National Edge AI Hub

Euclidean and Poincaré space ensemble Xgboost

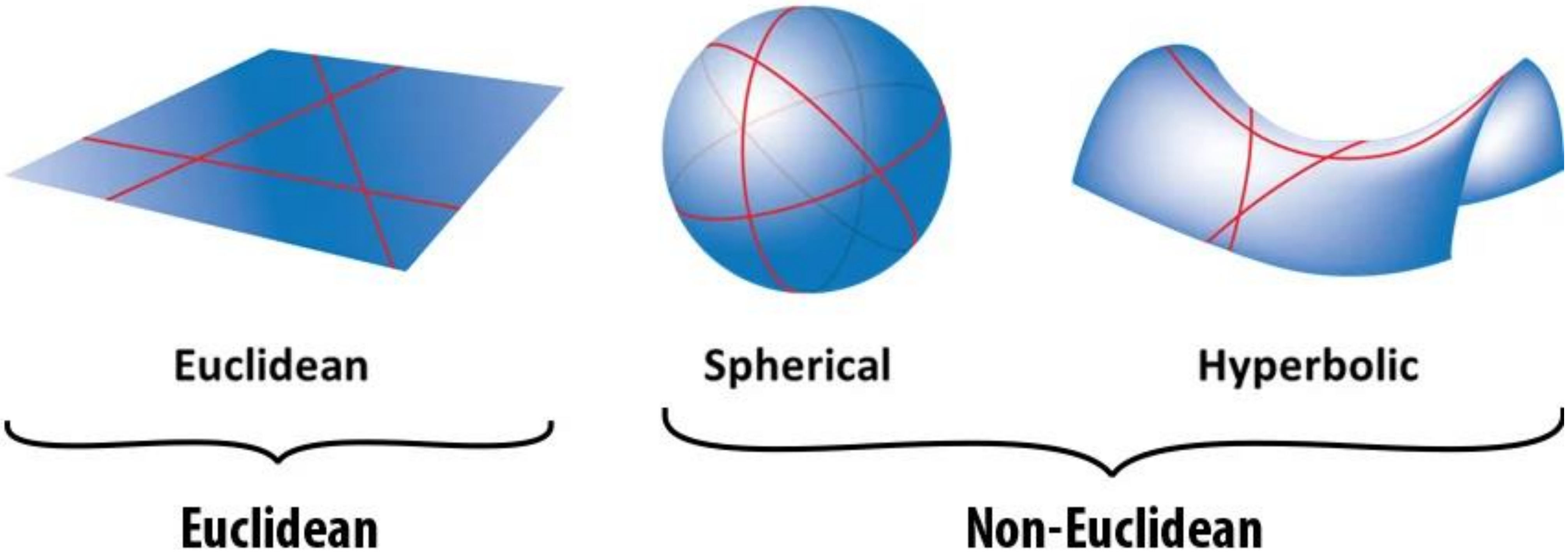
Varun Ojha
National Edge AI Hub

School of Computing, Newcastle University

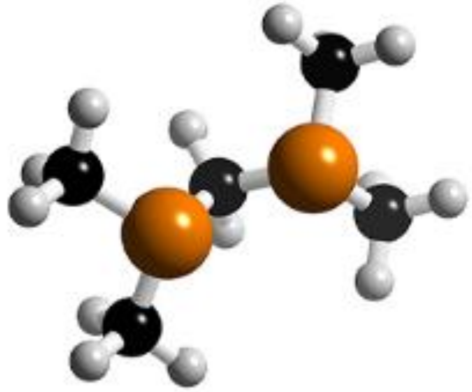
UKAIRS
8-9 September 2025

https://doi.org/10.1016/j.inffus.2024.102746

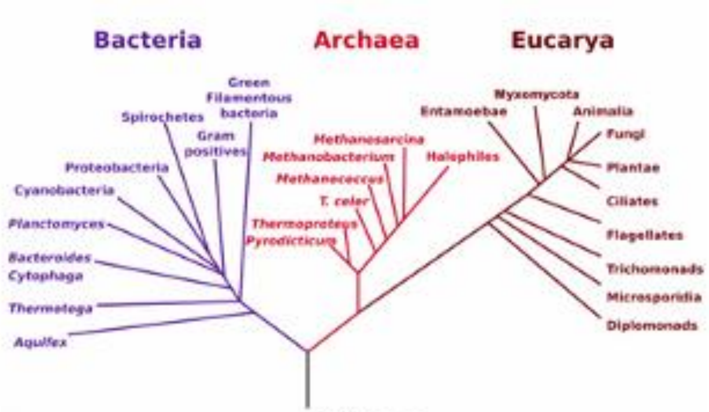
Euclidean and Non-Euclidean data



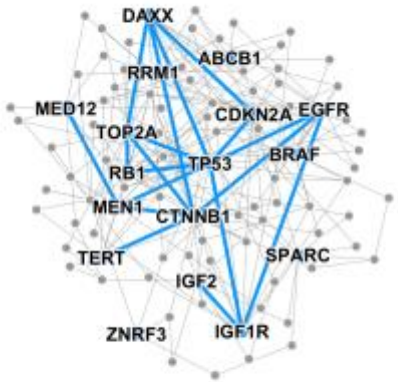
