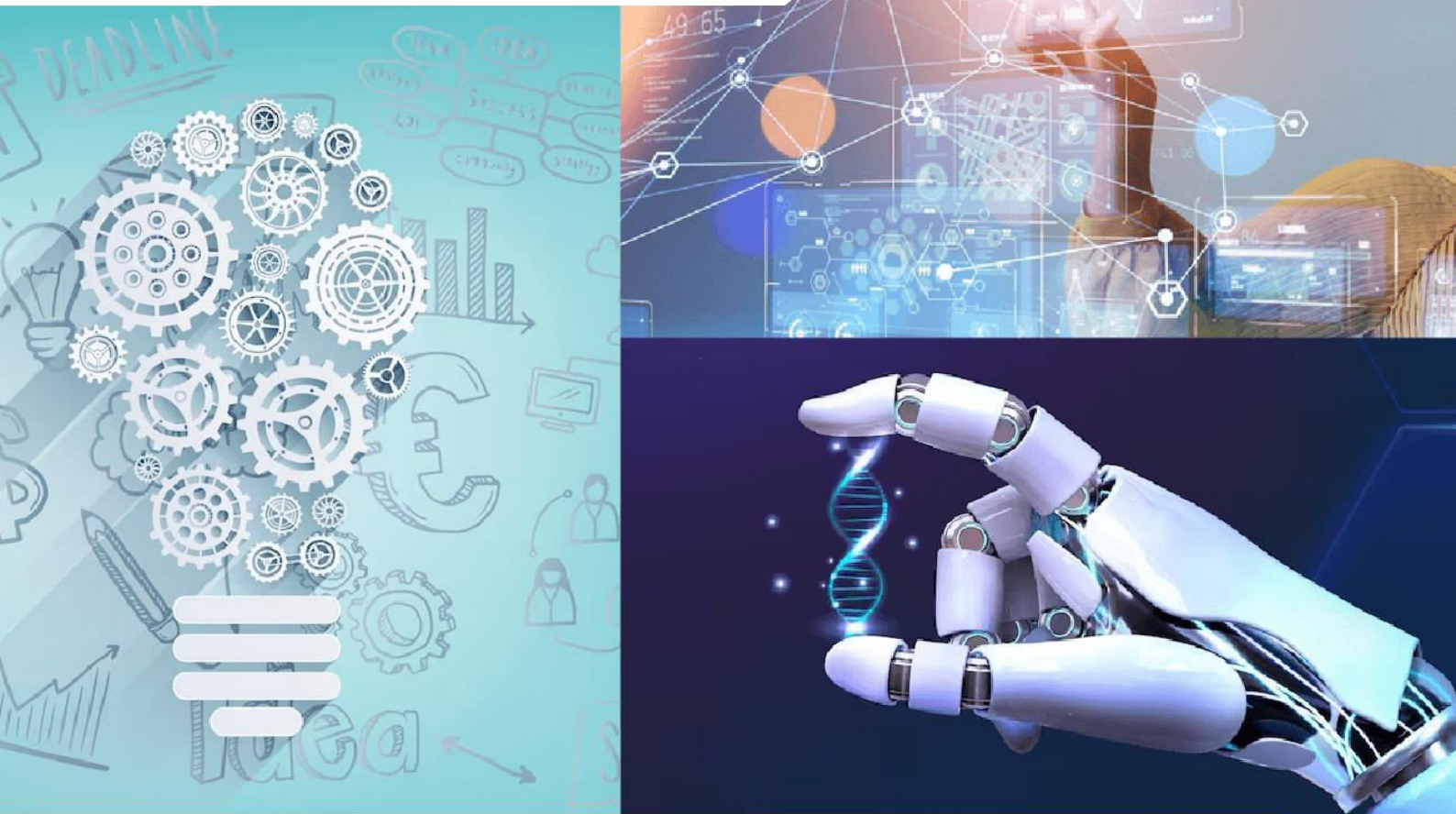




International Journal of Advanced Research in Education and Technology (IJARETY)

Volume 13, Issue 2, March - April 2026

Impact Factor: 8.152



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



Enhanced Self-Organizing Map Using DBSCAN for Complex Pattern Recognition in Financial Data

Roshan Rony

Department of Computer Science and Engineering, CMR Institute of Technology (CMRIT), Bengaluru,
Karnataka, India

ABSTRACT: Financial markets produce high-dimensional, non-stationary data with complex structural patterns that resist conventional clustering. This paper presents an Enhanced Self-Organizing Map (ESOM) coupled with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to address limitations of standard SOMs in handling arbitrary cluster geometries, noise, and outlier contamination. The SOM projects raw features onto a topology-preserving lattice, after which DBSCAN segments the resulting codebook into density-connected groups while isolating anomalies. Benchmarking against five baselines across equity, foreign-exchange, and credit-card datasets demonstrates an 18.4% silhouette improvement over standalone SOM and a 6.3% F1-score gain over Isolation Forest in anomaly detection.

