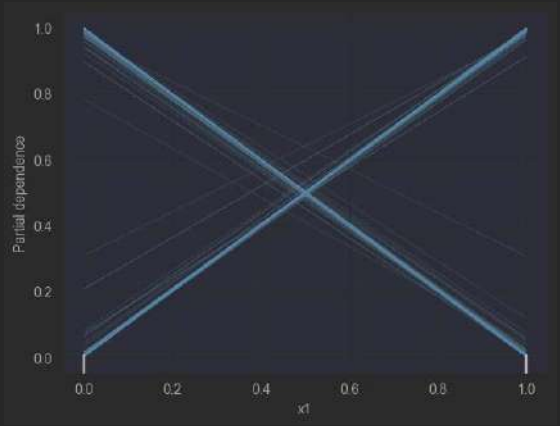
Model	Surrogate DT fidelity (depth 3)	Surrogate DT fidelity (unrestricted depth)
Decision Tree	0.535	1
XGBoost	0.510	0.965
MLP	0.510	0.965
Deep MLP	0.510	0.965
Bernoulli NB	0.770	0.985
1D-CNN	0.555	0.96
LSTM	0.550	0.96

Perturbation Methods

- LIME — локальний лінійний сурогат навколо одного прикладу
- Anchors — високоточні «якірні» правила для пояснення окремих рішень
- Permutation Feature Importance — оцінка втрати метрики при рандомізації кожної ознаки
- Partial Dependence (PDP) / ICE / c-ICE / ALE
  - PDP — усереднена залежність прогнозу від однієї ознаки
  - ICE / c-ICE — індивідуальні криві для кожного зразка (центровані або ні)
  - ALE — локальна похідна функція залежності

Perturbation Methods





Anchor: 0.00 < extra_bin_2 <= 1.00 AND 0.00 < extra_bin_3 <= 1.00 AND 0.00 < extra_bin_5 <= 1.00 AND 0.00 < extra_bin_6 <= 1.00 AND 0.00 < extra_bin_1 <= 1.00 AND extra_bin_8 <= 0.00 AND extra_bin_7 <= 1.00 AND extra_bin_4 <= 1.00 AND x1 = 0 AND x2 = 1

Precision: 1.00

Coverage: 0.00

Permutation Importance

Model	Runtime (s)	Comp	MeanImp
Decision Tree	0.1407	9	0.0926
XGBoost	1.1168	2	0.1004
MLP	0.2100	2	0.1004
Deep MLP	0.5318	2	0.1004
Bernoulli NB	0.2654	10	-0.0009
1D-CNN	1.0208	5	0.0950
LSTM	1.0546	2	0.1010
TabTransformer	0.7138	2	0.1010

Partial Dependence

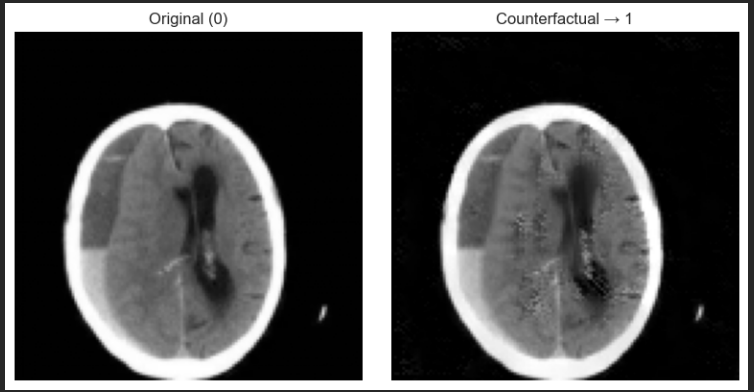
Model	Runtime (s)	Mean Range
Decision Tree	0.0022	0.03
XGBoost	0.0386	0.0153
MLP	0.0024	0.0127
Deep MLP	0.0077	0.0113
Bernoulli NB	0.0061	0.0340
1D-CNN	0.1133	0.0296
LSTM	0.1474	0.0112
TabTransformer	0.1293	0.0111