Non-Euclidean data is all around us



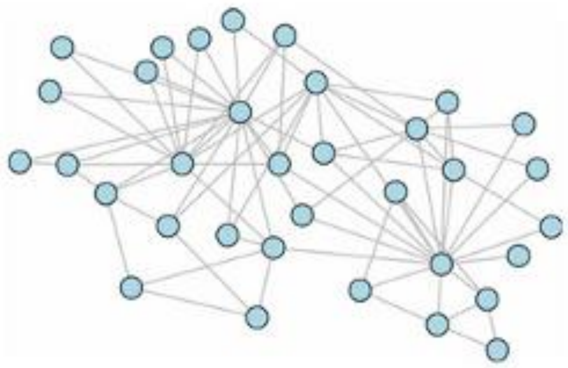
Molecules



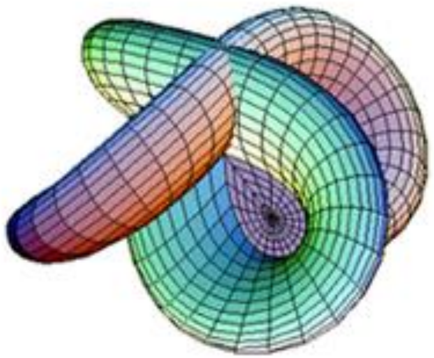
Trees



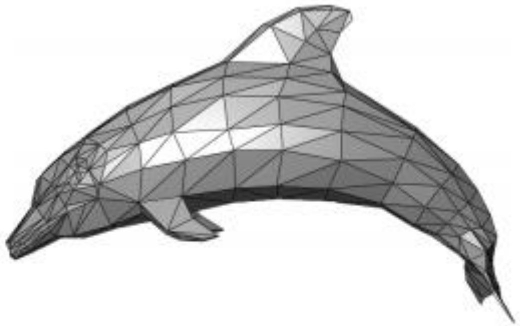
Disease pathways



Networks



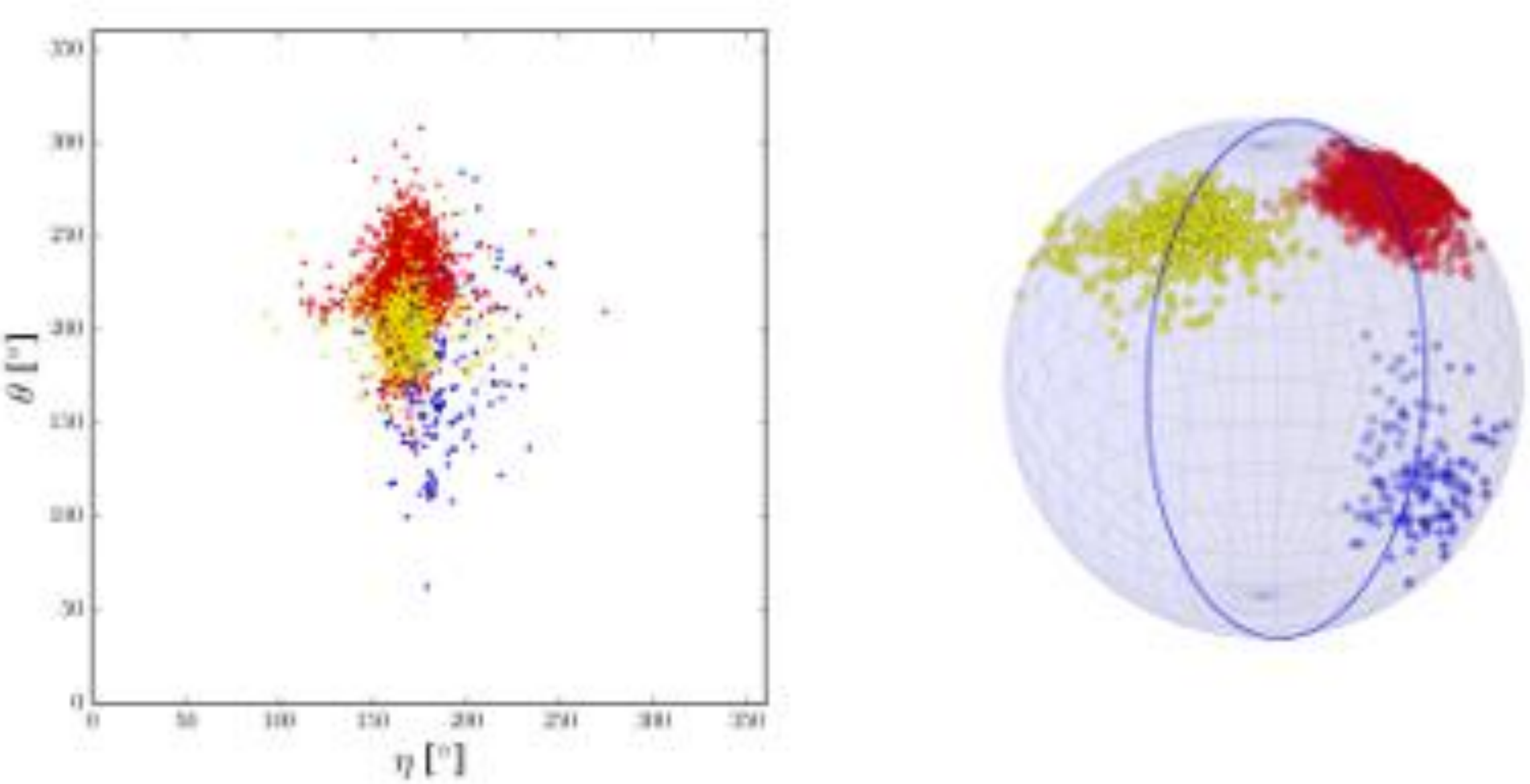
Manifolds



3D Shapes

Image Source: https://hackmd.io/@deep2233/BkeHKgcVd

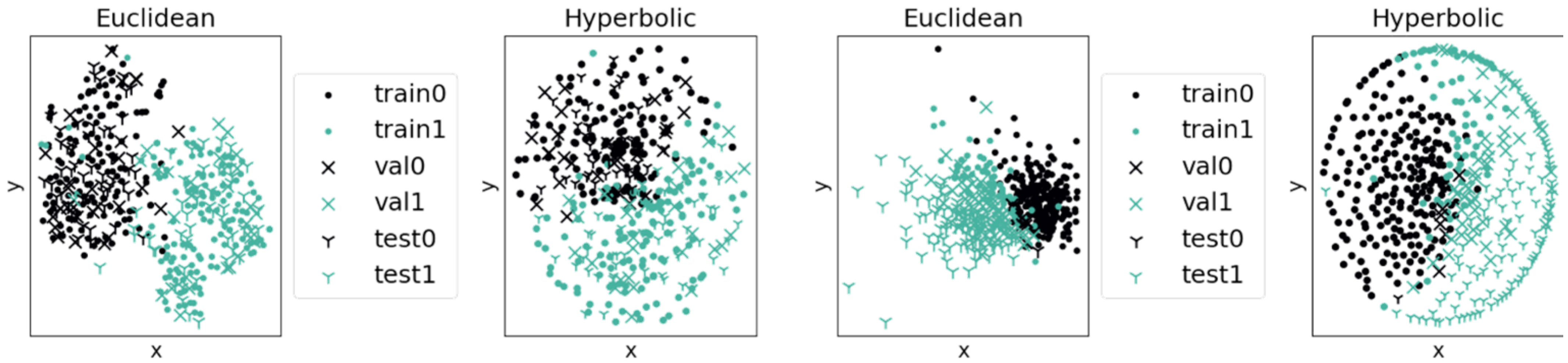
Statistics on non-Euclidian Spaces



