1. Dataset

The complete dataset used in this study is summarized in Supplemental Table S1, which provides detailed information on sample size and key variables. This Supplemental Table Sincludes different datasets to ensure transparency and reproducibility. Additional methodological details and raw data are available upon request.

Supplemental Table S1. Comprehensive overview of the dataset, including sample sizes, demographic distribution, and key variables analyzed in the study.

Dataset	Description
Heart Disease Dataset	The dataset includes 303 patient records, each with 14 clinical features.
Breast Cancer Dataset	The dataset consists of 569 records with 30 numerical features extracted from FNA images, and a binary target indicating diagnosis (357 benign, 212 malignant).
Online News Popularity Dataset	The dataset includes 4,954 news articles, each described by 60 features, with a target variable indicating the number of social media shares.
Iris Dataset	The dataset contains 150 samples from three iris species, each with four numerical features and a species label, commonly used for classification tasks.
Car Evaluation Dataset	The dataset contains 1,728 instances with 6 categorical features describing car characteristics, and a multi-class target representing overall acceptability.
Ultrasound-MRI	The dataset comprises 1,409 2D slices (128×128) from 24 patients, including ultrasound images and their corresponding synthetic MRIs generated using a Pix2Pix model, with real MRIs as ground truth.
Wine Quality Dataset	The dataset includes red Portuguese "Vinho Verde" wines, each described by 11 physicochemical features, with a target variable representing quality scores from 0 to 10.
Room Occupancy Dataset	The dataset contains 5 environmental features and a binary target indicating room occupancy based on indoor sensor data.
Weather Dataset	The dataset contains approximately 96,000 observations with 12 features describing historical weather conditions, including temperature, humidity, wind, and categorical weather descriptions.
House Dataset	The dataset includes 8 features describing property characteristics, used to predict house prices.
Multi Center Lung CT Dataset	The dataset contains 240 lung CT tumor segmentations from 8 multicenter sources (LCTSC, LUNG CT Diagnosis, NSCLC, QIN, Lung-Fused-CT-Pathology, RIDER, SPIE-AAPM, and TCGA).
computer-generated Synthetic Noisy Image Pairs	The dataset includes grayscale image pairs (ground truth and noisy variants) for standardized image quality assessment, with ground truth images being random 8-bit (64×64 pixels) and predictions as noise-corrupted versions using Gaussian noise (μ=0, σ=30)

2. Accuracy & Reliability

In this section, we present a comprehensive overview of all results obtained using various datasets and libraries to evaluate the accuracy and reliability of AllMetrics. The results are organized according to the specific tasks performed, with corresponding tables provided for each task to facilitate clear comparison and analysis.

2.1 Binary Classification

2.1.1 Heart Disease Dataset. We used the Cleveland subset of the Heart Disease Dataset, which contains 303 patient records. Each record includes 14 selected clinical features related to demographics, diagnostic test results, and risk factors. The target variable indicates the presence (1) or absence (0) of heart disease. All personally identifiable information was removed and replaced with dummy values.

Supplemental Table S2. Accuracy and Reliability Comparison of Binary Classification Models on the Heart Disease Dataset

Metric	Library	Value
Accuracy	AllMetrics	0.808441558441558
	Scikit-Learn	0.808441558441558
	PyTorch	0.808441579341888
	TensorFlow	0.808441579341888
	PyCM	0.808441558441558
Precision	AllMetrics	{0: 0.8623188405797102, 1: 0.7647058823529411}
	Scikit-Learn	0.764705882352941
	PyTorch	0.764705896377564

	TensorFlow	0.764705896377564
	PyCM	{0: 0.8623188405797102, 1: 0.7647058823529411}
Recall	AllMetrics	{0: 0.7484276729559748, 1: 0.87248322147651}
	Scikit-Learn	0.87248322147651
	PyTorch	0.872483193874359
	TensorFlow	0.872483193874359
	PyCM	{0: 0.7484276729559748, 1: 0.87248322147651}
F1 Score	AllMetrics	{0: 0.8013468013468014, 1: 0.8150470219435737}
	Scikit-Learn	0.815047021943574
	PyTorch	0.815047025680542
	TensorFlow	0.815047025680542
	PyCM	{0: 0.8013468013468014, 1: 0.8150470219435737}
Balanced Accuracy	AllMetrics	0.810455447216242
	Scikit-Learn	0.810455447216242
Matthews Correlation Coefficient	AllMetrics	{0: 0.6239603204911299, 1: 0.6239603204911299}
	Scikit-Learn	0.62396032049113
	PyCM	{0: 0.6239603204911299, 1: 0.6239603204911299}
Cohens Kappa	AllMetrics	0.61817112119684
	Scikit-Learn	0.61817112119684
	PyCM	0.61817112119684
F-Beta Score	AllMetrics	0.784077201447527
	Scikit-Learn	0.784077201447527
Jaccard Index	AllMetrics	{0: 0.6685393258426966, 1: 0.6878306878306878}
	Scikit-Learn	0.687830687830688
	TensorFlow	0.678184986114502
	PyCM	{0: 0.6685393258426966, 1: 0.6878306878306878}

2.1.2 Breast Cancer Wisconsin Dataset. We use the Breast Cancer Wisconsin (Diagnostic) Dataset, which includes 569 records derived from digitized images of fine needle aspirate (FNA) samples of breast masses. Each record contains 30 numerical features describing characteristics of cell nuclei, such as radius, texture, perimeter, area, and others, computed as the mean, standard error, and "worst" value for each measurement. The target variable is binary, indicating diagnosis as malignant (M) or benign (B). There are no missing values in the dataset. The class distribution includes 357 benign and 212 malignant cases.

Supplemental Table S3. Accuracy and Reliability Comparison of Binary Classification Models on the Breast Cancer Dataset

Metric	Library	Value
Accuracy	AllMetrics	0.947368421052632
	Scikit-Learn	0.947368421052632
	PyTorch	0.947368443012238
	TensorFlow	0.947368443012238
	PyCM	0.947368421052632
Precision	AllMetrics	{0: 0.922077922077922, 1: 1.0}
	Scikit-Learn	1
	PyTorch	1
	TensorFlow	1
	PyCM	{0: 0.922077922077922, 1: 1.0}
Recall	AllMetrics	{0: 1.0, 1: 0.8604651162790697}
	Scikit-Learn	0.86046511627907
	PyTorch	0.860465109348297
	TensorFlow	0.860465109348297
	PyCM	{0: 1.0, 1: 0.8604651162790697}
F1 Score	AllMetrics	{0: 0.9594594594594594, 1: 0.925}
	Scikit-Learn	0.925
	PyTorch	0.925000011920929
	TensorFlow	0.925000011920929
	PyCM	{0: 0.9594594594594594, 1: 0.925}
Balanced Accuracy	AllMetrics	0.930232558139535
	Scikit-Learn	0.930232558139535
Matthews Correlation Coefficient	AllMetrics	{0: 0.8907389552720495, 1: 0.8907389552720495}
	Scikit-Learn	0.89073895527205
	PyCM	{0: 0.8907389552720495, 1: 0.8907389552720495}
Cohens Kappa	AllMetrics	0.88480970023577
	Scikit-Learn	0.88480970023577
	PyCM	0.88480970023577

F-Beta Score	AllMetrics	0.968586387434555
	Scikit-Learn	0.968586387434555
Jaccard Index	AllMetrics	{0: 0.922077922077922, 1: 0.8604651162790697}
	Scikit-Learn	0.86046511627907
	TensorFlow	0.891271471977234
	PyCM	{0: 0.922077922077922, 1: 0.8604651162790697}

2.1.3 Online News Popularity Reduced dataset. We use the Online News Popularity Reduced dataset, which contains 4,954 news articles. Each entry includes 60 numerical and categorical features describing the content, structure, and metadata of the articles, such as word count, sentiment scores, topic distribution, and publication day. The target variable, shares, indicates the number of times each article was shared on social media. There are no missing values in the dataset.