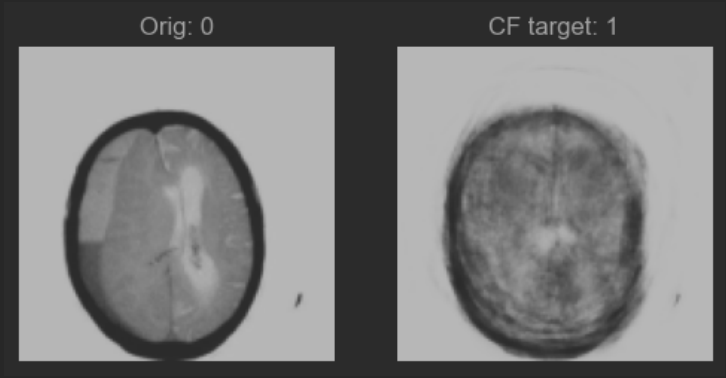
Counterfactual Explanation

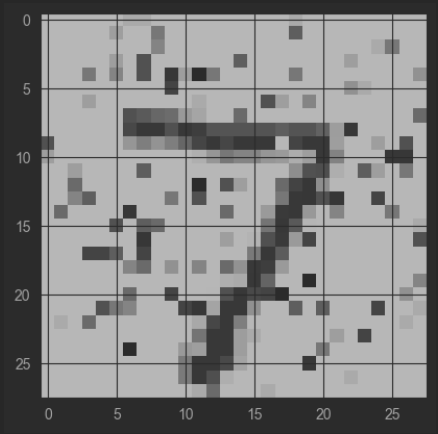
Ідея: знайти мінімальні зміни вхідних ознаках, необхідні для зміни прогнозу моделі

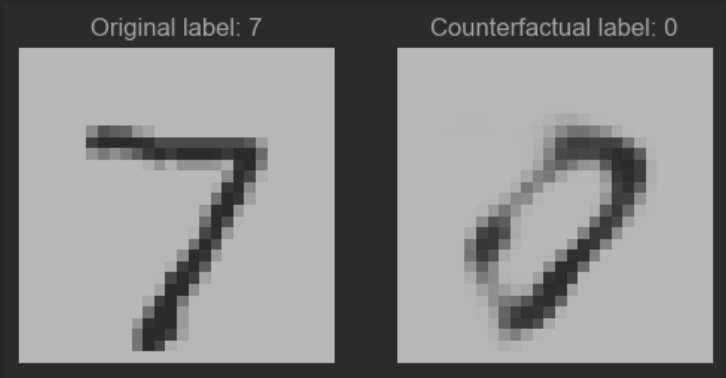
- DiCE
  - Пошук набору контрфактів через оптимізацію цільової функції та функції регуляризації
  - Швидкі
  - Часто виходить шум, мало реалістичності
- VAE-контрфакти
  - Генерація через прихований простір автоенкодера
  - Реалістичніші, зберігають структуру даних
  - Потребують додаткового навчання й можуть узагальнювати занадто широко