Three clusters of different RNA backbone geometries. They overlap in the classic pseudo-torsion representation (left) but can easily be separated using non-Euclidean statistical methods (right).

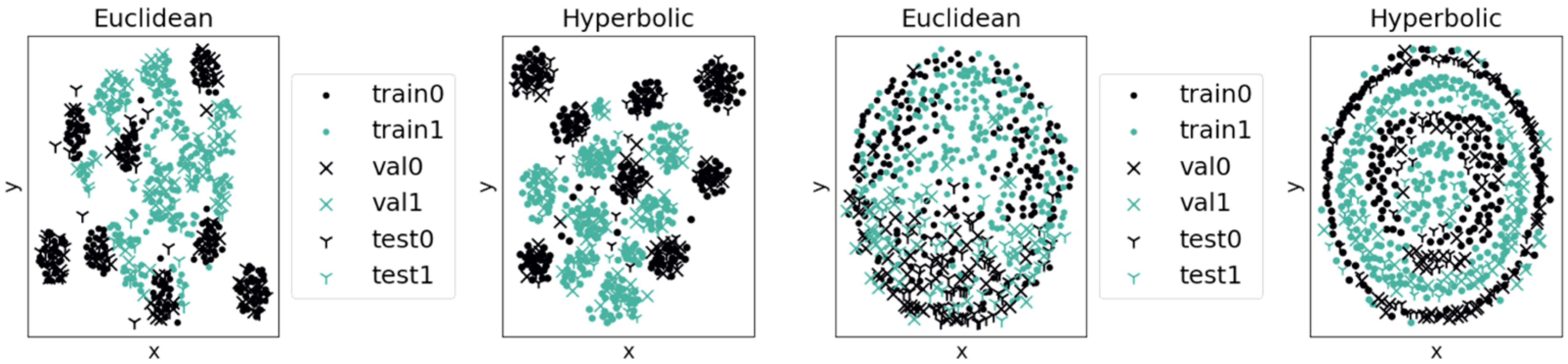
http://www.statistics.uni-goettingen.de/index.php?id=20

Landscape of data points in both Euclidean and Hyperbolic space



(a) t-SNE (data: *breast-cancer-wisc-diag*)

(b) Sammon mapping (data: *breast-cancer-wisc-diag*)