Supplemental Table S4. Accuracy and Reliability Comparison of Binary Classification Models on the Online News Popularity Reduced Dataset

Metric	Library	Value
Accuracy	AllMetrics	0.73365617433414
	Scikit-Learn	0.73365617433414
	PyTorch	0.733656167984009
	TensorFlow	0.733656167984009
	PyCM	0.73365617433414
Precision	AllMetrics	{0: 0.7334410339256866, 1: 1.0}
	Scikit-Learn	1
	PyTorch	1
	TensorFlow	1
	PyCM	{0: 0.7334410339256866, 1: 1.0}
Recall	AllMetrics	{0: 1.0, 1: 0.0030211480362537764}
	Scikit-Learn	0.00302114803625378
	PyTorch	0.00302114803344011
	TensorFlow	0.00302114803344011
	PyCM	{0: 1.0, 1: 0.0030211480362537764}
F1 Score	AllMetrics	{0: 0.8462255358807083, 1: 0.006024096385542169}
	Scikit-Learn	0.00602409638554217
	PyTorch	0.00602409616112709
	TensorFlow	0.00602409662678838
	PyCM	{0: 0.8462255358807083, 1: 0.006024096385542169}
Balanced Accuracy	AllMetrics	0.501510574018127
	Scikit-Learn	0.501510574018127
Matthews Correlation Coefficient	AllMetrics	{0: 0.047072645340500326, 1: 0.047072645340500326}
	Scikit-Learn	0.0470726453405003
	PyCM	{0: 0.047072645340500326, 1: 0.047072645340500326}
Cohens Kappa	AllMetrics	0.00442186974963834
	Scikit-Learn	0.00442186974963834
	PyCM	0.00442186974963834
F-Beta Score	AllMetrics	0.0149253731343284
	Scikit-Learn	0.0149253731343284
Jaccard Index	AllMetrics	{0: 0.7334410339256866, 1: 0.0030211480362537764}
	Scikit-Learn	0.00302114803625378
	TensorFlow	0.36823108792305
	PyCM	{0: 0.7334410339256866, 1: 0.0030211480362537764}

2.2 Multi-Class Classification

2.2.1 Iris dataset. We use the Iris dataset, which contains 150 samples from three species of iris flowers: *setosa*, *versicolor*, and *virginica*. Each sample includes four numerical features—sepal length, sepal width, petal length, and petal width along with the species label. The dataset is commonly used for classification tasks and is noSupplemental Table Sfor its simplicity and balanced class distribution.

Supplemental Table S5. Accuracy and Reliability Comparison of Binary Classification Models on the Iris Dataset

Metric	Library	Value
Accuracy	AllMetrics	0.633333333333333
	Scikit-Learn	0.633333333333333
	PyTorch	0.633333325386047

	TensorFlow	0.633333325386047
	PyCM	0.633333333333333
Precision	AllMetrics	{0: 1.0, 1: 0.4, 2: 0.5454545454545454}
	Scikit-Learn	0.648484848484849
	PyTorch	0.648484885692596
	TensorFlow	0.648484885692596
	PyCM	{0: 1.0, 1: 0.4, 2: 0.5454545454545454}
Recall	AllMetrics	{0: 0.9, 1: 0.4444444444444444, 2: 0.5454545454545454}
	Scikit-Learn	0.62996632996633
	PyTorch	0.62996631860733
	TensorFlow	1
	PyCM	{0: 0.9, 1: 0.4444444444444444, 2: 0.5454545454545454}
F1 Score	AllMetrics	{0: 0.9473684210526315, 1: 0.42105263157894735, 2: 0.5454545454545454}
	Scikit-Learn	0.637958532695375
	PyTorch	0.637958526611328
	TensorFlow	0.97560977935791
	PyCM	{0: 0.9473684210526315, 1: 0.42105263157894735, 2: 0.5454545454545454}
Balanced Accuracy	AllMetrics	0.62996632996633
	Scikit-Learn	0.62996632996633
Matthews Correlation Coefficient	AllMetrics	{0: 0.9258200997725514, 1: 0.1543033499620919, 2: 0.2822966507177033}
	Scikit-Learn	0.449832775919732
	PyCM	{0: 0.9258200997725514, 1: 0.1543033499620919, 2: 0.2822966507177033}
Cohens Kappa	AllMetrics	0.449081803005008
	Scikit-Learn	0.449081803005008
	PyCM	0.449081803005008
F-Beta Score	AllMetrics	0.643959560108628
	Scikit-Learn	0.643959560108628
Jaccard Index	AllMetrics	{0: 0.9, 1: 0.26666666666666666, 2: 0.375}
	Scikit-Learn	0.513888888888889
	TensorFlow	0.513888895511627
	PyCM	{0: 0.9, 1: 0.26666666666666666, 2: 0.375}

2.2.2 Car Evaluation Dataset. We use the Car Evaluation dataset, which contains 1,728 instances and 6 categorical attributes: buying price, maintenance cost, number of doors, passenger capacity, luggage boot size, and safety. The target variable represents the overall acceptability of a car, classified into multiple categories. The dataset was derived from a hierarchical decision model designed for multi-attribute decision making and contains no missing values. All possible combinations of attribute values are represented, making it useful for evaluating classification and rule-learning algorithms.

Supplemental Table S6. Accuracy and Reliability Comparison of Multi-Class Classification Models on the Car Evaluation Dataset

Metric	Library	Value
Accuracy	AllMetrics	0.935185185185185
	Scikit-Learn	0.935185185185185
	PyTorch	0.935185194015503
	TensorFlow	0.935185194015503
	PyCM	0.935185185185185
Precision	AllMetrics	{0: 0.8854166666666666, 1: 0.75, 2: 0.9761904761904762, 3: 0.7692307692307693}
	Scikit-Learn	0.845209478021978
	PyTorch	0.84520947933197
	TensorFlow	0.967261910438538
	PyCM	{0: 0.8854166666666666, 1: 0.75, 2: 0.9761904761904762, 3: 0.7692307692307693}
Recall	AllMetrics	{0: 0.8854166666666666, 1: 0.6666666666666666, 2: 0.9630872483221476, 3: 1.0}
	Scikit-Learn	0.87879264541387
	PyTorch	0.878792643547058
	TensorFlow	0.967261910438538
	PyCM	{0: 0.8854166666666666, 1: 0.6666666666666666, 2: 0.9630872483221476, 3: 1.0}
F1 Score	AllMetrics	{0: 0.8854166666666666, 1: 0.7058823529411765, 2: 0.9695945945945946, 3: 0.8695652173913043}
	Scikit-Learn	0.857614707898436
	PyTorch	0.857614696025848
	TensorFlow	0.967261910438538
	PyCM	{0: 0.8854166666666666, 1: 0.7058823529411765, 2: 0.9695945945945946, 3: 0.8695652173913043}
Balanced Accuracy	AllMetrics	0.87879264541387

	Scikit-Learn	0.87879264541387
Matthews Correlation Coefficient	AllMetrics	{0: 0.8526785714285714, 1: 0.6951816970931011, 2: 0.9036358300388782, 3: 0.8706482523887603}
	Scikit-Learn	0.864407228542852
	PyCM	{0: 0.8526785714285714, 1: 0.6951816970931011, 2: 0.9036358300388782, 3: 0.8706482523887603}
Cohens Kappa	AllMetrics	0.864071560210366
	Scikit-Learn	0.864071560210366
	PyCM	0.864071560210366
F-Beta Score	AllMetrics	0.849279245158052
	Scikit-Learn	0.849279245158052
Jaccard Index	AllMetrics	{0: 0.794392523364486, 1: 0.5454545454545454, 2: 0.940983606557377, 3: 0.7692307692307693}
	Scikit-Learn	0.762515361151794
	TensorFlow	0.762515366077423
	PyCM	{0: 0.794392523364486, 1: 0.5454545454545454, 2: 0.940983606557377, 3: 0.7692307692307693}

2.2.3 2D ultrasound. We utilized a dataset comprising 2D ultrasound (US) and their corresponding synthetic Magnetic Resonance Imaging (MRI) slices for prostate cancer. A sample includes 24 patients with 1409 slices of size 128*128, generated using a Pix2Pix-based image-to-image translation model trained on paired US-MRI images. These synthetic MRIs serve as predictions, while the real MRIs are treated as ground truth for evaluation.