Counterfactual Explanation

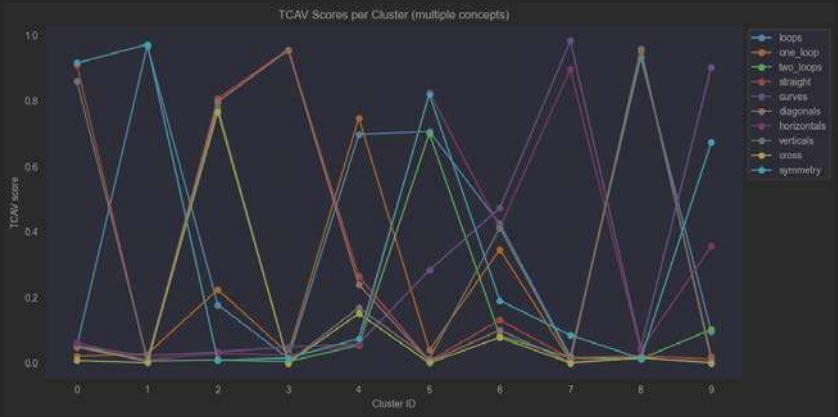




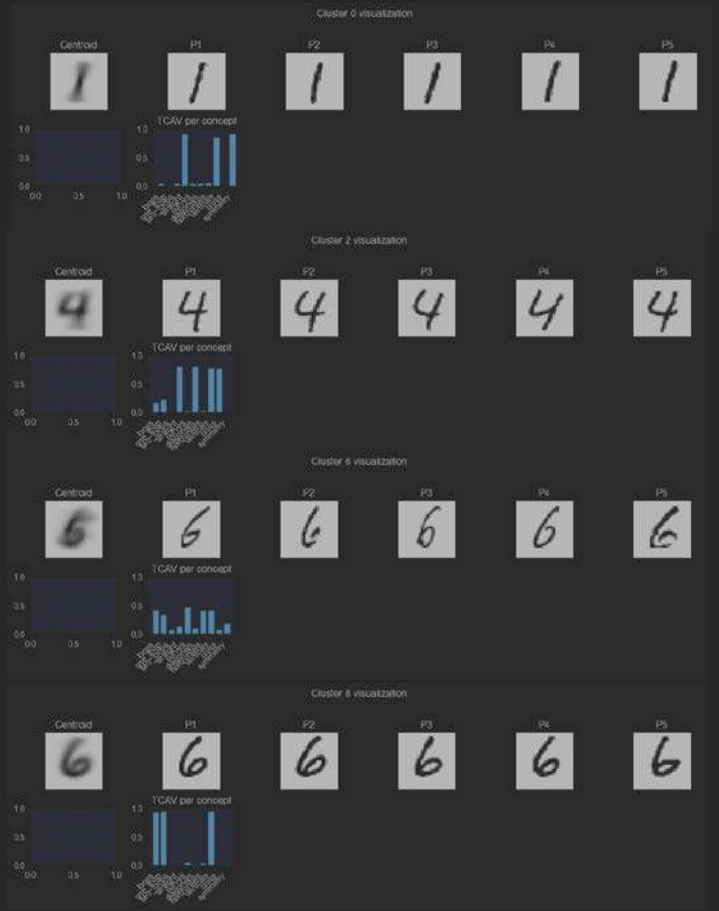




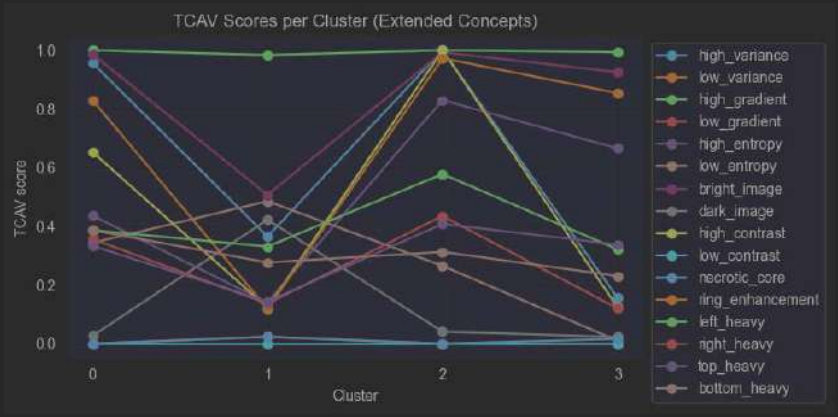
TCAV (MNIST)

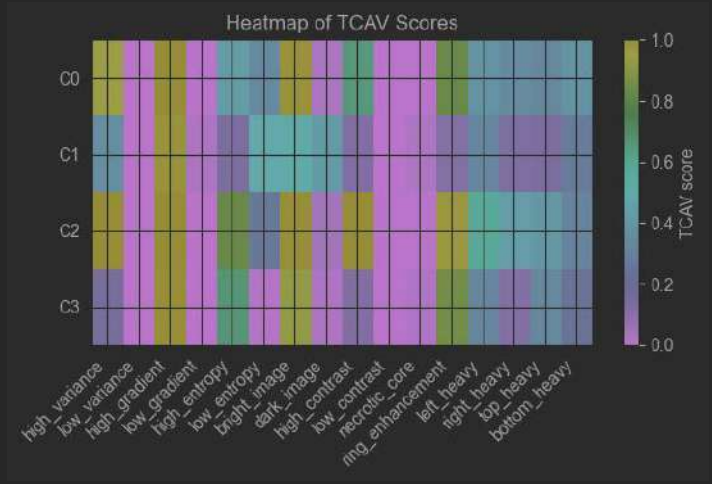

TCAV Scores per Cluster (multiple concepts)
loops
one_loop
two_loops
straight
curves
diagonals
horizontals
verticals
cross
symmetry
TCAV score
Cluster ID


Heatmap of TCAV Scores
C0
C1
C2
C3
C4
C5
C6
C7
C8
C9
loops
one_loop
two_loops
straight
curves
diagonals
horizontals
verticals
cross
symmetry
TCAV score