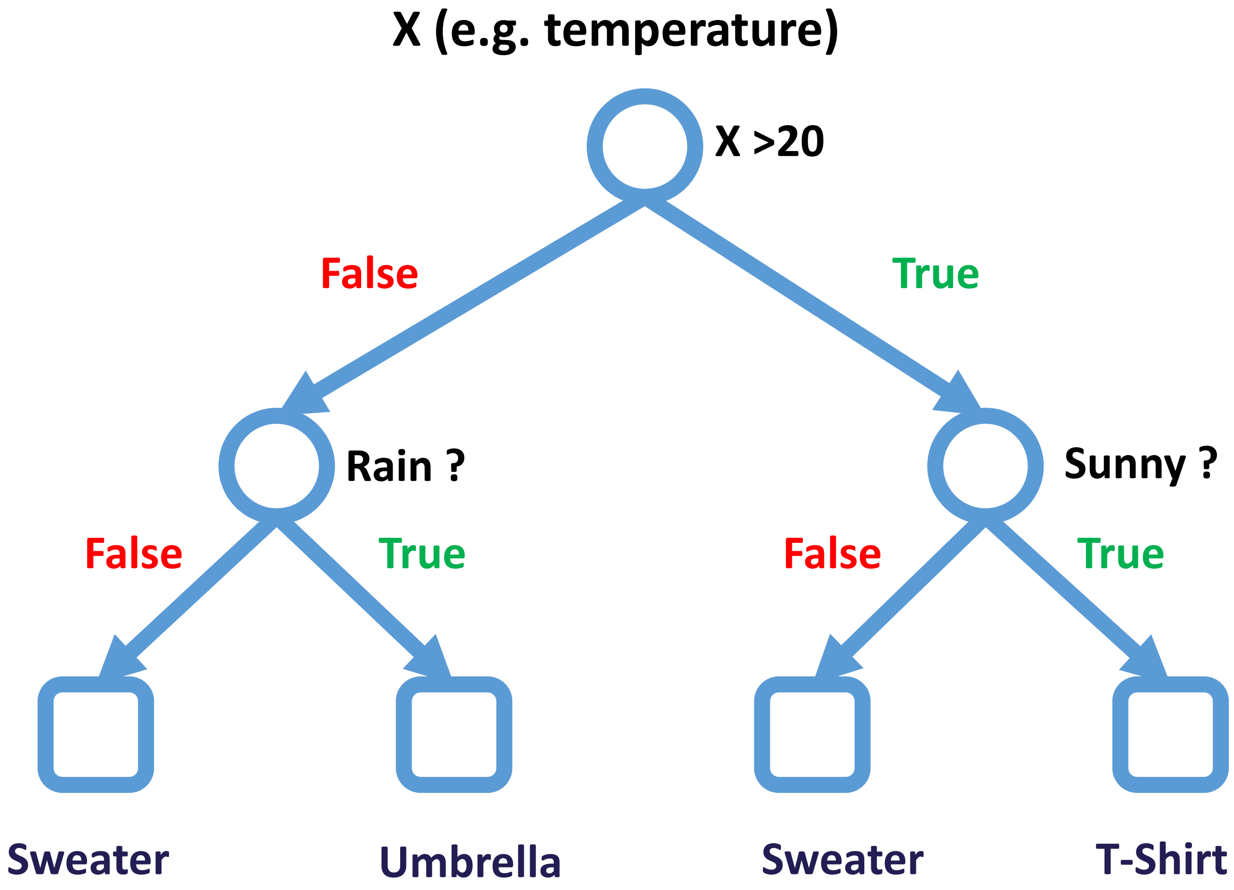


(c) t-SNE (data: *tic-tac-toe*)

(d) Sammon mapping (data: *tic-tac-toe*)