Supplemental Table S7. Accuracy and Reliability Comparison of Multi-Class Classification Models on the MRI Dataset

Metric	Library	Value
Accuracy	AllMetrics	0.788135593220339
	Scikit-Learn	0.788135593220339
	PyTorch	0.788135588169098
	TensorFlow	0.788135588169098
	PyCM	0.788135593220339
Precision	AllMetrics	{3: 0.9375, 4: 0.6, 5: 0.9}
	Scikit-Learn	0.8125
	PyTorch	-
	TensorFlow	1
	PyCM	{3: 0.9375, 4: 0.6, 5: 0.9}
Recall	AllMetrics	{3: 0.9, 4: 0.8823529411764706, 5: 0.5294117647058824}
	Scikit-Learn	0.770588235294118
	PyTorch	-
	TensorFlow	1
	PyCM	{3: 0.9, 4: 0.8823529411764706, 5: 0.5294117647058824}
F1 Score	AllMetrics	{3: 0.9183673469387755, 4: 0.7142857142857143, 5: 0.6666666666666666}
	Scikit-Learn	0.766439909297052
	PyTorch	-
	TensorFlow	1
	PyCM	{3: 0.9183673469387755, 4: 0.7142857142857143, 5: 0.6666666666666666}
Balanced Accuracy	AllMetrics	0.770588235294118
	Scikit-Learn	0.770588235294118
Matthews Correlation Coefficient	AllMetrics	{3: 0.8609618180272807, 4: 0.5904719382856264, 5: 0.6103234488126624}
	Scikit-Learn	0.694874065551233
	PyCM	{3: 0.8609618180272807, 4: 0.5904719382856264, 5: 0.6103234488126624}
Cohens Kappa	AllMetrics	0.677384076990376
	Scikit-Learn	0.677384076990376
	PyCM	0.677384076990376
F-Beta Score	AllMetrics	0.786750463783957
	Scikit-Learn	0.786750463783957
Jaccard Index	AllMetrics	{3: 0.8490566037735849, 4: 0.5555555555555556, 5: 0.5}
	Scikit-Learn	0.63487071977638
	TensorFlow	-
	PyCM	{3: 0.8490566037735849, 4: 0.5555555555555556, 5: 0.5}

2.3 Clustering

2.3.1 Heart Disease dataset. We used the same Heart Disease dataset that was employed for binary classification.

Supplemental Table S8. Accuracy and Reliability Comparison of Clustering Models on the Heart Disease Dataset

Metric	Library	Value
Fowlkes-Mallows Score	AllMetrics	0.508815592231628

	Scikit-Learn	0.508815592231628
V-Measure Score	AllMetrics	0.00618281999411902
	Scikit-Learn	0.00618281999411797
Completeness Score	AllMetrics	0.00625636237371619
	Scikit-Learn	0.00625636237371497
Homogeneity Score	AllMetrics	0.0061109864811818
	Scikit-Learn	0.00611098648118091
Normalized Mutual Information	AllMetrics	0.00618324719783199
	Scikit-Learn	0.00618281999411797
Mutual Info Score	AllMetrics	0.00423259156407365
	Scikit-Learn	0.00423259156407291
Adjusted Rand Score	AllMetrics	0.00395740394571725
	Scikit-Learn	0.00395740394571722
Rand Score	AllMetrics	0.501945936799357
	Scikit-Learn	0.501945936799357

2.3.2 Breast Cancer Wisconsin Dataset. We used the same Breast Cancer Wisconsin (Diagnostic) dataset that was employed for binary classification.

Supplemental Table S9. Accuracy and Reliability Comparison of Clustering Models on the Breast Cancer Dataset

Metric	Library	Value
Fowlkes-Mallows Score	AllMetrics	0.775430240967485
	Scikit-Learn	0.775430240967485
V-Measure Score	AllMetrics	0.459918789726656
	Scikit-Learn	0.459918789726656
Completeness Score	AllMetrics	0.451037388686386
	Scikit-Learn	0.451037388686386
Homogeneity Score	AllMetrics	0.469156984688718
	Scikit-Learn	0.469156984688718
Normalized Mutual Information	AllMetrics	0.46000797969518
	Scikit-Learn	0.46000797969518
Mutual Info Score	AllMetrics	0.310897768874046
	Scikit-Learn	0.310897768874046
Adjusted Rand Score	AllMetrics	0.538917538905536
	Scikit-Learn	0.538917538905536
Rand Score	AllMetrics	0.769445738239404
	Scikit-Learn	0.769445738239404

2.3.3 Online News Popularity Reduced Dataset. We used the same Online News Popularity Reduced dataset that was employed for binary classification.

Supplemental Table S10. Accuracy and Reliability Comparison of Clustering Models on the Online News Popularity Dataset

Metric	Library	Value
Fowlkes-Mallows Score	AllMetrics	0.725200170971141
	Scikit-Learn	0.725200170971141
V-Measure Score	AllMetrics	0.0153181964427894
	Scikit-Learn	0.0153181964427895
Completeness Score	AllMetrics	0.0214012766776079
	Scikit-Learn	0.021401276677608
Homogeneity Score	AllMetrics	0.0119278381269261
	Scikit-Learn	0.0119278381269262
Normalized Mutual Information	AllMetrics	0.0159772013871594
	Scikit-Learn	0.0153181964427895
Mutual Info Score	AllMetrics	0.00692297034477501
	Scikit-Learn	0.00692297034477513
Adjusted Rand Score	AllMetrics	0.061057637163006
	Scikit-Learn	0.0610576371630062
Rand Score	AllMetrics	0.595708144433535
	Scikit-Learn	0.595708144433535

2.3.4 Car Evaluation Dataset. We used the same Car Evaluation dataset that was employed for binary classification.

Supplemental Table S11. Accuracy and Reliability Comparison of Clustering Models on the Car Evaluation Dataset

Metric	Library	Value
Fowlkes-Mallows Score	AllMetrics	0.422513140301071
	Scikit-Learn	0.422513140301071
V-Measure Score	AllMetrics	0.234802793016664
	Scikit-Learn	0.234802793016664
Completeness Score	AllMetrics	0.192694558830291
	Scikit-Learn	0.192694558830291
Homogeneity Score	AllMetrics	0.30046035534405
	Scikit-Learn	0.30046035534405
Normalized Mutual Information	AllMetrics	0.240618111668583
	Scikit-Learn	0.234802793016664
Mutual Info Score	AllMetrics	0.259917147258834
	Scikit-Learn	0.259917147258834
Adjusted Rand Score	AllMetrics	0.0675224925832546
	Scikit-Learn	0.0675224925832546
Rand Score	AllMetrics	0.521665807338661
	Scikit-Learn	0.521665807338661

2.3.5 Wine Quality Dataset. We use a dataset related to red variants of Portuguese *"Vinho Verde"* wine. It includes 11 physicochemical features such as acidity, sugar content, pH, and alcohol, describing the chemical composition of each sample. The target variable is wine quality, scored between 0 and 10. The classes are imbalanced, with most instances representing average-quality wines. The goal is to predict the perceived quality of wine based on its measurable properties.

Supplemental Table S12. Accuracy and Reliability Comparison of Clustering Models on the Wine Quality Dataset

Metric	Library	Value
Fowlkes-Mallows Score	AllMetrics	0.293964704315248
	Scikit-Learn	0.293964704315248
V-Measure Score	AllMetrics	0.060498175337052
	Scikit-Learn	0.0604981753370523
Completeness Score	AllMetrics	0.0512758580486296
	Scikit-Learn	0.0512758580486298
Homogeneity Score	AllMetrics	0.0737653903739568
	Scikit-Learn	0.073765390373957
Normalized Mutual Information	AllMetrics	0.0615010868896046
	Scikit-Learn	0.0604981753370523
Mutual Info Score	AllMetrics	0.0819410070484871
	Scikit-Learn	0.0819410070484874
Adjusted Rand Score	AllMetrics	0.0255803846271359
	Scikit-Learn	0.0255803846271358
Rand Score	AllMetrics	0.586416915651574
	Scikit-Learn	0.586416915651574

2.3.6 Room Occupancy Dataset. We use a dataset based on experimental data collected from indoor environmental sensors. It contains 5 features: Temperature, Humidity, Light, and Carbon Dioxide (CO_2), along with a target variable indicating room occupancy. The target is labeled as 1 if the room is likely to be occupied and 0 if not. The dataset is useful for analyzing how environmental conditions relate to human presence in indoor spaces.

