Paper ID: 1415

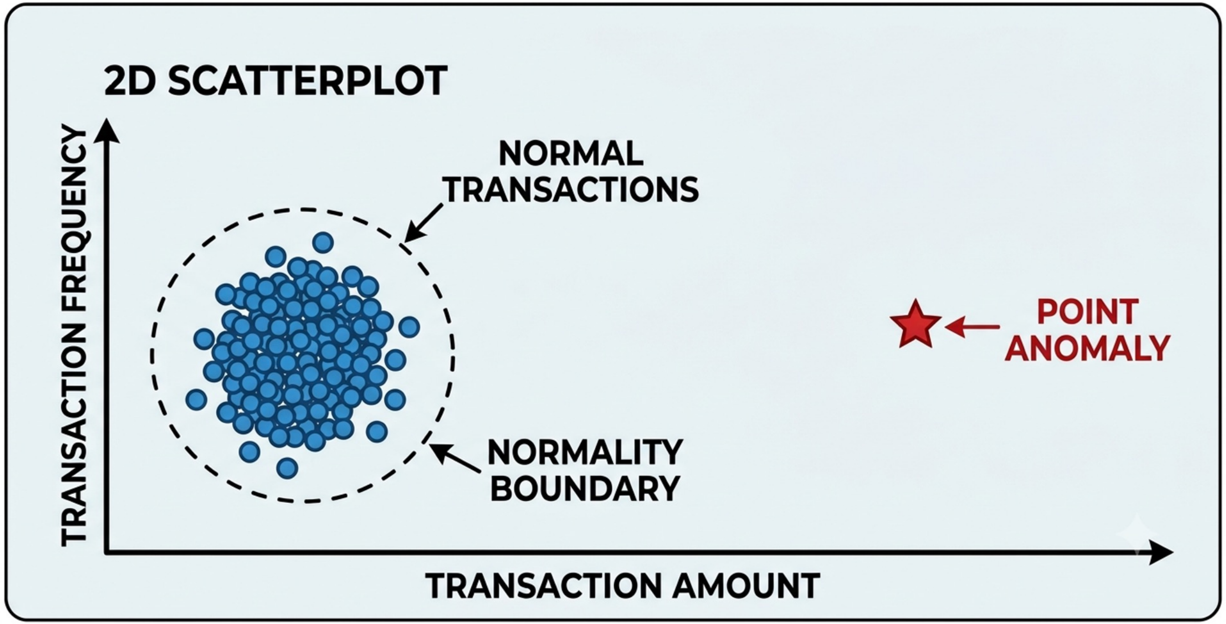
Paper Title: Anomaly Detection Using Density Adaptive Tree Based Clustering

Authors: Subhadip Boral, Sagnik Ghosal, Ashish Ghosh

Presenter: Subhadip Boral, Computer Science & Information Technology, ITER, Siksha O Anusandhan, subhadipboral@soa.ac.in

What is an Anomaly?

- ❖ Observation or a group of observations.
- ❖ Functions Distinctly Different from Other Observations.
- ❖ Generated by a Completely Different Mechanism.



What are the Challenges?

- Imbalance in the information available for normal data and anomalous data.
- Compounds when anomalies occur with diversity.

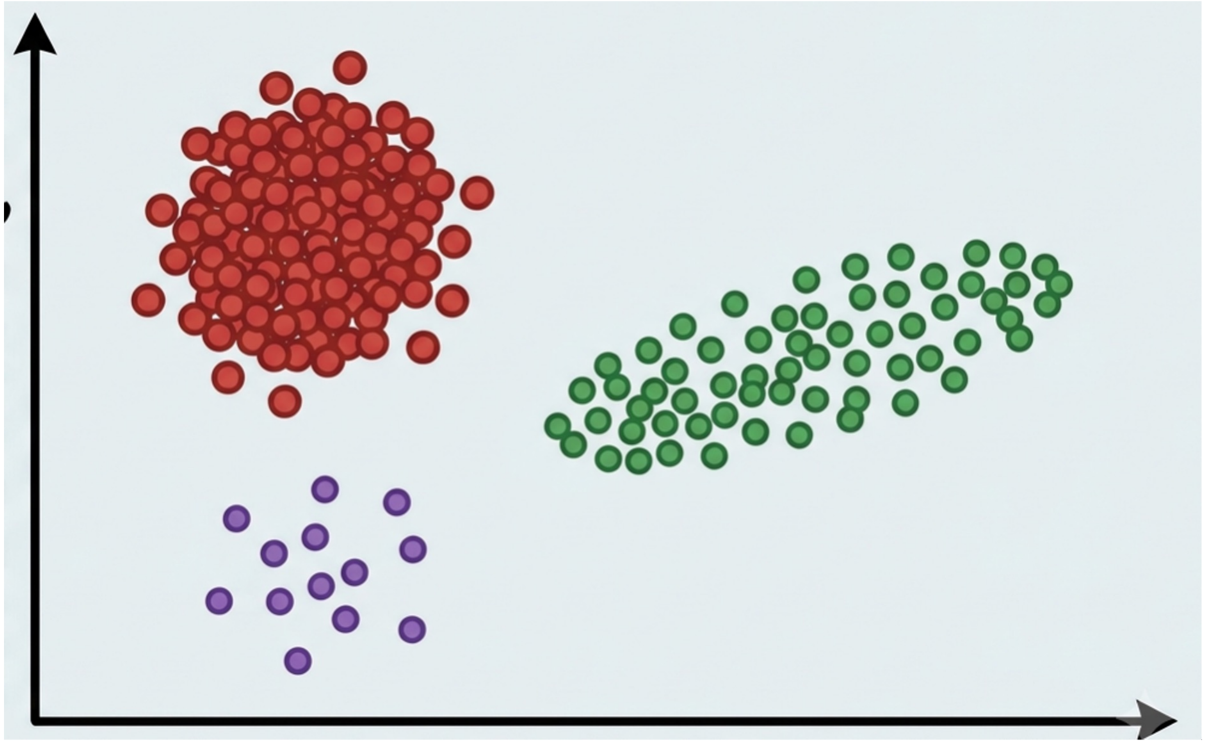
What could be the approach?

Unsupervised Learning

What it will learn?