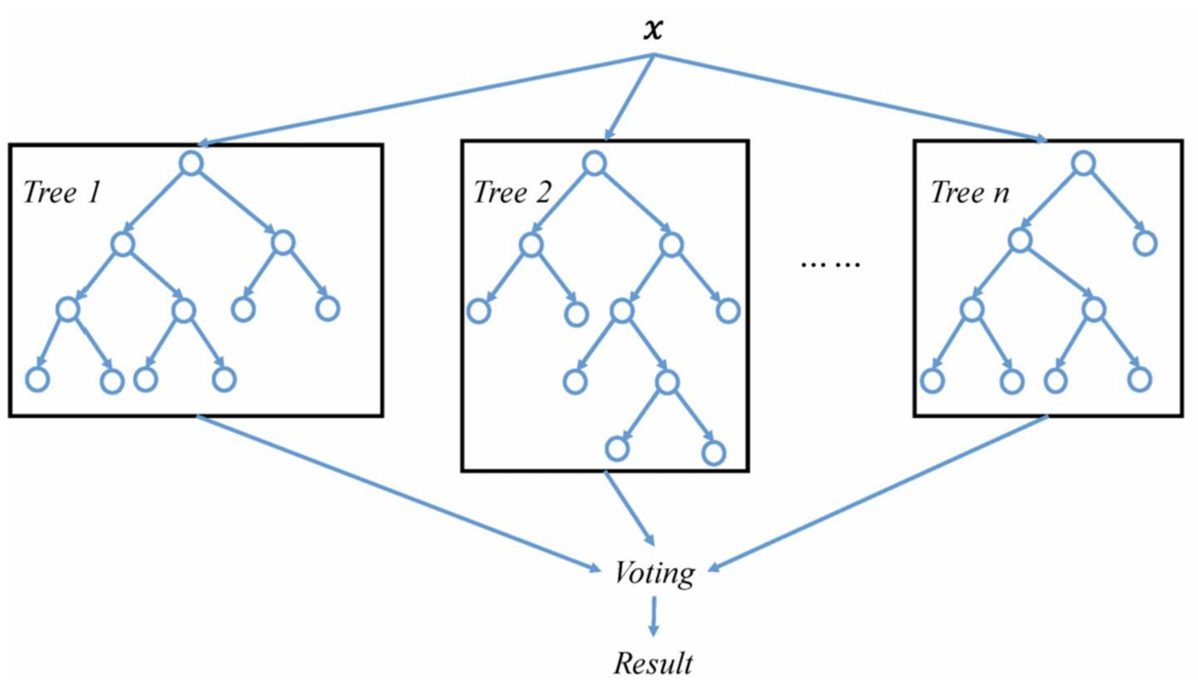
Decision Tree

a model that predicts the value of a target variable by learning simple decision rules inferred from the data features



eXtrema Gradient Boost (XGBoost)

Random Forest is a collection of decision tree where each tree is built from a random subset of the training set using bootstrap sampling. When splitting a node during the construction of the tree, normal bagging method would choose the best split among all features



XGBoost is an iterative decision tree algorithm with multiple decision trees. Every tree is learning from the residuals of all previous trees. *Rather than adopting most voting output results in Random Forest, the predicted output of XGBoost is the sum of all the results*

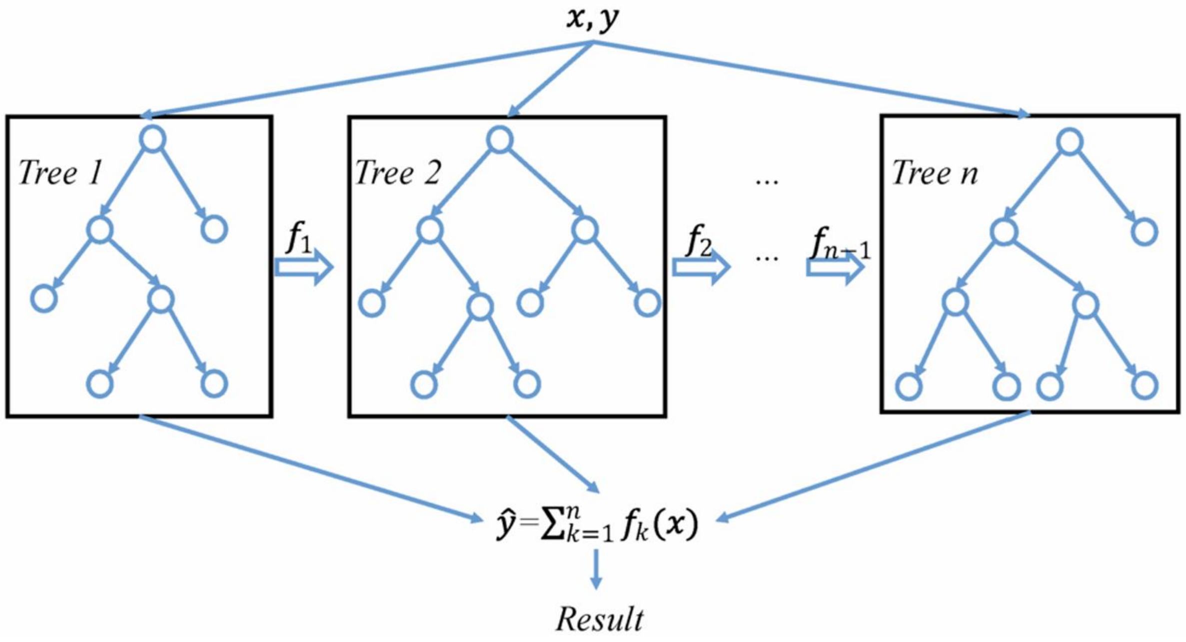
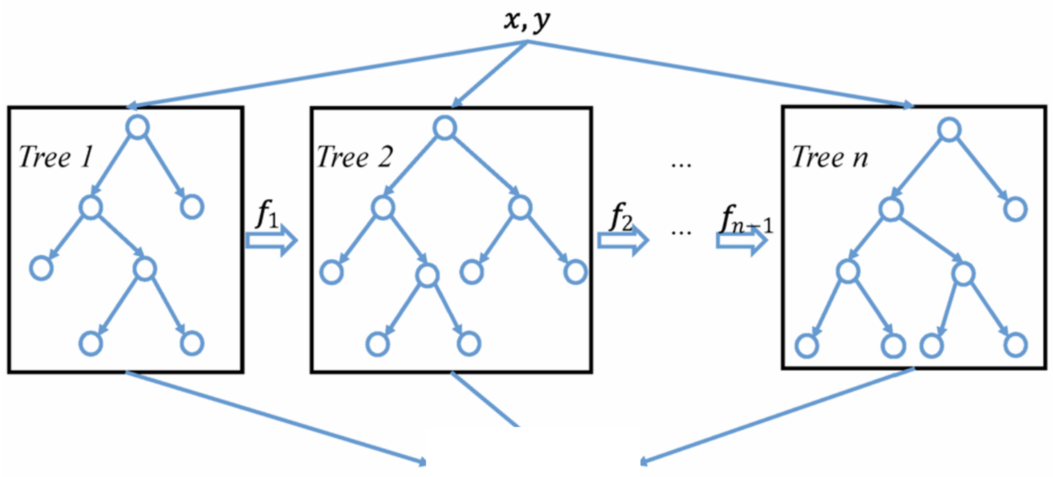