Supplemental Table S13. Accuracy and Reliability Comparison of Clustering Models on the Room Occupancy Dataset

Metric	Library	Value
Fowlkes-Mallows Score	AllMetrics	0.878008637290485
	Scikit-Learn	0.878008637290485
V-Measure Score	AllMetrics	0.61018437522509
	Scikit-Learn	0.61018437522509
Completeness Score	AllMetrics	0.606939765061789
	Scikit-Learn	0.606939765061789
Homogeneity Score	AllMetrics	0.613463862246443
	Scikit-Learn	0.613463862246443
Normalized Mutual Information	AllMetrics	0.610193094565674
	Scikit-Learn	0.61018437522509
Mutual Info Score	AllMetrics	0.395337335264829

	Scikit-Learn	0.395337335264829
Adjusted Rand Score	AllMetrics	0.732625082746845
	Scikit-Learn	0.732625082746845
Rand Score	AllMetrics	0.867327794157062
	Scikit-Learn	0.867327794157062

2.4 Regression

2.4.1 Weather Dataset. We use the weather History dataset, which contains approximately 96,000 observations and 12 features related to historical weather conditions. The features include timestamps, temperature, humidity, wind speed and direction, visibility, atmospheric pressure, and categorical weather descriptions such as precipitation type and summary.

Supplemental Table S14. Accuracy and Reliability Comparison of Regression Models on the Weather Dataset

Metric	Library	Value
Mean Absolute Error	AllMetrics	3.09026504168572
	Scikit-Learn	3.09026504168572
	TorchMetrics	3.09026479721069
	Tensorflow	3.09026503562927
Mean Squared Error	AllMetrics	15.6722940667374
	Scikit-Learn	15.6722940667374
	TorchMetrics	15.6722936630249
	Tensorflow	15.6722927093506
Mean Absolute Percentage Error	AllMetrics	71559651968089
	Scikit-Learn	71559651968089.3
	TorchMetrics	13581.021484375
	Tensorflow	-
Mean Bias Deviation	AllMetrics	-0.0251285367243777
	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Median AE	AllMetrics	2.59180657759829
	Scikit-Learn	2.59180657759829
Symmetric Mean Absolute Percentage Error	AllMetrics	23.0635078393929
	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Relative Squared Error	AllMetrics	0.227822146509458
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	-
R-squared	AllMetrics	0.772177853490543
	Scikit-Learn	0.772177853490543
	TorchMetrics	0.772177875041962
Explained Variance Score	AllMetrics	0.772187032541534
	Scikit-Learn	0.772187032541203
	TorchMetrics	0.772187113761902
Huber Loss Function	AllMetrics	2.62458453262229
	TensorFlow	2.62458467483521
Logarithm of the Hyperbolic Cosine Loss	AllMetrics	2.48038645586762
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	2.48038673400879
Maximum Absolute Error	AllMetrics	40.8512269831941
	Scikit-Learn	40.8512269831941
	TorchMetrics	-
	TensorFlow	40.8512268066406
Tweedie Deviance Score	AllMetrics	15.6722940667374
	Scikit-Learn	15.6722940667374
	TorchMetrics	-
	TensorFlow	-

2.4.2 House Dataset. We use a housing dataset consisting of 8 features related to property characteristics, including area, number of bedrooms, furnishing status, and proximity to the main road, among others. The goal is to predict house prices based on these attributes.

Supplemental Table S15. Accuracy and Reliability Comparison of Regression Models on the House Dataset

Metric	Library	Value
Mean Absolute Error	AllMetrics	1.562126568694
	Scikit-Learn	1.562126568694
	TorchMetrics	1.56212651729584
	Tensorflow	1.56212651729584
Mean Squared Error	AllMetrics	3.6392650706761
	Scikit-Learn	3.6392650706761
	TorchMetrics	3.63926506042481
	Tensorflow	3.63926506042481
Mean Absolute Percentage Error	AllMetrics	35.59005802046
	Scikit-Learn	0.355900580216156
	TorchMetrics	0.355900585651398
	Tensorflow	-
Mean Squared Logarithmic Error	AllMetrics	0.0985411458568938
	Scikit-Learn	0.0985411458568938
	TorchMetrics	0.0985411405563355
	Tensorflow	0.0985411480069161
Mean Bias Deviation	AllMetrics	0.0358957257867641
	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Median AE	AllMetrics	1.35611875393634
	Scikit-Learn	1.35611875393634
Symmetric Mean Absolute Percentage Error	AllMetrics	29.4128390496168
	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Relative Squared Error	AllMetrics	0.749304622646288
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	-
R-squared	AllMetrics	0.250695377353718
	Scikit-Learn	0.250695377353712
	TorchMetrics	0.250696182250977
Explained Variance Score	AllMetrics	0.250960673062817
	Scikit-Learn	0.250960673047395
	TorchMetrics	0.250961363315582
Huber Loss Function	AllMetrics	1.12374408708483
	TensorFlow	1.12374413013458
Logarithm of the Hyperbolic Cosine Loss	AllMetrics	1.01733732462693
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	1.01733732462693
Maximum Absolute Error	AllMetrics	6.9853439587012
	Scikit-Learn	6.9853439587012
	TorchMetrics	-
	TensorFlow	6.98534393310547
Tweedie Deviance Score	AllMetrics	3.6392650706761
	Scikit-Learn	3.6392650706761
	TorchMetrics	-
	TensorFlow	-

2.4.3 Head and Neck Dataset. We use the CT modality from the Head and Neck (CITISCAN) dataset.

Supplemental Table S16. Accuracy and Reliability Comparison of Regression Models on the CT Head and Neck Dataset

Metric	Library	Value
Mean Absolute Error	AllMetrics	693.101685393258
	Scikit-Learn	693.101685393258

	TorchMetrics	693.101623535156
	Tensorflow	693.101684570313
Mean Squared Error	AllMetrics	723329.200869663
	Scikit-Learn	723329.200869663
	TorchMetrics	723329.1875
	Tensorflow	723329.1875
Mean Absolute Percentage Error	AllMetrics	71.7676895873158
	Scikit-Learn	0.717676895873312
	TorchMetrics	0.717676937580109
	Tensorflow	-
Mean Squared Logarithmic Error	AllMetrics	0.399902194265692
	Scikit-Learn	0.399902194265692
	TorchMetrics	0.399902164936066
	Tensorflow	0.39990222454071
Mean Bias Deviation	AllMetrics	-4.66955056179775
	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Median AE	AllMetrics	601.125
	Scikit-Learn	601.125
Symmetric Mean Absolute Percentage Error	AllMetrics	43.4693435769297
	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Relative Squared Error	AllMetrics	0.996398669458336
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	-
R-squared	AllMetrics	0.00360133054166378
	Scikit-Learn	0.00360133054166378
	TorchMetrics	0.00360107421875
Explained Variance Score	AllMetrics	0.00363136690198518
	Scikit-Learn	0.00363136690198507
	TorchMetrics	0.00363129377365112
Huber Loss Function	AllMetrics	692.601685393258
	TensorFlow	692.601684570313
Logarithm of the Hyperbolic Cosine Loss	AllMetrics	Inf
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	inf
Maximum Absolute Error	AllMetrics	2267.08
	Scikit-Learn	2267.08
	TorchMetrics	-
	TensorFlow	2267.080078125
Tweedie Deviance Score	AllMetrics	723329.200869663
	Scikit-Learn	723329.200869663
	TorchMetrics	-
	TensorFlow	-

2.5 Segmentation

2.5.1 3D segmentation. For 3D segmentation evaluation, we utilize a validated collection of 240 lung CT tumor segmentations sourced from eight independent multi-center datasets, including: LCTSC (#35), Lung CT Diagnosis (#49), NSCLC-Radiogenomics (#39), QIN LUNG CT (#38), Lung-Fused-CT-Pathology (#6), RIDER (#16), SPIE-AAPM Lung Challenge (#44), and TCGA (#13).