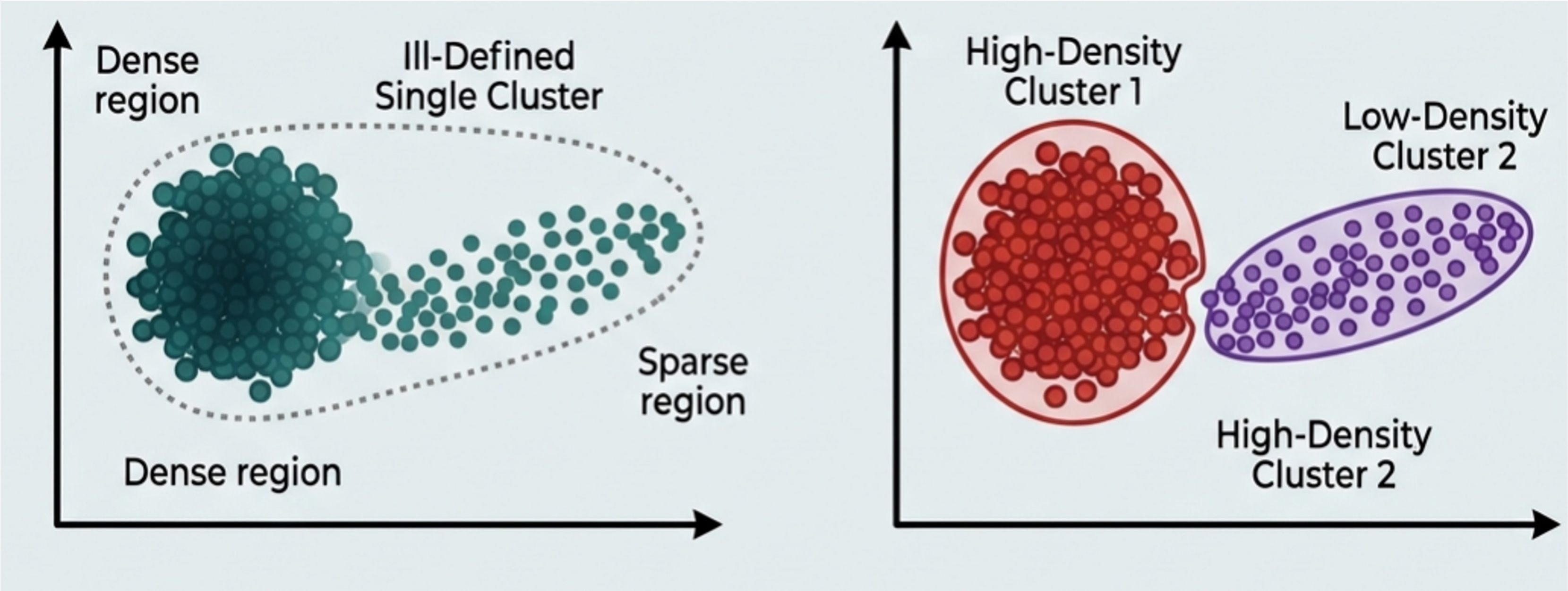
Learn the structure of the normal Data and represent them with clusters