Image source: Wang et al. Astrophys Space Sci 364, 139 (2019). https://doi.org/10.1007/s10509-019-3602-4

XGBoost in nutshell



$$\hat{y}_m = \sum_{m=1}^{n} f_m$$

(1) create an initial model f_1

(2) build a new model f_2 to fit the residuals from the previous model

(3) ensemble models at m-th step with a learning rate η as:

$$f_m = f_{m-1} + \eta * \frac{\partial Loss}{\partial f_{m-1}}$$

We solve a loss at m-th step computed as follows:

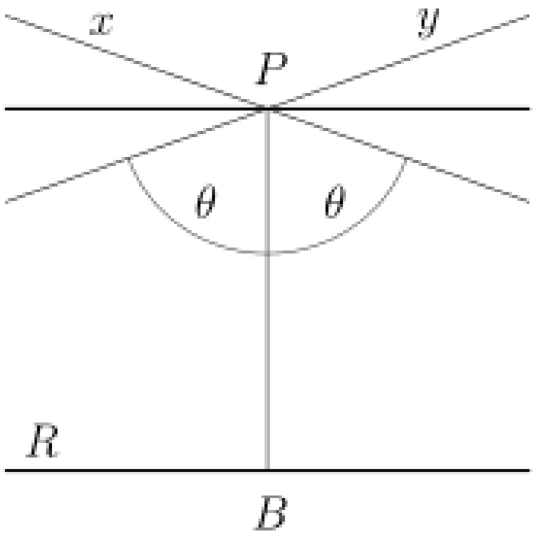
$$L_m = \sum_{i=1}^{N} \left[g_i * f_m + \frac{1}{2} h_i * f_m^2 \right] + \Omega f_m$$

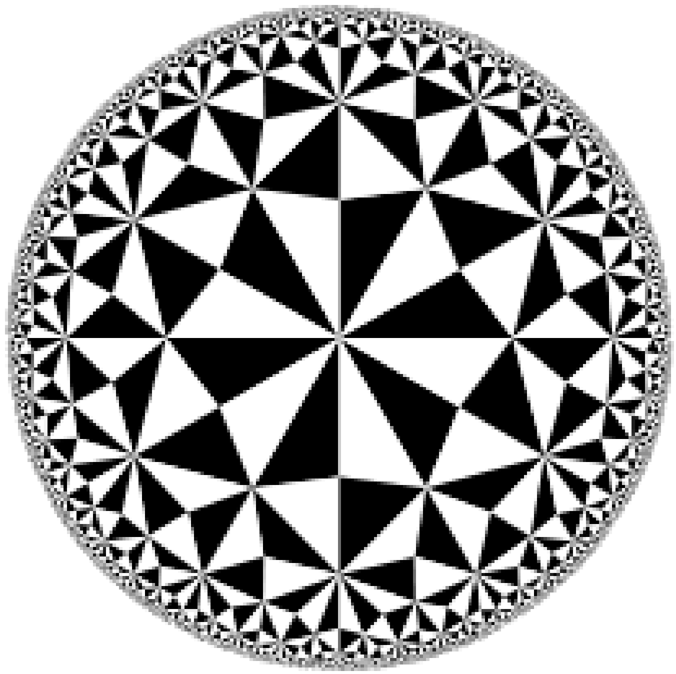
Basically, we compute

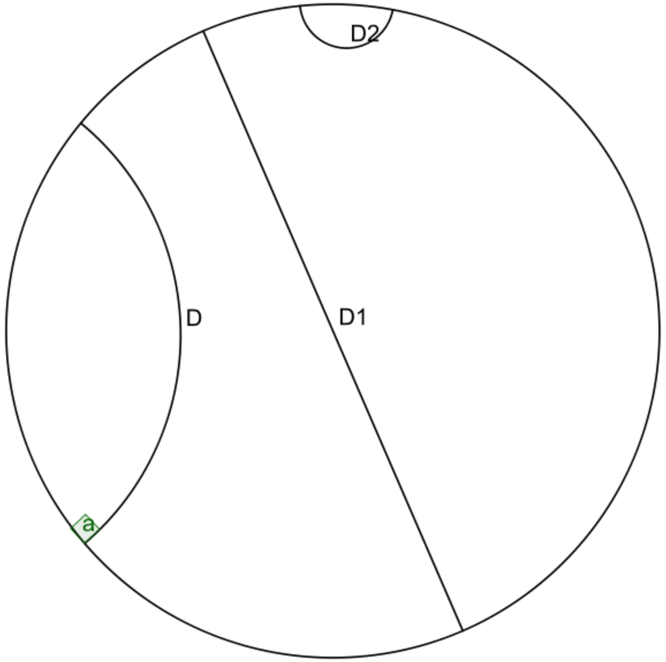
Gradient $g_i = \frac{\partial L(y_i - \hat{y}_{m-1})}{\partial \hat{y}_{m-1}}$ and **Hessian** as $h_i = \frac{\partial^2 L(y_i - \hat{y}_{m-1})}{\partial \hat{y}_{m-1}^2}$