Supplemental Table S17. Accuracy and Reliability Comparison of 3D Segmentation Models on the Lung CT Tumor

Metric	Library	Value
Dice Score	AllMetrics	0.3056159819
	scikit-learn	0.3056159819
	PyTorch	0.3056159824
	TensorFlow	0.3056159616
	MedPy	0.3056159819
	MONAI	0.3056159824

IoU	AllMetrics	0.1984215915
	scikit-learn	0.1984215914
	PyTorch	0.1984215885
	TensorFlow	0.1984215826
	MedPy	0.1984215914
	MONAI	0.1984215885
Sensitivity	AllMetrics	0.9959124976
	scikit-learn	0.9959124976
	PyTorch	-
	TensorFlow	-
	MedPy	0.9959124976
	MONAI	-
Specificity	AllMetrics	0.8923930645
	scikit-learn	0.8923930645
	PyTorch	-
	TensorFlow	-
	MedPy	0.8923930645
	MONAI	-
Precision	AllMetrics	0.1986794051
	scikit-learn	0.1986794100
	PyTorch	-
	TensorFlow	-
	MedPy	0.1986794051
	MONAI	-
Hausdorff	AllMetrics	35.94114328
	scikit-learn	-
	PyTorch	-
	TensorFlow	-
	MedPy	35.94114328
	MONAI	22.43082142

2.5.2 2D Segmentation. To evaluate segmentation metrics across libraries, we employ two synthetic 2D datasets designed to simulate common challenges in medical image analysis.

Dataset 1: Complex Overlapping Shapes
Comprising 256×256 images with 3–7 randomly placed geometric shapes (circles, rectangles, ellipses, polygons), this dataset simulates boundary noise, affine distortions, and partial occlusions. Predictions include random affine transformations, 5% noise, and 30% missing regions (30×30 pixels).

Dataset 2: Textured Anatomical Structures
This dataset includes biologically inspired shapes (e.g., kidney-like structures) with internal textures (~30% pixel density). Prediction artifacts include Gaussian blur ($\sigma = 1$), 3% false positives, and 40% random pixel dropouts.

Both datasets generate binary masks (uint8 NumPy arrays) with controlled statistical properties and foreground occupancy, supporting reproducible and privacy-preserving evaluation of segmentation performance.

Supplemental Table S18. Accuracy and Reliability Comparison of 2D Segmentation Models on the Complex Overlapping Shapes

Metric	Library	Value
Dice Score	AllMetrics	0.739394972050
	scikit-learn	0.739394972036
	PyTorch	0.739394968748
	TensorFlow	0.739394962788
	MedPy	0.739394972036
	MONAI	0.739394968748
IoU	AllMetrics	0.612176970426
	scikit-learn	0.612176970397
	PyTorch	0.612176972032
	TensorFlow	0.612176954746
	MedPy	0.612176954746
	MONAI	0.612176972032
Sensitivity	AllMetrics	0.844163576616
	scikit-learn	0.844163576597
	PyTorch	-

	TensorFlow	-
	MedPy	0.844163576597
	MONAI	-
Specificity	AllMetrics	0.922501787070
	scikit-learn	0.922501787069
	PyTorch	-
	TensorFlow	-
	MedPy	0.922501787069
	MONAI	-
Precision	AllMetrics	0.663045481808
	scikit-learn	0.663045481778
	PyTorch	-
	TensorFlow	-
	MedPy	0.663045481778
	MONAI	-
Hausdorff	AllMetrics	123.133121785085
	scikit-learn	-
	PyTorch	-
	TensorFlow	-
	MedPy	123.133121785085
	MONAI	86.236210021973

Supplemental Table S19. Accuracy and Reliability Comparison of 2D Segmentation Models on the Textured Anatomical Structures

Metric	Library	Value
Dice Score	AllMetrics	0.324286747424
	scikit-learn	0.324286747233
	PyTorch	0.324286749065
	TensorFlow	0.324286729097
	MedPy	0.324286747233
	MONAI	0.324286749065
IoU	AllMetrics	0.207500312212
	scikit-learn	0.207500311954
	PyTorch	0.207500311956
	TensorFlow	0.207500308752
	MedPy	0.207500311954
	MONAI	0.207500311956
Sensitivity	AllMetrics	0.559939210894
	scikit-learn	0.559939210458
	PyTorch	-
	TensorFlow	-
	MedPy	0.559939210458
	MONAI	-
Specificity	AllMetrics	0.969495466909
	scikit-learn	0.969495466908
	PyTorch	-
	TensorFlow	-
	MedPy	0.969495466908
	MONAI	-
Precision	AllMetrics	0.234121522403
	scikit-learn	0.234121522095
	PyTorch	-
	TensorFlow	-
	MedPy	0.234121522095
	MONAI	-
Hausdorff	AllMetrics	142.738365295379
	scikit-learn	-
	PyTorch	-
	TensorFlow	-
	MedPy	142.738365295379
	MONAI	111.983593292236

2.6. Image to Image Translation

We utilized a dataset comprising 2D ultrasound (US) and their corresponding synthetic Magnetic Resonance Imaging (MRI) slices for prostate cancer. A sample includes 24 patients with 1409 slices of size 128*128, generated using a

Pix2Pix-based image-to-image translation model trained on paired US-MRI images. These synthetic MRIs serve as predictions, while the real MRIs are treated as ground truth for evaluation.

Additionally, a computer-generated benchmark dataset includes grayscale image pairs (ground truth and noisy variants) for standardized image quality assessment, with ground truth images being random 8-bit (64×64 pixels) and predictions as noise-corrupted versions using Gaussian noise (μ=0, σ=30). Pixel values are scaled to [0,1], and the dataset supports configurable dimensions, adjusSupplemental Table Snoise levels, and deterministic generation for reproducible benchmarking of metrics like PSNR and SSIM across Python imaging libraries.

Supplemental Table S20. Accuracy and Reliability Comparison of Image To Image Translation on 2D Ultrasound-MRI for Prostate Cancer

Metric	Library	Value
SSIM	AllMetrics	0.1433720633
	Scikit-Image	0.1051427986
	PyTorch	0.1339276796
	TensorFlow	0.1273871213
PSNR	AllMetrics	4.438936710
	Scikit-Image	4.438936805
	PyTorch	4.438936799
	TensorFlow	4.438936710

Supplemental Table S21. Accuracy and Reliability Comparison of Image to Image Translation on computer-generated Synthetic Noisy Image Pairs

Metric	Library	Value
SSIM	AllMetrics	0.9325634397
	Scikit-Image	0.9303055623
	PyTorch	0.9285293818
	TensorFlow	0.9288128614
PSNR	AllMetrics	19.28897095
	Scikit-Image	19.28897030
	PyTorch	19.28897047
	TensorFlow	19.28897095

3. Efficiency

This section presents an additional perspective of our evaluation, with a focus on computational efficiency. We compare the runtime performance of AllMetrics against existing libraries to assess its speed and resource utilization.

The results for different datasets set in below tables related to specific dataset.

Supplemental Table S22. Runtime Comparison for Binary Classification on the Heart Disease Dataset

Metric	Library	Time (s)
Accuracy	AllMetrics	0.000347137451171875
	Scikit-Learn	0.00112080574035645
	PyTorch	0.00348711013793945
	TensorFlow	0.0403339862823486
	PyCM	0.0309188365936279
Precision	AllMetrics	0.000417470932006836
	Scikit-Learn	0.00342750549316406
	PyTorch	0.000952005386352539
	TensorFlow	0.0229766368865967
	PyCM	0.0075833797454834
Recall	AllMetrics	0.0004119873046875
	Scikit-Learn	0.0140869617462158
	PyTorch	0.00101184844970703
	TensorFlow	0.0704507827758789
	PyCM	0.00752997398376465
F1 Score	AllMetrics	0.00063323974609375
	Scikit-Learn	0.00417375564575195
	PyTorch	0.00130987167358398

	TensorFlow	0.103707551956177
	PyCM	0.00882792472839356
Balanced Accuracy	AllMetrics	0.000375747680664063
	Scikit-Learn	0.00158882141113281
Matthews Correlation Coefficient	AllMetrics	0.000883340835571289
	Scikit-Learn	0.00315451622009277
	PyCM	0.00752043724060059
Cohens Kappa	AllMetrics	0.000816583633422852
	Scikit-Learn	0.0016179084777832
	PyCM	0.00763344764709473
F-Beta Score	AllMetrics	0.000420570373535156
	Scikit-Learn	0.00351190567016602
Jaccard Index	AllMetrics	0.000441789627075195
	Scikit-Learn	0.00319957733154297
	TensorFlow	0.0398294925689697
	PyCM	0.00772166252136231