Intention



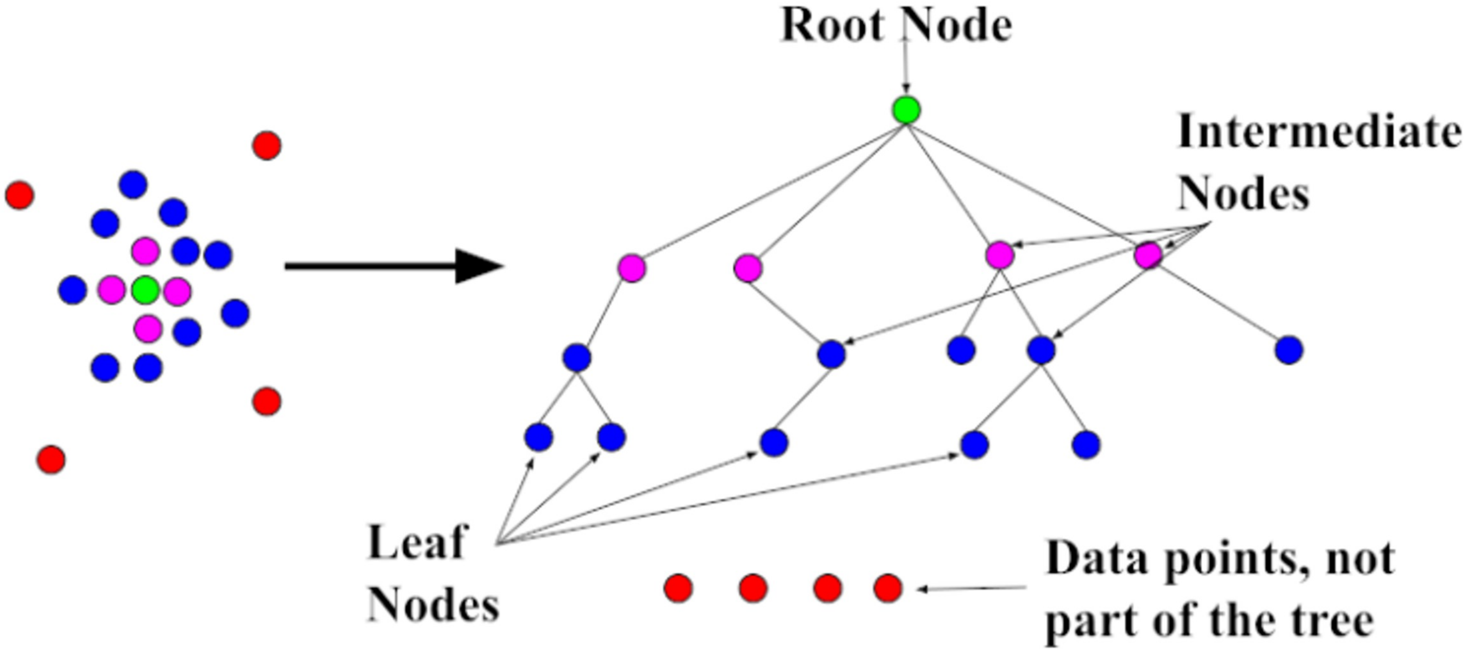
Dataset with diverse density clusters

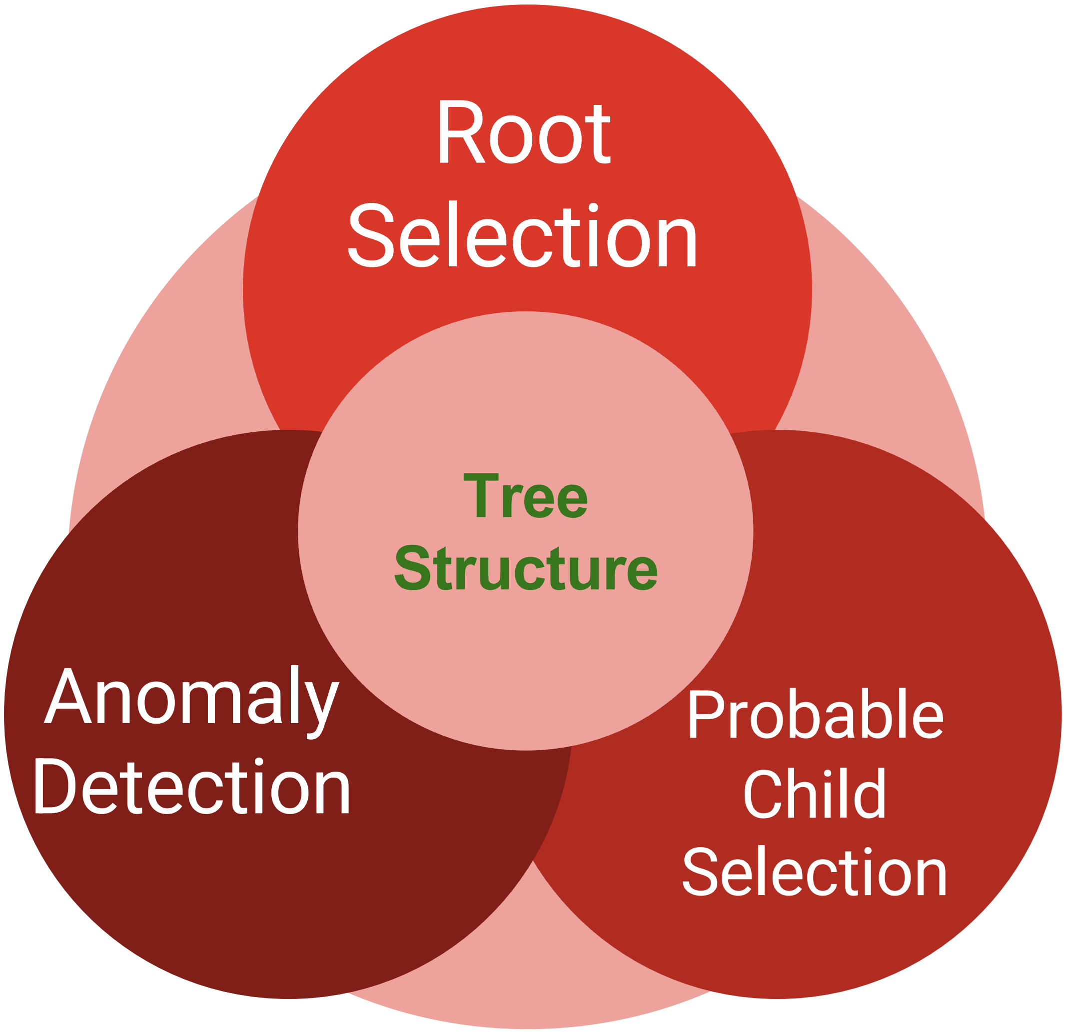
Intention



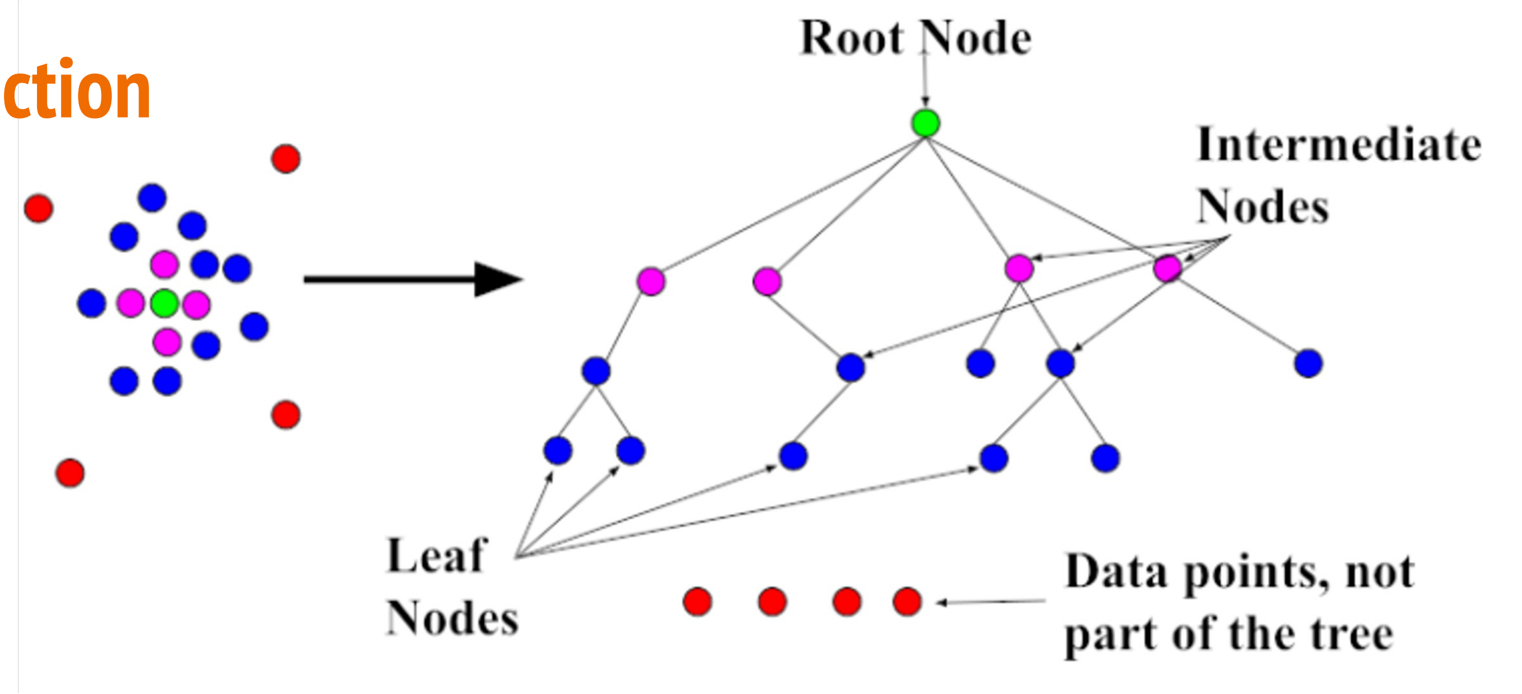
Note: Significant density change signifies different clusters.