Hyperbolic Geometry: **Poincaré** disk model

The Poincaré hyperbolic disk is a two-dimensional space having hyperbolic geometry defined as the disk





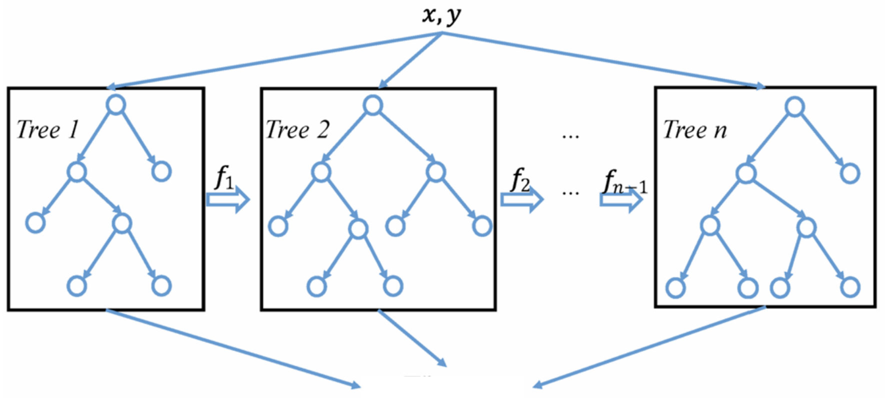


Hyperbolic Geometry: For any given line R and point P not on R, in the plane containing both line R and point P there are at least two distinct lines through P that do not intersect R.

The Poincare ball model is a model of n-dimensional hyperbolic geometry in which all points are embedded in an n-dimensional sphere (or in a circle in the 2D case which is called the Poincaré disk model)

We can represent common geometric concepts by points on the unit circle. Starting with a line, if we project the geodesic line from the hyperboloid to the unit circle, we get an arcs along the unit circle with each one approaching the circumference at a 90 degree angle.

Poincaré XGBoost (PXGBoost) in nutshell



$$\hat{y}_m = \sum_{m=1}^{n} f_m$$

We solve a loss at m-th step computed as follows:

$$L_m = \sum_{i=1}^{N} \left[g_i * f_m + \frac{1}{2} h_i * f_m^2 \right] + \Omega f_m$$

Basically, we compute

Gradient $g_i = \frac{\partial L(y_i - \hat{y}_{m-1})}{\partial \hat{y}_{m-1}}$ and **Hessian** as $h_i = \frac{\partial^2 L(y_i - \hat{y}_{m-1})}{\partial \hat{y}_{m-1}^2}$

Replace Euclidean Gradient and Hessian with Hyperbolic ones

Riemannian gradient $g(x) = \nabla_R f(x) = \frac{1}{1 - ||x||} < \nabla f(x), x > \cdot x - \nabla f(x),$

Riemannian Hessian $h(x) = \frac{1}{1 - ||x||} (g(x) \nabla^2 f(x) + \nabla f_i(x) \nabla f_j(x)),$

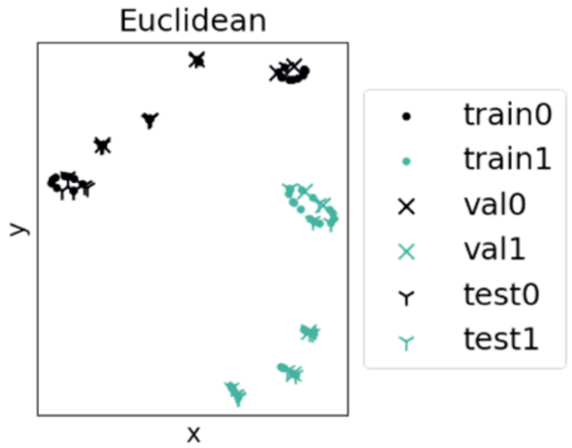
Results (F1 Score on classification)

Dataset	Name (ID)	classes (ℓ)	Space	Features (m)	Instance (n)
UCI	64 datasets	[2 – 15]	Euclidean	[3 – 60]	[24 – 4177]
H-UCI	64 datasets	[2 – 15]	Hyperbolic	[3 – 60]	[24 – 4177]