Supplemental Table S23. Runtime Comparison for Binary Classification on the Breast Cancer Dataset

Metric	Library	Time (s)
Accuracy	AllMetrics	0.000243902206420898
	Scikit-Learn	0.00117063522338867
	PyTorch	0.00120878219604492
	TensorFlow	0.0241031646728516
	PyCM	0.00362348556518555
Precision	AllMetrics	0.000392675399780273
	Scikit-Learn	0.0230166912078857
	PyTorch	0.00546956062316895
	TensorFlow	0.0987479686737061
	PyCM	0.0112321376800537
Recall	AllMetrics	0.000382661819458008
	Scikit-Learn	0.00733780860900879
	PyTorch	0.00149774551391602
	TensorFlow	0.089756965637207
	PyCM	0.0098259449005127
F1 Score	AllMetrics	0.00478839874267578
	Scikit-Learn	0.0207140445709229
	PyTorch	0.0124413967132568
	TensorFlow	0.179150581359863
	PyCM	0.00172233581542969
Balanced Accuracy	AllMetrics	0.0168275833129883
	Scikit-Learn	0.0103757381439209
Matthews Correlation Coefficient	AllMetrics	0.000581264495849609
	Scikit-Learn	0.0318403244018555
	PyCM	0.00708150863647461
Cohens Kappa	AllMetrics	0.00051569938659668
	Scikit-Learn	0.00160026550292969
	PyCM	0.01279616355896
F-Beta Score	AllMetrics	0.00040435791015625
	Scikit-Learn	0.00581479072570801
Jaccard Index	AllMetrics	0.000471591949462891
	Scikit-Learn	0.0119214057922363
	TensorFlow	0.0880427360534668
	PyCM	0.00274491310119629

Supplemental Table S24. Runtime Comparison for Binary Classification on the Online News Popularity Reduced Dataset

Metric	Library	Time (s)
Accuracy	AllMetrics	0.000272750854492188
	Scikit-Learn	0.000732898712158203
	PyTorch	0.00125694274902344
	TensorFlow	0.00342893600463867
	PyCM	0.0136573314666748
Precision	AllMetrics	0.000329732894897461
	Scikit-Learn	0.0024416446685791
	PyTorch	0.00079345703125

	TensorFlow	0.0133674144744873
	PyCM	0.0127029418945313
Recall	AllMetrics	0.000299692153930664
	Scikit-Learn	0.0022120475769043
	PyTorch	0.000805139541625977
	TensorFlow	0.0127439498901367
	PyCM	0.0132308006286621
F1 Score	AllMetrics	0.000464916229248047
	Scikit-Learn	0.00236415863037109
	PyTorch	0.000917434692382813
	TensorFlow	0.0262441635131836
	PyCM	0.013054370880127
Balanced Accuracy	AllMetrics	0.000318050384521484
	Scikit-Learn	0.00121378898620605
Matthews Correlation Coefficient	AllMetrics	0.0011293888092041
	Scikit-Learn	0.00243592262268066
	PyCM	0.0130164623260498
Cohens Kappa	AllMetrics	0.00119900703430176
	Scikit-Learn	0.00111556053161621
	PyCM	0.0133411884307861
F-Beta Score	AllMetrics	0.000347614288330078
	Scikit-Learn	0.0028526782989502
Jaccard Index	AllMetrics	0.000443696975708008
	Scikit-Learn	0.00267720222473145
	TensorFlow	0.0183439254760742
	PyCM	0.0127668380737305

Supplemental Table S25. Runtime Comparison for Multi-Class Classification on the Iris Dataset

Metric	Library	Time (s)
Accuracy	AllMetrics	0.00131726264953613
	Scikit-Learn	0.00107979774475098
	PyTorch	0.00159239768981934
	TensorFlow	0.00630688667297363
	PyCM	0.000925302505493164
Precision	AllMetrics	0.00161242485046387
	Scikit-Learn	0.00392818450927734
	PyTorch	0.00144839286804199
	TensorFlow	0.0261349678039551
	PyCM	0.000757455825805664
Recall	AllMetrics	0.000451564788818359
	Scikit-Learn	0.00310349464416504
	PyTorch	0.0010836124420166
	TensorFlow	0.0220322608947754
	PyCM	0.0007781982421875
F1 Score	AllMetrics	0.000775337219238281
	Scikit-Learn	0.00380206108093262
	PyTorch	0.00181818008422852
	TensorFlow	0.0698540210723877
	PyCM	0.00072026252746582
Balanced Accuracy	AllMetrics	0.000448465347290039
	Scikit-Learn	0.00167942047119141
Matthews Correlation Coefficient	AllMetrics	0.000447988510131836
	Scikit-Learn	0.0130820274353027
	PyCM	0.000194787979125977
Cohens Kappa	AllMetrics	0.000493764877319336
	Scikit-Learn	0.00147342681884766
	PyCM	0.00077056884765625
F-Beta Score	AllMetrics	0.000409603118896484
	Scikit-Learn	0.00315213203430176
Jaccard Index	AllMetrics	0.000450372695922852
	Scikit-Learn	0.00337958335876465
	TensorFlow	0.0368115901947022
	PyCM	0.00078582763671875

Supplemental Table S26. Runtime Comparison for Multi-Class Classification on the Car Evaluation Dataset

Metric	Library	Time (s)
Accuracy	AllMetrics	0.000275135040283203
	Scikit-Learn	0.000977754592895508
	PyTorch	0.00826001167297363
	TensorFlow	0.0178611278533936
	PyCM	0.02071213722229
Precision	AllMetrics	0.000552892684936523
	Scikit-Learn	0.00394058227539063
	PyTorch	0.00163602828979492
	TensorFlow	0.0283169746398926
	PyCM	0.0164015293121338
Recall	AllMetrics	0.000594615936279297
	Scikit-Learn	0.00388073921203613
	PyTorch	0.00144243240356445
	TensorFlow	0.0231454372406006
	PyCM	0.0163133144378662
F1 Score	AllMetrics	0.000837802886962891
	Scikit-Learn	0.00336241722106934
	PyTorch	0.0012812614440918
	TensorFlow	0.0485951900482178
	PyCM	0.0164587497711182
Balanced Accuracy	AllMetrics	0.000694751739501953
	Scikit-Learn	0.00170445442199707
Matthews Correlation Coefficient	AllMetrics	0.00109338760375977
	Scikit-Learn	0.00346255302429199
	PyCM	0.0344085693359375
Cohens Kappa	AllMetrics	0.00101494789123535
	Scikit-Learn	0.00165724754333496
	PyCM	0.0379264354705811
F-Beta Score	AllMetrics	0.00062251091003418
	Scikit-Learn	0.00333118438720703
Jaccard Index	AllMetrics	0.00069737434387207
	Scikit-Learn	0.0102458000183106
	TensorFlow	0.0344150066375732
	PyCM	0.0154941082000732

Supplemental Table S27. Runtime Comparison for Multi-Class Classification on the MRI Dataset

Metric	Library	Time (s)
Accuracy	AllMetrics	0.0019371509552002
	Scikit-Learn	0.00148653984069824
	PyTorch	0.000989437103271484
	TensorFlow	0.00631213188171387
	PyCM	0.00175952911376953
Precision	AllMetrics	0.000530004501342773
	Scikit-Learn	0.00495147705078125
	PyTorch	
	TensorFlow	0.0318925380706787
	PyCM	0.00253510475158691
Recall	AllMetrics	0.000626087188720703
	Scikit-Learn	0.00335502624511719
	PyTorch	-
	TensorFlow	0.0188195705413818
	PyCM	0.00161147117614746
F1 Score	AllMetrics	0.000659465789794922
	Scikit-Learn	0.00364470481872559
	PyTorch	-
	TensorFlow	0.038097620010376
	PyCM	0.00192117691040039
Balanced Accuracy	AllMetrics	0.000529766082763672
	Scikit-Learn	0.00185942649841309
Matthews Correlation Coefficient	AllMetrics	0.000633955001831055
	Scikit-Learn	0.00277853012084961