Proposed Cluster Structure



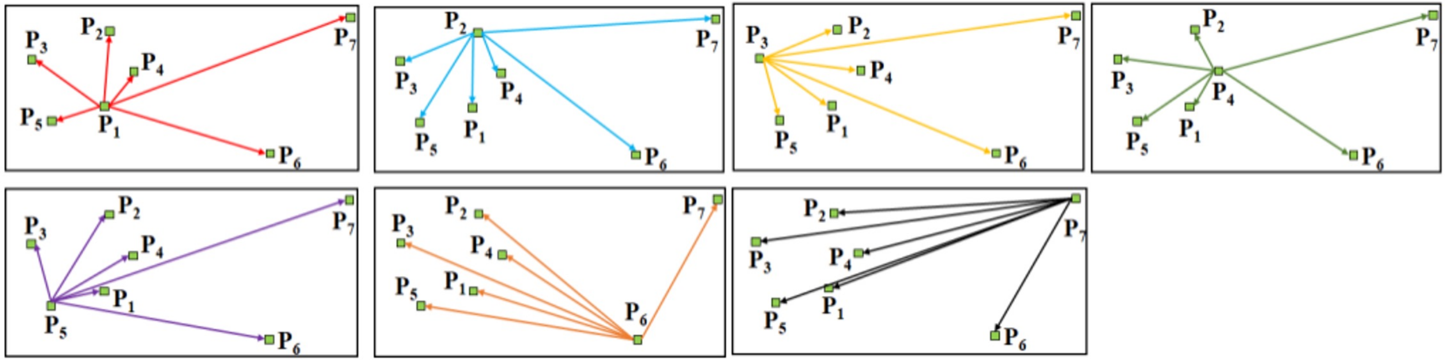

Root Selection
Tree Structure
Anomaly Detection
Probable Child Selection

Root Selection



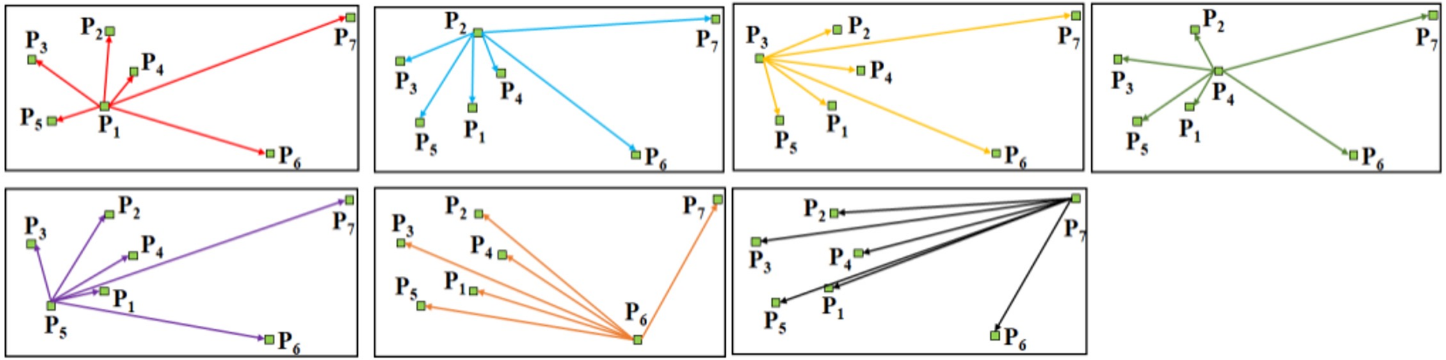
$$Den = \frac{\Sigma_{i=1}^{k_{Opt}} E(\boldsymbol{p}, \boldsymbol{p_i})}{\pi E(\boldsymbol{p}, \boldsymbol{p}_{k_{Opt}})^2}$$

Neighbourhood Selection



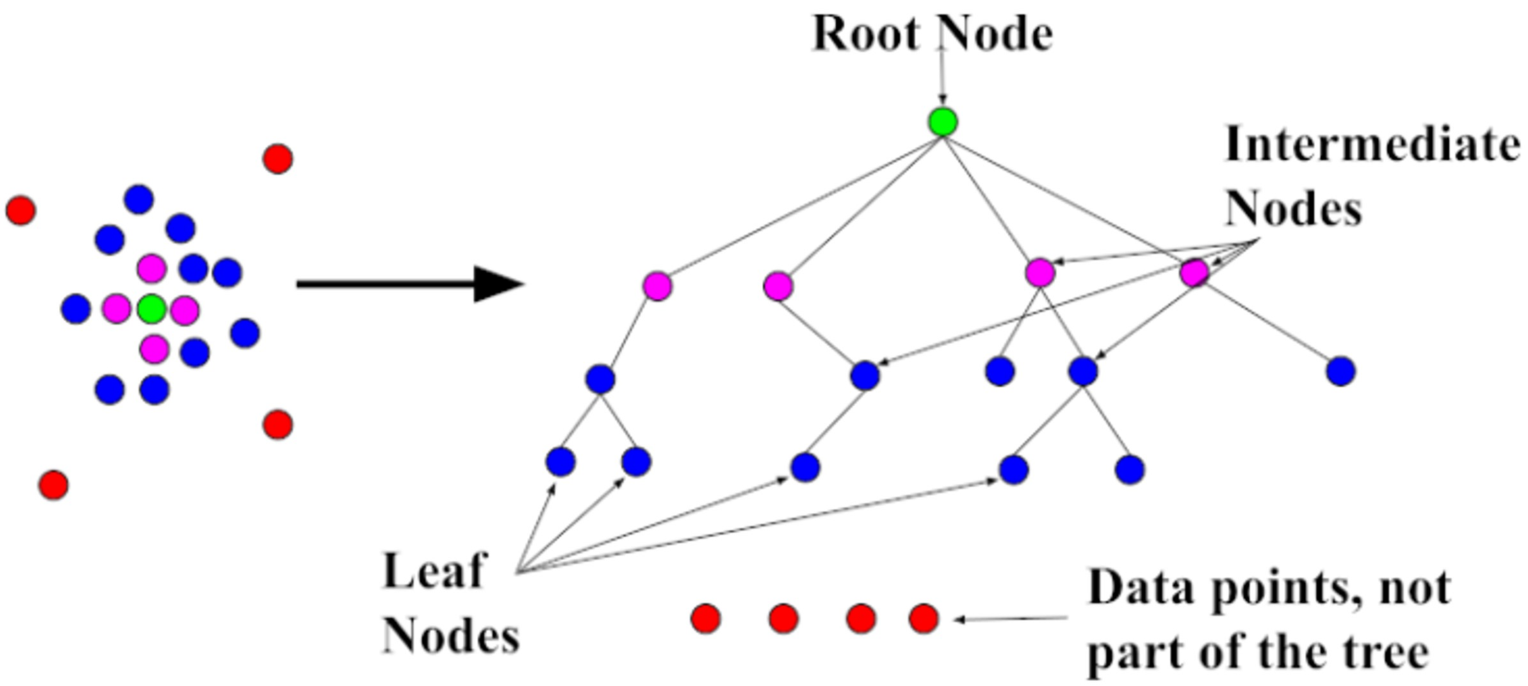
$$Er = |\boldsymbol{p} - \Sigma_{i=1}^{k} \Omega_i \boldsymbol{p_i}|^2$$