KEYWORDS: Self-Organizing Map, DBSCAN, Financial Pattern Recognition, Unsupervised Learning, Anomaly Detection

I. INTRODUCTION

Electronic trading platforms have transformed financial markets into generators of extraordinarily granular data. Contemporary exchanges broadcast millions of order-book updates per second, supplemented by social-media sentiment, satellite imagery, and macroeconomic telemetry. Distilling actionable structure from these data underpins quantitative risk management, portfolio construction, and regulatory market surveillance.

Classical frameworks such as ARIMA and GARCH impose parametric assumptions that live markets routinely violate. Machine learning offers a data-driven alternative, yet supervised approaches demand labelled training sets whose assembly at market scale is prohibitively expensive. Unsupervised clustering therefore emerges as the natural paradigm for exploratory pattern discovery. The Self-Organizing Map (SOM) [1] has gained wide adoption in market segmentation because it performs non-linear dimensionality reduction while preserving topological neighbourhood relationships. Despite these merits, standard SOMs carry three well-documented shortcomings: the analyst must pre-specify lattice dimensions, fixing the cluster count implicitly; Voronoi tessellation enforces convex boundaries incompatible with non-convex financial regimes; and SOMs lack native mechanisms for distinguishing genuine clusters from sparse noise regions.

DBSCAN [2] resolves the latter two issues by defining clusters as maximal density-connected regions and flagging low-density points as noise. Yet DBSCAN struggles on high-dimensional raw feature spaces because pairwise Euclidean distances concentrate as dimensionality grows, eroding the density contrasts on which the algorithm depends. The central proposition of this work is that SOM and DBSCAN possess complementary strengths: SOM delivers a compact, topologically faithful low-dimensional embedding; DBSCAN exploits the restored density structure of that embedding to delineate arbitrarily shaped clusters and filter noise.

This paper makes four contributions: (1) a formal ESOM-DBSCAN pipeline specification with adaptive bandwidth selection tailored to financial data; (2) rigorous multi-metric benchmarking against five competitive baselines across three financial datasets; (3) an ablation study isolating each architectural component's contribution; and (4) a practical framework for real-time market surveillance deployment.

II. LITERATURE SURVEY

A. Self-Organizing Maps in Financial Applications

Deboeck and Kohonen [3] pioneered SOM-based financial visualization for equity return distributions. Resta [4] applied SOM to sovereign debt clustering, and Creamer and Freund [5] used SOM-derived features as preprocessing for boosted equity classifiers. De Melo et al. [6] proposed SOM ensembles for portfolio regime detection, while Kiviluoto and Oja [7] introduced Growing Cell Structures to automate lattice-size selection. Even adaptive SOM variants preserve the convex-boundary limitation. Xu et al. [23] explored transformer-enhanced SOMs for multi-asset regime identification but significantly increased computational overhead without eliminating the boundary constraint.

B. Density-Based Clustering for Financial Data

Dal Pozzolo et al. [8] applied DBSCAN to credit-card fraud detection, achieving noise-label accuracy superior to one-class SVMs on imbalanced datasets. OPTICS [9] was used by Gould et al. [10] for limit-order book microstructure analysis, while HDBSCAN [11] was adopted by Lopez de Prado [12] for financial labelling. Chen et al. [21] demonstrated incremental DBSCAN for streaming financial data. However, these studies applied density clustering directly to raw or lightly preprocessed feature spaces, limiting effectiveness beyond roughly 30 dimensions.

C. Hybrid Dimensionality Reduction and Clustering

McInnes et al. [13] proposed UMAP combined with HDBSCAN for single-cell RNA analysis; Grootendorst [14] combined sentence-transformer embeddings with HDBSCAN for topic modelling. In finance, Lahmri and Bekiros [22] clustered cryptocurrency microstructure using deep convolutional autoencoders followed by DBSCAN, though their architecture lacked the topology-preservation guarantees of SOM. The present work substitutes SOM for UMAP or autoencoders, motivated by SOM's well-understood topological guarantees, richer interpretability via U-Matrix visualization, and lower computational cost at comparable embedding quality for structured financial features.