	PyCM	0.00163173675537109
Cohens Kappa	AllMetrics	0.000519990921020508
	Scikit-Learn	0.00184130668640137
	PyCM	0.000189304351806641
F-Beta Score	AllMetrics	0.00046849250793457
	Scikit-Learn	0.00301957130432129
Jaccard Index	AllMetrics	0.000438451766967773
	Scikit-Learn	0.00244331359863281
	TensorFlow	-
	PyCM	0.00246477127075195

Supplemental Table S28. Runtime Comparison for Clustering on the Heart Disease

Metric	Library	Time (s)
Fowlkes-Mallows Score	AllMetrics	0.00573968887329102
	Scikit-Learn	0.00573968887329102
V-Measure Score	AllMetrics	0.00075221061706543
	Scikit-Learn	0.00261878967285156
Completeness Score	AllMetrics	0.000327110290527344
	Scikit-Learn	0.00238609313964844
Homogeneity Score	AllMetrics	0.000277519226074219
	Scikit-Learn	0.00941085815429688
Normalized Mutual Information	AllMetrics	0.000555753707885742
	Scikit-Learn	0.0064089298248291
Mutual Info Score	AllMetrics	0.000293254852294922
	Scikit-Learn	0.00219202041625977
Adjusted Rand Score	AllMetrics	0.000432729721069336
	Scikit-Learn	0.0139811038970947
Rand Score	AllMetrics	0.0855348110198975
	Scikit-Learn	0.0102057456970215

Supplemental Table S29. Runtime Comparison for Clustering on the Breast Cancer

Metric	Library	Time (s)
Fowlkes-Mallows Score	AllMetrics	0.000964641571044922
	Scikit-Learn	0.00304317474365234
V-Measure Score	AllMetrics	0.0005645751953125
	Scikit-Learn	0.00282025337219238
Completeness Score	AllMetrics	0.000373125076293945
	Scikit-Learn	0.00429344177246094
Homogeneity Score	AllMetrics	0.00072932243347168
	Scikit-Learn	0.0026092529296875
Normalized Mutual Information	AllMetrics	0.000413656234741211
	Scikit-Learn	0.00253868103027344
Mutual Info Score	AllMetrics	0.000304937362670898
	Scikit-Learn	0.00173759460449219
Adjusted Rand Score	AllMetrics	0.000253200531005859
	Scikit-Learn	0.00120377540588379
Rand Score	AllMetrics	0.00349879264831543
	Scikit-Learn	0.00125885009765625

Supplemental Table S30. Runtime Comparison for Clustering on the Online News Popularity

Metric	Library	Time (s)
Fowlkes-Mallows Score	AllMetrics	0.000864028930664063
	Scikit-Learn	0.00198054313659668
V-Measure Score	AllMetrics	0.000418663024902344
	Scikit-Learn	0.00191926956176758
Completeness Score	AllMetrics	0.000224113464355469
	Scikit-Learn	0.00162649154663086
Homogeneity Score	AllMetrics	0.000204801559448242
	Scikit-Learn	0.00161170959472656

Normalized Mutual Information	AllMetrics	0.000337600708007813
	Scikit-Learn	0.00172328948974609
Mutual Info Score	AllMetrics	0.000200271606445313
	Scikit-Learn	0.00161647796630859
Adjusted Rand Score	AllMetrics	0.00025629997253418
	Scikit-Learn	0.00105762481689453
Rand Score	AllMetrics	0.415280818939209
	Scikit-Learn	0.00156283378601074

Supplemental Table S31. Runtime Comparison for Clustering on the Car Evaluation Clustering

Metric	Library	Time (s)
Fowlkes-Mallows Score	AllMetrics	0.001033
	Scikit-Learn	0.005610
V-Measure Score	AllMetrics	0.0010988712310791
	Scikit-Learn	0.00318026542663574
Completeness Score	AllMetrics	0.000512123107910156
	Scikit-Learn	0.00235962867736816
Homogeneity Score	AllMetrics	0.000483989715576172
	Scikit-Learn	0.0024712085723877
Normalized Mutual Information	AllMetrics	0.000677108764648438
	Scikit-Learn	0.00272274017333984
Mutual Info Score	AllMetrics	0.000489711761474609
	Scikit-Learn	0.00200533866882324
Adjusted Rand Score	AllMetrics	0.00067591667175293
	Scikit-Learn	0.00182509422302246
Rand Score	AllMetrics	0.0938005447387695
	Scikit-Learn	0.00240206718444824

Supplemental Table S32. Runtime Comparison for Clustering on the Wine Quality Dataset

Metric	Library	Time (s)
Fowlkes-Mallows Score	AllMetrics	0.001976
	Scikit-Learn	0.004192
V-Measure Score	AllMetrics	0.001920
	Scikit-Learn	0.003950
Completeness Score	AllMetrics	0.000953
	Scikit-Learn	0.003052
Homogeneity Score Normalized Mutual Information	AllMetrics	0.000936
	Scikit-Learn	0.002971
	AllMetrics	0.001320
	Scikit-Learn	0.003560
Mutual Info Score	AllMetrics	0.000848
	Scikit-Learn	0.002532
Adjusted Rand Score	AllMetrics	0.001363
	Scikit-Learn	0.002116
Rand Score	AllMetrics	0.026585
	Scikit-Learn	0.002362

Supplemental Table S33. Runtime Comparison for Clustering on the Room Occupancy Dataset

Metric	Library	Time (s)
Fowlkes-Mallows Score	AllMetrics	0.000780
	Scikit-Learn	0.003098
V-Measure Score	AllMetrics	0.000642
	Scikit-Learn	0.009948
Completeness Score	AllMetrics	0.000362
	Scikit-Learn	0.002479
Homogeneity Score Normalized Mutual Information	AllMetrics	0.000328
	Scikit-Learn	0.013636
	AllMetrics	0.000791
	Scikit-Learn	0.003863
Mutual Info Score	AllMetrics	0.008247
	Scikit-Learn	0.002269
Adjusted Rand Score	AllMetrics	0.000440
	Scikit-Learn	0.001865
Rand Score	AllMetrics	0.172333

	Scikit-Learn	0.002171

Supplemental Table S34. Runtime Comparison for Clustering on the Weather Dataset

Metric	Library	Time(s)
Mean Absolute Error	AllMetrics	0.000790596008300781
	Scikit-Learn	0.00104546546936035
	TorchMetrics	0.00153851509094238
	Tensorflow	0.00155258178710938
Mean Squared Error	AllMetrics	0.00062870979309082
	Scikit-Learn	0.0112264156341553
	TorchMetrics	0.0016942024230957
	Tensorflow	0.00110650062561035
Mean Absolute Percentage Error	AllMetrics	-
	Scikit-Learn	0.0011744499206543
	TorchMetrics	0.00593900680541992
	Tensorflow	-
Mean Bias Deviation	AllMetrics	0.000711917877197266
	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Median AE	AllMetrics	0.0195522308349609
	Scikit-Learn	0.00150465965270996
Symmetric Mean Absolute Percentage Error	AllMetrics	0.000742435455322266
	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Relative Squared Error	AllMetrics	0.0226497650146484
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	-
R-squared	AllMetrics	0.000886917114257813
	Scikit-Learn	0.0165190696716309
	TorchMetrics	0.0123879909515381
Explained Variance Score	AllMetrics	0.00102114677429199
	Scikit-Learn	0.0137367248535156
	TorchMetrics	0.0121393203735352
Huber Loss Function	AllMetrics	0.00116109848022461
	TensorFlow	0.0309348106384277
Logarithm of the Hyperbolic Cosine Loss	AllMetrics	0.01934814453125
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	0.0101869106292725
Maximum Absolute Error	AllMetrics	0.000751733779907227
	Scikit-Learn	0.000919818878173828
	TorchMetrics	-
	TensorFlow	0.0174417495727539
Tweedie Deviance Score	AllMetrics	0.000723838806152344
	Scikit-Learn	0.0010535717010498
	TorchMetrics	-
	TensorFlow	-