Neighbourhood Selection



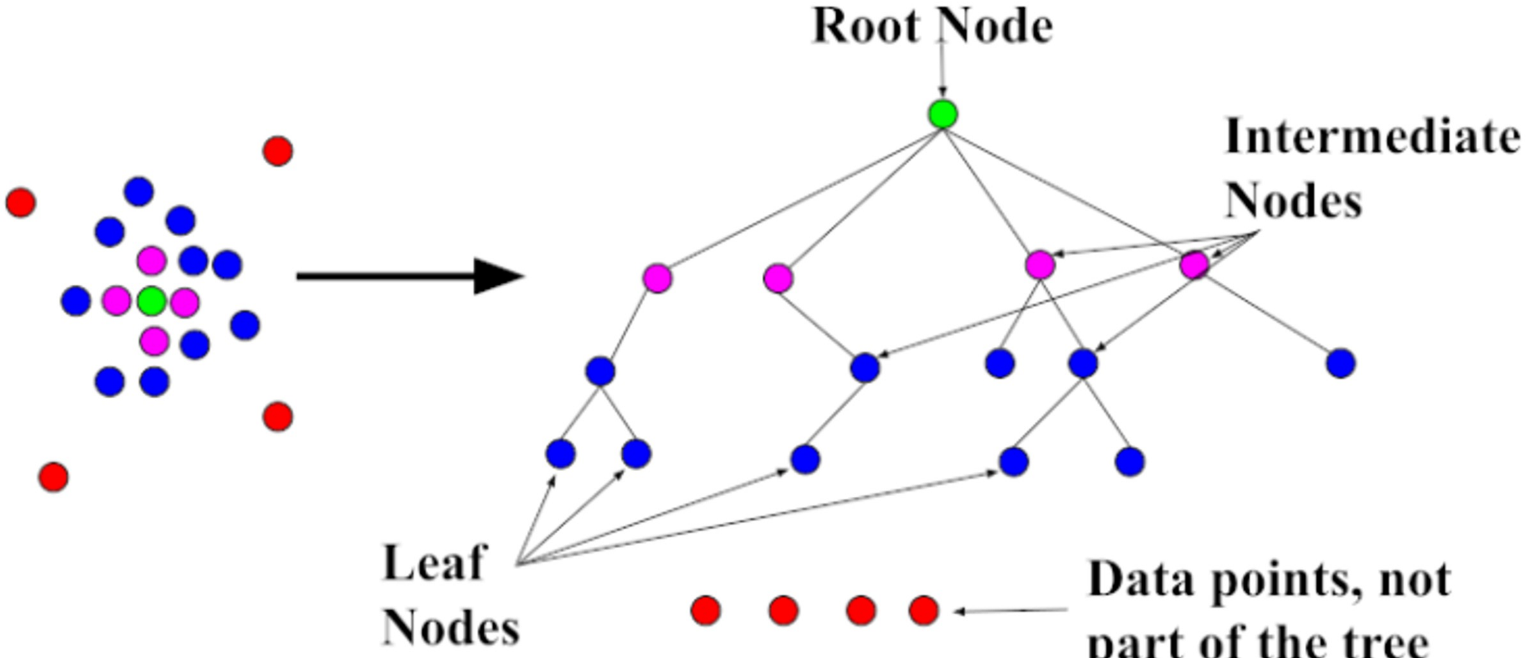
$$
W = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \\ P_7 \end{array}
\begin{array}{ccccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 & P_7 \end{array}
\begin{pmatrix}
0 & 0.22204 & 0.21179 & 0.24391 & 0.23223 & \mathbf{0.06349} & \mathbf{0.02639} \\
0.23568 & 0 & 0.23659 & 0.25683 & 0.21450 & \mathbf{0.03025} & \mathbf{0.02540} \\
0.22982 & 0.23052 & 0 & 0.22026 & 0.23257 & \mathbf{0.05632} & \mathbf{0.02916} \\
0.23698 & 0.23321 & 0.22503 & 0 & 0.22983 & \mathbf{0.04987} & \mathbf{0.02476} \\
0.24987 & 0.20526 & 0.23145 & 0.22025 & 0 & \mathbf{0.07306} & \mathbf{0.01997} \\
0.18781 & 0.14220 & 0.13621 & 0.19791 & 0.16315 & 0 & 0.17174 \\
0.15563 & 0.17267 & 0.15358 & 0.18136 & 0.15126 & 0.18364 & 0
\end{pmatrix}
$$

Probable to Prominent Child Selection



Prominent Child Selection

$$Tree\_Den(\boldsymbol{c}) = \frac{\Sigma E(\boldsymbol{c}, \boldsymbol{p_i})}{\pi E(\boldsymbol{q}, \boldsymbol{q}_{far})^2}, \forall \boldsymbol{p_i} | E(\boldsymbol{c}, \boldsymbol{p_i}) \leq E(\boldsymbol{q}, \boldsymbol{q}_{far})$$