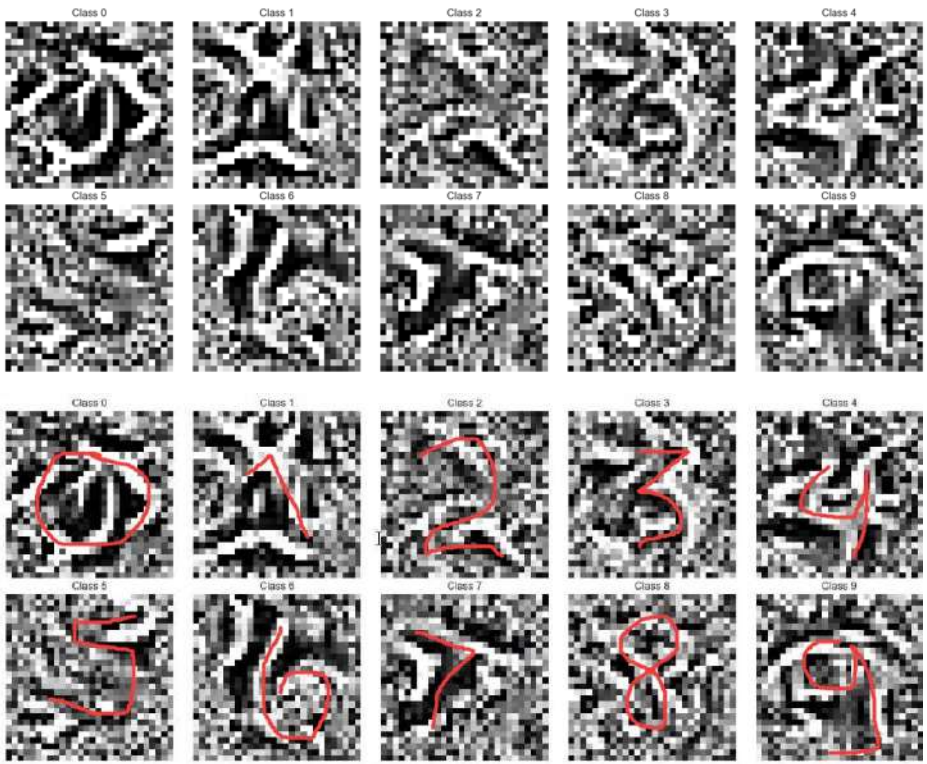

Cluster 0 visualization
Centroid
P1
P2
P3
P4
P5
TCAV per concept
Cluster 2 visualization
Cluster 6 visualization
Cluster 8 visualization

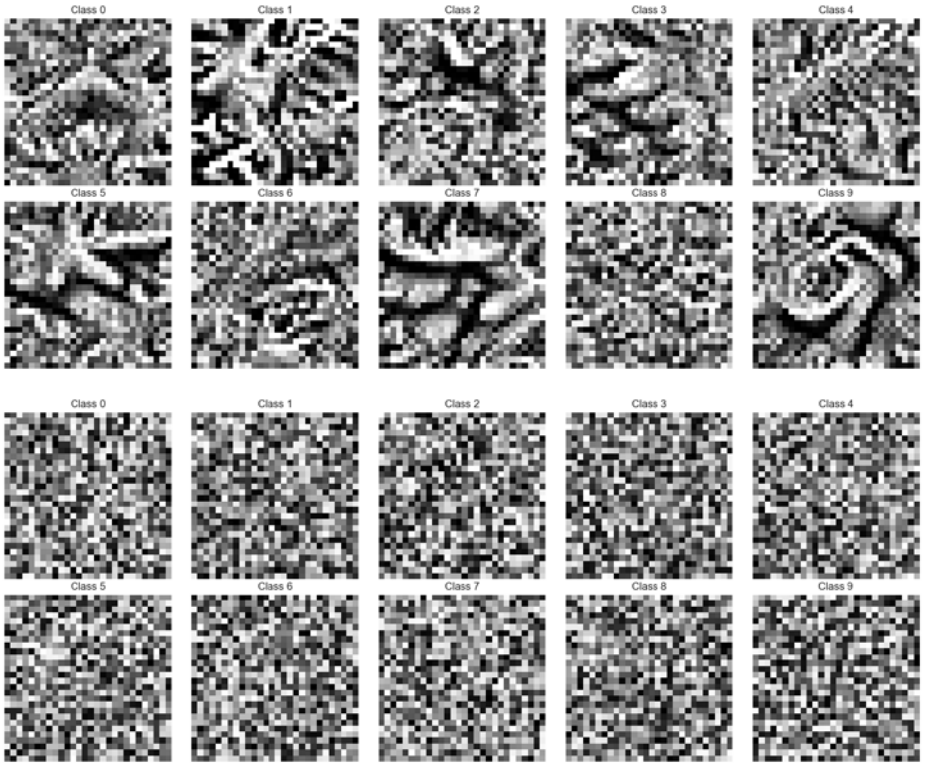
TCAV (Brain Tumor)





Etalon Images





Висновки

- Проведено систематичну оцінку популярних XAI-методів на 3 різних наборах даних. Жоден підхід не домінує за всіма метриками; вибір залежить від задачі, типу даних і моделі.
- Лінійні/ дерева рішень – швидкі й інтерпретовані для табличних задач (XOR), але слабкі в складних сценаріях (MNIST, MRI).
- Perturbation методи – гнучкі й модель-агностичні, проте обчислювально дорогі та чутливі до кореляцій ознак.
- Методика синтетичних експериментів показала фундаментальну розбіжність між точністю та інтерпретованістю.

Висновки

- Використання високорівневих концепцій є дуже корисним для генерації зрозумілих людям пояснень, проте теж є досить дорогими з точки зору обчислень.
- Для невеликих моделей (або певних частин моделі) візуалізація еталонних результатів дозволяє приблизно зрозуміти, на що спирається модель при ухвалені рішень.

Дякую за увагу