Supplemental Table S35. Runtime Comparison for Clustering on the House Dataset

Metric	Library	Time(s)
Mean Absolute Error	AllMetrics	0.000587
	Scikit-Learn	0.001040
	TorchMetrics	0.001647
	Tensorflow	0.001647
Mean Squared Error	AllMetrics	0.000356
	Scikit-Learn	0.000892
	TorchMetrics	0.001552
	Tensorflow	0.000914
Mean Absolute Percentage Error	AllMetrics	0.000414
	Scikit-Learn	0.000920
	TorchMetrics	0.002932

	Tensorflow	-
Mean Squared Logarithmic Error	AllMetrics	0.000591
	Scikit-Learn	0.001319
	TorchMetrics	0.001513
	Tensorflow	0.001388
Mean Bias Deviation	AllMetrics	0.000381
	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Median AE	AllMetrics	0.000423
	Scikit-Learn	0.006073
Symmetric Mean Absolute Percentage Error	AllMetrics	0.000341
	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Relative Squared Error	AllMetrics	0.000277
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	-
R-squared	AllMetrics	0.000309
	Scikit-Learn	0.001068
	TorchMetrics	0.007840
Explained Variance Score	AllMetrics	0.000537
	Scikit-Learn	0.000997
	TorchMetrics	0.031442
Huber Loss Function	AllMetrics	0.001631
	TensorFlow	0.017705
Logarithm of the Hyperbolic Cosine Loss	AllMetrics	0.000622
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	0.001121
Maximum Absolute Error	AllMetrics	0.000341
	Scikit-Learn	0.000710
	TorchMetrics	-
	TensorFlow	0.000837
Tweedie Deviance Score	AllMetrics	0.000837
	Scikit-Learn	0.000786
	TorchMetrics	-
	TensorFlow	-

Supplemental Table S36. Runtime Comparison for Clustering on the CT Head and Neck Dataset

Metric	Library	Time(s)
Mean Absolute Error	AllMetrics	0.000383
	Scikit-Learn	0.000965
	TorchMetrics	0.001417
	Tensorflow	0.005392
Mean Squared Error	AllMetrics	0.000391
	Scikit-Learn	0.000862
	TorchMetrics	0.001767
	Tensorflow	0.000747
Mean Absolute Percentage Error	AllMetrics	0.000377
	Scikit-Learn	0.000786
	TorchMetrics	0.001410
	Tensorflow	-
Mean Squared Logarithmic Error	AllMetrics	0.000336
	Scikit-Learn	0.001223
	TorchMetrics	0.001503
	Tensorflow	0.001090
Mean Bias Deviation	AllMetrics	0.000310
	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Median AE	AllMetrics	0.000256
	Scikit-Learn	0.000577
Symmetric Mean Absolute Percentage Error	AllMetrics	0.000228

	Scikit-Learn	-
	TorchMetrics	-
	Tensorflow	-
Relative Squared Error	AllMetrics	0.003895
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	-
R-squared	AllMetrics	0.000323
	Scikit-Learn	0.001076
	TorchMetrics	0.029328
Explained Variance Score	AllMetrics	0.000576
	Scikit-Learn	0.006355
	TorchMetrics	0.005081
Huber Loss Function	AllMetrics	0.000446
	TensorFlow	0.003698
Logarithm of the Hyperbolic Cosine Loss	AllMetrics	0.000998
	Scikit-Learn	-
	TorchMetrics	-
	TensorFlow	0.000970
Maximum Absolute Error	AllMetrics	0.000366
	Scikit-Learn	0.000615
	TorchMetrics	-
	TensorFlow	0.000809
Tweedie Deviance Score	AllMetrics	0.000280
	Scikit-Learn	0.000760
	TorchMetrics	-
	TensorFlow	-

Supplemental Table S37. Runtime Comparison for 3D Segmentation

Metric	Library	Time
Dice Score	AllMetrics	0.000968
	scikit-learn	0.000006
	PyTorch	0.000511
	TensorFlow	0.004973
	MedPy	0.000218
	MONAI	0.016017
IoU	AllMetrics	0.001027
	scikit-learn	0.000001
	PyTorch	0.00003
	TensorFlow	0.001118
	MedPy	0.000155
	MONAI	0.000998
Sensitivity	AllMetrics	0.007555
	scikit-learn	0.000001
	PyTorch	-
	TensorFlow	-
	MedPy	0.000194
	MONAI	-
Specificity	AllMetrics	0.008504
	scikit-learn	0.000001
	PyTorch	-
	TensorFlow	-
	MedPy	0.00023
	MONAI	-
Precision	AllMetrics	0.007450
	scikit-learn	0.000002
	PyTorch	-
	TensorFlow	-
	MedPy	0.000156
	MONAI	-
Hausdorff	AllMetrics	0.444600
	scikit-learn	-
	PyTorch	-
	TensorFlow	-
	MedPy	0.103074

	MONAI	0.107905

Supplemental Table S38. Runtime Comparison for 2D Segmentation on the Complex Overlapping Shapes

Metric	Library	Time
Dice Score	AllMetrics	0.000469
	scikit-learn	0.000005
	PyTorch	0.000258
	TensorFlow	0.002037
	MedPy	0.000111
	MONAI	0.001826
IoU	AllMetrics	0.000305
	scikit-learn	0.000001
	PyTorch	0.000028
	TensorFlow	0.000354
	MedPy	0.000040
	MONAI	0.000778
Sensitivity	AllMetrics	0.002797
	scikit-learn	0.000001
	PyTorch	-
	TensorFlow	-
	MedPy	0.000054
	MONAI	-
Specificity	AllMetrics	0.00309
	scikit-learn	0.000001
	PyTorch	-
	TensorFlow	-
	MedPy	0.00006
	MONAI	-
Precision	AllMetrics	0.002889
	scikit-learn	0.000003
	PyTorch	-
	TensorFlow	-
	MedPy	0.000037
	MONAI	-
Hausdorff	AllMetrics	0.144313
	scikit-learn	-
	PyTorch	-
	TensorFlow	-
	MedPy	0.015555
	MONAI	0.027129

Supplemental Table S39. Runtime Comparison for 2D Segmentation on the Textured Anatomical Structures

Metric	Library	Time
Dice Score	AllMetrics	0.000596
	scikit-learn	0.000005
	PyTorch	0.000298
	TensorFlow	0.002312
	MedPy	0.000138
	MONAI	0.001927
IoU	AllMetrics	0.000421
	scikit-learn	0.000001
	PyTorch	0.000038
	TensorFlow	0.000546
	MedPy	0.000064
	MONAI	0.000837
Sensitivity	AllMetrics	0.003863
	scikit-learn	0.000001
	PyTorch	-
	TensorFlow	-
	MedPy	0.00007
	MONAI	-
Specificity	AllMetrics	0.003696
	scikit-learn	0.000001
	PyTorch	-

	TensorFlow	-
	MedPy	0.000074
	MONAI	-
Precision	AllMetrics	0.003704
	scikit-learn	0.000003
	PyTorch	-
	TensorFlow	-
	MedPy	0.000059
	MONAI	-
Hausdorff	AllMetrics	0.010361
	scikit-learn	-
	PyTorch	-
	TensorFlow	-
	MedPy	0.019720
	MONAI	0.031599

Supplemental Table S40. Runtime Comparison for Image To Image Translation on 2D Ultrasound-MRI for Prostate Cancer

Metric	Library	Time
SSIM	AllMetrics	4.706770920
	Scikit-Image	1.491764740
	PyTorch	12.00770435
	TensorFlow	105.7376791
PSNR	AllMetrics	0.18661660
	Scikit-Image	0.17459364
	PyTorch	0.23110290
	TensorFlow	3.53893665

Supplemental Table S41. Runtime Comparison for Image to Image Translation on computer-generated Synthetic Noisy Image Pairs

Metric	Library	Time
SSIM	AllMetrics	2.1338462830
	Scikit-Image	0.5930423737
	PyTorch	4.2533397670
	TensorFlow	33.786582950
PSNR	AllMetrics	0.1774787903
	Scikit-Image	0.1215934753
	PyTorch	0.2395629883
	TensorFlow	3.5048484800