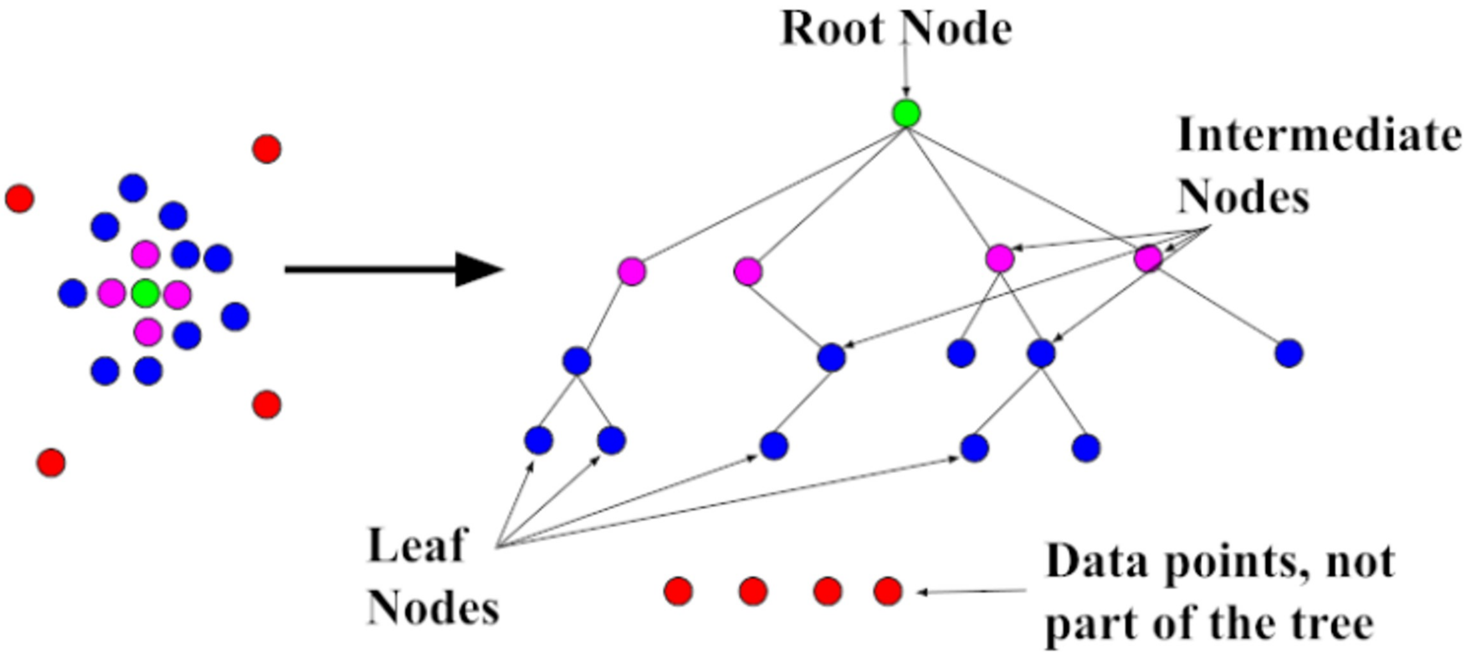
Prominent Child Selection

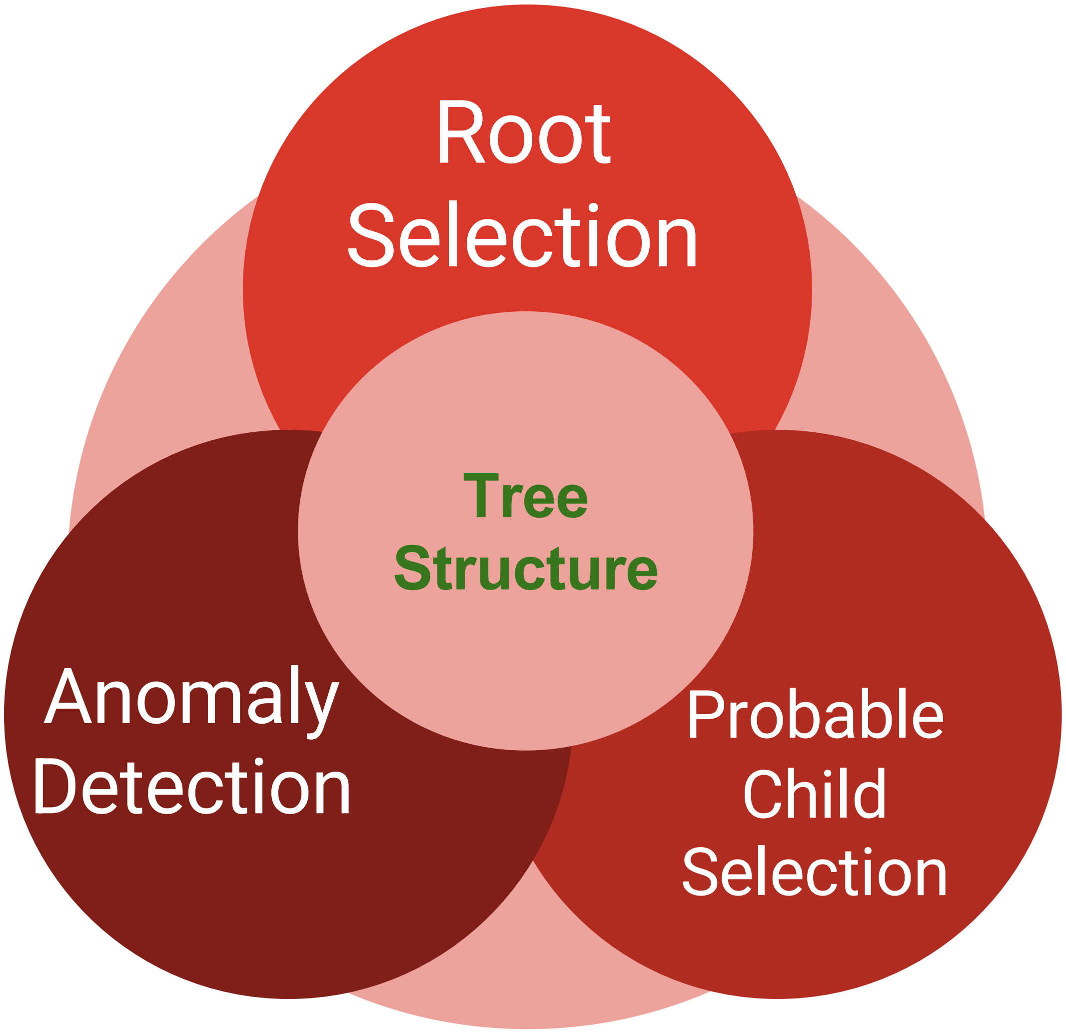
$$Tree\_Den(\boldsymbol{c}) = \frac{\Sigma E(\boldsymbol{c}, \boldsymbol{p_i})}{\pi E(\boldsymbol{q}, \boldsymbol{q}_{far})^2}, \forall \boldsymbol{p_i} | E(\boldsymbol{c}, \boldsymbol{p_i}) \leq E(\boldsymbol{q}, \boldsymbol{q}_{far})$$

$$\begin{aligned} EwMA_t &= \beta EwMA_{t-1} + (1-\beta)P_t, \beta \in [0,1] \\ &= \beta(\beta EwMA_{t-2} + (1-\beta)P_{t-1}) + (1-\beta)P_t \end{aligned}$$