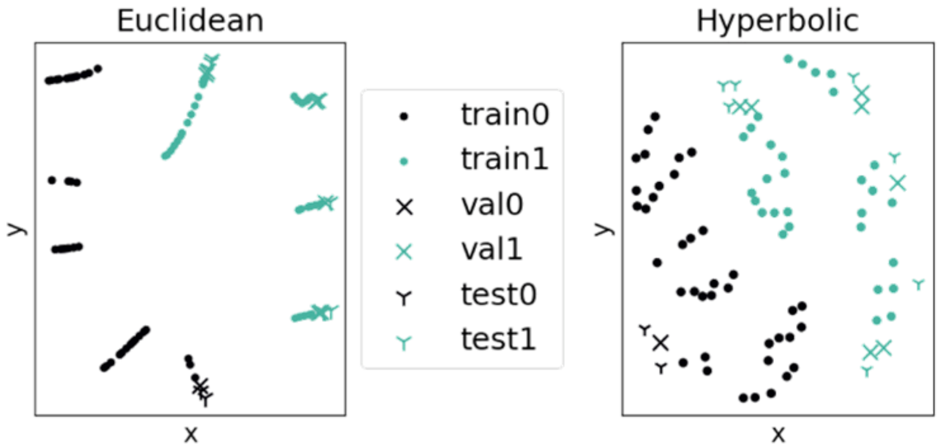
Dataset	Accuracy				F1-macro			
	HLSVM	HoroRF	Xgboostt⋆	PXgboost	HLSVM	HoroRF	Xgboostt⋆	PXgboost
H-abalone	0.6312	0.5172	0.6410	0.6377	0.6170	0.5152	0.6316	0.6307
H-acute-inflammation	1.0000	0.9500	0.9583	0.9750	1.0000	0.9492	0.9582	0.9749
:	:	:	:	:	:	:	:	:
H-vertebral-column-3clases	0.8052	0.5584	0.8214	0.8182	0.7253	0.4089	0.7510	0.7551
H-wine	0.9830	0.5284	0.9489	0.9489	0.9827	0.5027	0.9496	0.9471
H-wine-quality-red	0.5869	0.4669	0.6463	0.6619	0.2270	0.1814	0.3348	0.3445
H-zoo	0.9400	0.6800	0.8400	0.8200	0.8140	0.3844	0.5211	0.5572
Win-Tie-Lose of PXgboost	41-1-22	59-0-5	34-7-23		35-0-29	58-0-6	38-4-22	

Results of Xgboost and PXgboost

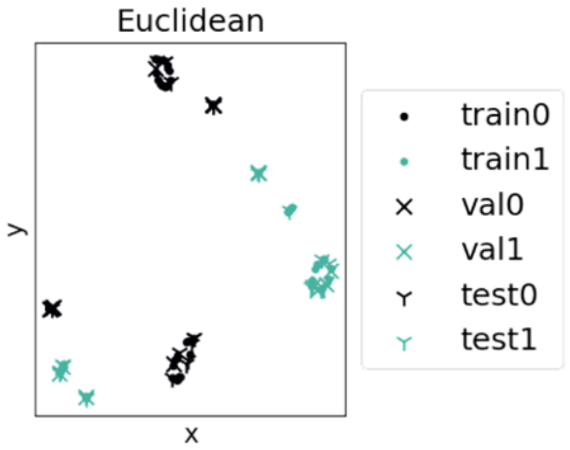
0.9750 and 0.9950 by Pxgboost and 0.9583 and 0.9833 by Xgboost, where the data in classes have a clear boundary



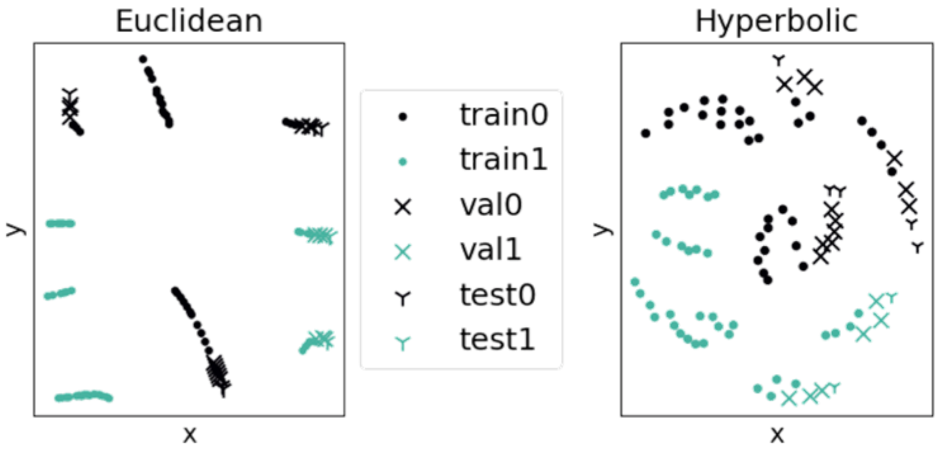
(a) t-SNE (data: *acute-nephritis*)



(b) Sammon mapping (data: *acute-nephritis*)



(c) t-SNE (data: *acute-inflammation*)



(d) Sammon mapping (data: *acute-inflammation*)

National Edge AI Hub

Get in touch

Address

Urban Sciences Building, 1 Science Square,
Newcastle upon Tyne NE4 5TG, UK

Email

Varun.Ojha@newcastle.ac.uk

Web

https://ojhavk.github.io/

Paper

Euclidean and Poincaré Space Ensemble Xgboost
Information Fusion, Elsevier. (2024)
Suganthan PN, Kong L, Snasel V, Ojha V