$$\frac{EwMA_q - Tree\_Den(\boldsymbol{c})}{EwMA_q} \leq \delta$$

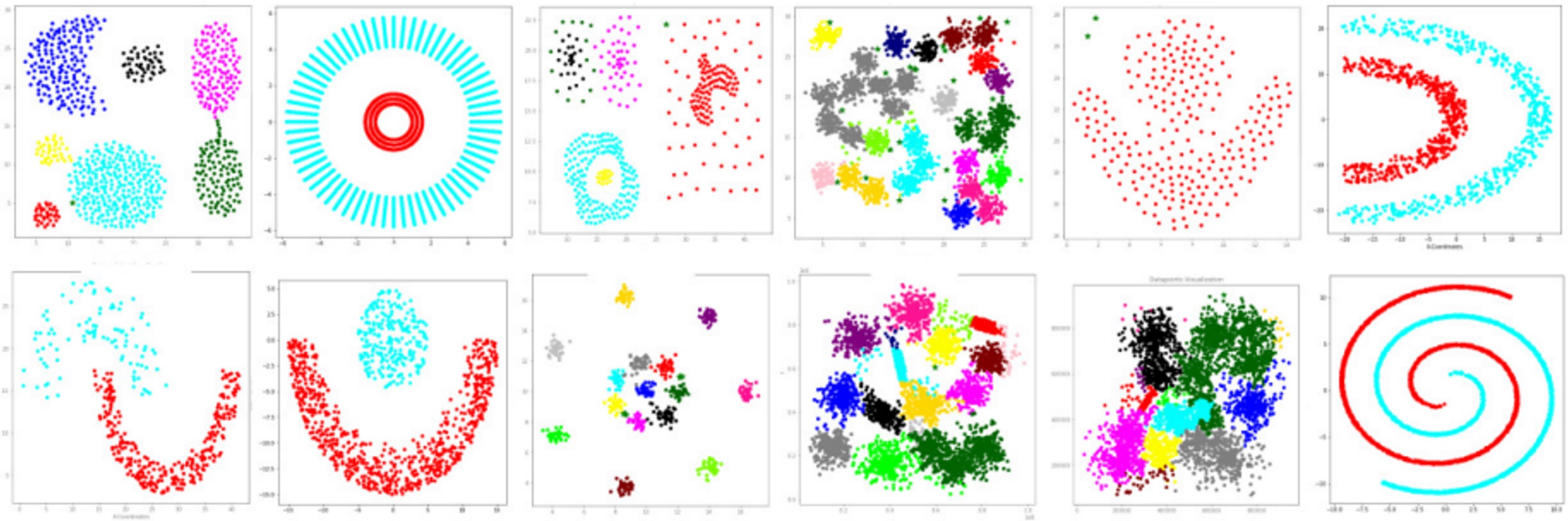
Proposed Cluster Structure




Root
Selection
Tree
Structure
Anomaly
Detection
Probable
Child
Selection

Performance Analysis

DATASET	$F - score$	$g - Mean$	$Accuracy$	DB	DUNN	SIL	CAL-HAR
A3	0.834	0.848	82.973	1.412	0.002	0.423	7236.720
AGGREGATION	0.986	0.986	98.984	0.526	0.041	0.413	1044.238
CLUSTER IN CLUSTER	1	1	100	492.699	0.227	0.234	0.015
COMPOUND	0.809	0.834	72.932	4.866	0.050	0.158	419.403
CORNERS	1	1	100	0.766	0.292	0.453	987.081
D31	0.895	0.902	87.968	1.358	0.002	0.346	2047.410
DOUBLE MOON1	1	1	100	0.843	0.177	0.483	12153.558
DOUBLE MOON2	1	1	100	0.754	0.413	0.543	15614.521
FLAME	1	1	100	0.580	0.193	0.295	6.894
HALF KERNEL	1	1	100	2.721	0.114	0.201	114.067
IMAGE1	1	1	100	0.295	0.266	0.791	90712.755
IMAGE2	0.999	0.999	99.999	0.332	0.025	0.763	63701.060
IMAGE3	0.999	0.999	99.950	0.354	0.009	0.750	59140.142
JAIN	1	1	100	0.899	0.092	0.402	279.483
MOON	1	1	100	1.628	0.176	0.206	141.495
OUTLIER	1	1	100	0.406	0.434	0.563	886.247
R15	1	1	86.833	1.326	0.025	0.736	3552.621
S1	0.963	0.965	94.420	1.657	0.003	0.627	13733.068
S2	0.870	0.879	86.56	1.709	0.001	0.346	4262.244
S3	0.685	0.712	64.1	0.868	0.003	0.158	1961.286
TWO SPIRALS	1	1	100	3.426	0.139	0.061	393.227
UNBALANCE	0.913	0.914	98.154	1.176	0.0004	0.707	63995.676



Clustering Efficiency Visualization

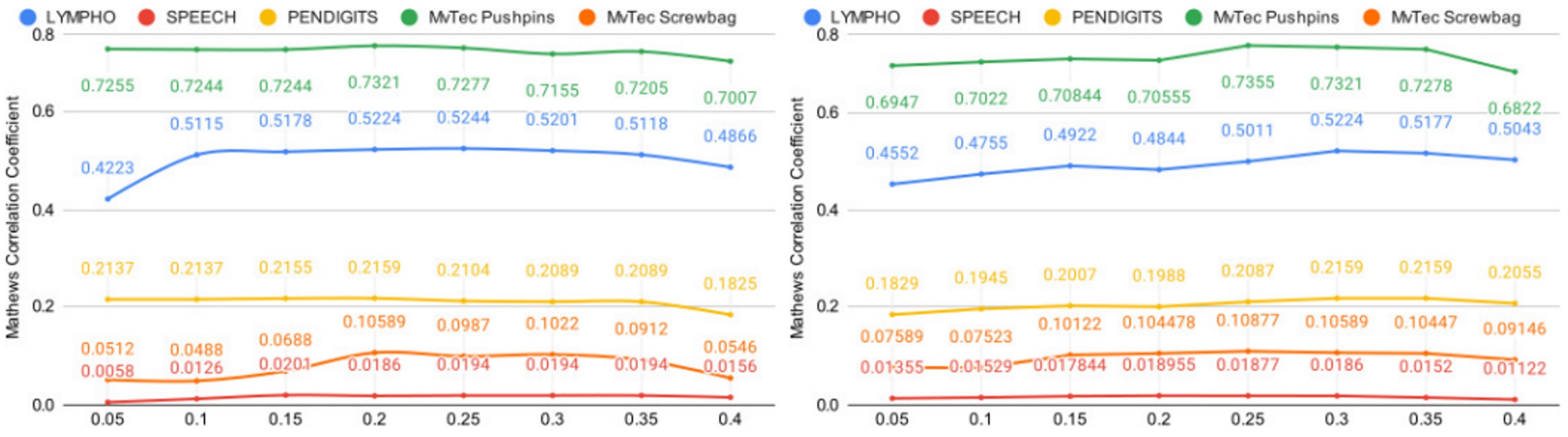
Table 4: *Accuracy* of the Proposed method and other state-of-the-art algorithms over real-world datasets

Dataset Number	LOF[5]	iForest[12]	Hbos[9]	LSCP[19]	SUOD[18]	ECOD[11]	WFRDA[17]	DIF[16]	Proposed
1	0.93	**0.94**	**0.94**	**0.94**	**0.94**	**0.94**	0.54	0.92	**0.97**
2	**0.74**	**0.74**	0.58	**0.74**	**0.74**	0.72	0.7	0.73	**0.78**
3	**0.87**	0.85	0.85	0.85	0.85	**0.87**	0.61	0.84	**0.86**
4	**0.91**	0.88	0.88	0.89	0.89	0.88	0.5	0.81	**0.97**
5	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**	0.75	0.81	**0.84**
6	0.88	0.88	0.9	0.88	0.88	0.88	0.72	**0.92**	**0.97**
7	0.88	0.87	0.84	0.86	0.86	0.84	0.65	**0.93**	**0.96**
8	0.88	0.91	0.91	0.9	0.91	0.91	0.85	**0.94**	**0.9**
9	0.89	0.88	0.88	0.88	0.88	0.88	0.8	**0.94**	**0.97**
10	0.9	0.9	0.9	0.9	0.9	0.9	0.84	**0.93**	**0.98**
11	0.9	0.9	**0.93**	0.9	0.9	0.9	0.78	**0.93**	**0.98**
12	0.86	0.88	**0.89**	0.87	**0.89**	0.87	0.82	0.86	**0.93**
13	0.75	0.73	**0.76**	**0.76**	0.73	0.75	0.63	0.75	**0.82**
14	0.77	0.78	0.79	**0.81**	0.79	**0.81**	0.68	0.73	**0.83**
Friedman test statistic p-value	1.87E-03								
Wilcoxon signed-rank test p-value	1.01E-03	3.61E-03	6.29E-03	3.48E-03	5.65E-03	8.17E-03	1.11E-03	2.32E-03	-

Table 6: *MCC* of the Proposed method and other state-of-the-art algorithms over real-world datasets in the normal environment

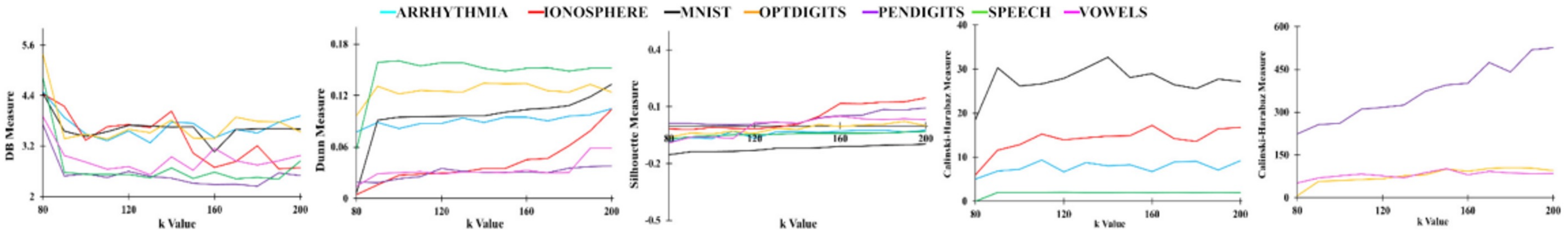
Dataset Number	LOF[5]	iForest[12]	Hbos[9]	LSCP[19]	SUOD[18]	ECOD[11]	WFRDA[17]	DIF[16]	Proposed
1	0.49	**0.61**	**0.61**	**0.61**	**0.61**	**0.61**	0.1889	0.4327	**0.6271**
2	0.4447	0.4249	-0.1103	0.4249	0.4447	0.3852	0.4167	**0.483**	**0.5276**
3	**0.4203**	0.3374	0.3374	0.3167	0.3374	0.3996	0.3229	0.3772	**0.4292**
4	**0.389**	0.0501	0.0877	0.1253	0.1505	0.0877	0.0077	0.2836	**0.4507**
5	0.0134	0.0134	-0.0007	0.0063	0.0063	0.0063	-0.0002	**0.1256**	**0.0344**
6	-0.015	0.0611	**0.25**	0.019	0.049	-0.015	0.0171	0.1933	**0.2066**
7	0.024	0.2879	**0.3237**	0.252	**0.2976**	0.2846	0.1231	0.076	0.278
8	0.3106	0.2257	0.092	0.1742	0.2151	0.1029	0.0732	**0.3148**	**0.3733**
9	**0.1058**	0.0154	0.0126	0.06	0.0443	-0.0003	0.0038	0.032	**0.0848**
10	0.0512	0.1239	0.0576	0.1184	0.1261	**0.1904**	0.0608	0.072	**0.1447**
11	-0.0096	0.1876	0.2175	0.1877	0.1876	0.1876	0.0016	**0.3920**	**0.9739**
12	0.5266	0.4788	0.504	0.5146	0.504	0.4899	0.3892	**0.5572**	**0.7644**
13	0.01341	0.0128	-0.0382	0.0022	0.0116	0.0088	0.1314	**0.2031**	**0.2614**
14	0.0113	0.0112	-0.0358	0.01255	0.0149	0.0157	-0.0129	**0.0932**	**0.1257**
Friedman test statistic p-value	2.83E-04								
Wilcoxon signed-rank test p-value	4.27E-04	2.14E-04	1.31E-03	3.05E-05	4.27E-04	5.16E-03	1.66E-05	3.93E-05	-

Sensitivity Study



(a) Over an ϵ value range

(b) Over a δ value range


ARRHYTHMIA
IONOSPHERE
MNIST
OPTDIGITS
PENDIGITS
SPEECH
VOWELS
DB Measure
5.6
4.4
3.2
2
80
120
160
200
k Value
Dunn Measure
0.18
0.12
0.06
0
80
120
160
200
k Value
Silhouette Measure
0.4
0.1
-0.2
-0.5
80
120
160
200
k Value
Calinski-Harabaz Measure
40
30
20
10
0
80
120
160
200
k Value
Calinski-Harabaz Measure
600
450
300
150
0
80
120
160
200
k Value

Contributions

- ★ A novel neighborhood selection technique.
- ★ A novel density-adaptive tree-based clustering method.
- ★ Effective anomaly detection across lower and higher dimensions.

Limitations

- ★ Effective cluster validity index where clusters share the same space.
- ★ Explore the relation between exponential moving average and high-dimensional clusters.
- ★ Comparative effectiveness in higher dimensions should be improved.

Thank you