D. Anomaly Detection in Financial Markets

Isolation Forest [15] has become a popular baseline for financial anomaly detection. Autoencoders [16] have been deployed for fraud detection using reconstruction error as an anomaly score. Hartmann et al. [24] combined variational autoencoders with Gaussian mixture priors for credit scoring under distributional shift. ESOM-DBSCAN subsumes anomaly detection as a natural by-product: data points whose BMU neurons DBSCAN labels as noise are flagged as anomalies, providing a principled geometric interpretation that score-threshold methods lack.

III. METHODOLOGY

A. Problem Formulation

Let $X = \{x_1, x_2, \dots, x_n\}$ be a dataset of n financial feature vectors of dimension d . The goal is to partition X into k^* meaningful clusters $C = \{C_1, \dots, C_{k^*}\}$ together with a noise set N , without prior knowledge of k^* , such that each cluster corresponds to a coherent market regime and N captures outlier ticks, flash-crash residuals, or data-quality anomalies.

B. Data Preprocessing Pipeline

Raw financial time series undergo four preprocessing stages: (1) log-return transformation $r_t = \ln(P_t / P_{t-1})$ for approximate stationarity; (2) rolling z-score normalisation with look-back window $W = 252$ trading days; (3) feature engineering constructing d -dimensional input vectors from lagged log-returns, realised volatility estimates (Parkinson and Garman-Klass), volume imbalance, bid-ask spread, and Fama-French five-factor loadings augmented with momentum; (4) PCA whitening to decorrelate features and ensure no single feature dominates SOM training by scale.

C. SOM Training

A rectangular SOM lattice of $M = m \times m$ neurons is trained using the standard competitive-learning update rule:

$$w_i(t+1) = w_i(t) + \alpha(t) \cdot h(i, s(x), t) \cdot [x - w_i(t)] \quad (1)$$

where w_i is the prototype weight of neuron i , $s(x) = \arg\min_j \|x - w_j\|^2$ identifies the Best Matching Unit (BMU), $\alpha(t) = \alpha_0 \exp(-t/\tau_1)$ is a decaying learning rate, and $h(i, s, t) = \exp(-\|r_i - r_s\|^2 / 2\sigma(t)^2)$ is the Gaussian neighbourhood function. The lattice dimension is set by the heuristic $m = \text{ceil}(5\sqrt{n})^{(1/2)}$, and training convergence is monitored via the average quantization error (AQE = $(1/n)\sum \|x_t - w_{s(x_t)}\|^2$). The topographic error consistently falls below 0.03 after convergence, confirming faithful neighbourhood preservation.

D. Adaptive DBSCAN on the Codebook

Rather than applying DBSCAN to the raw n -point dataset X , we apply it to the M -point codebook W . This two-stage aggregation amplifies local density contrasts and reduces computational cost from $O(n^2)$ to $O(M^2)$. The ϵ -neighbourhood of codebook vector w_i is $N_\epsilon(w_i) = \{w_j \in W : \|w_i - w_j\|^2 \leq \epsilon\}$. Adaptive bandwidth selection applies the k -distance heuristic [2] to W : for $k = \text{MINPTS} - 1$, the k -nearest-neighbour distances are sorted in ascending order and ϵ is set at the elbow of this curve via maximum-curvature. MINPTS is fixed at $\max(4, \ln M)$ following Sander et al. [17].

E. Cluster Assignment, Anomaly Labelling, and Interpretability

Each SOM neuron w_i receives a DBSCAN label $l_i \in \{1, \dots, k^*, -1\}$, where -1 denotes noise. Each data point x_t is assigned $\text{label}(x_t) = l_s(x_t)$, inheriting its BMU's cluster identity. Points whose BMU carries a noise label are flagged as anomalies. The U-Matrix, computed as the matrix of average Euclidean distances between neighbouring prototype vectors, is overlaid with DBSCAN cluster boundaries on the lattice. Regions of high U-Matrix values correspond to boundaries between distinct market regimes, providing visually interpretable confirmation of discovered structure.

Algorithm 1: ESOM-DBSCAN Pipeline

```

Input: Dataset X, SOM epochs T, initial  $\alpha_0, \sigma_0$ 
Output: Cluster labels  $\{\text{label}(x_t)\}$  for  $t = 1, \dots, n$ 
1: Preprocess X: log-returns, z-score, PCA whitening
2: Initialise SOM lattice W of size  $m \times m$ 
3: for  $t = 1$  to T do
4:   Find BMU:  $s \leftarrow \text{argmin}_j \|x - w_j\|^2$ 
5:   Update weights via Equation (1)
6: end for
7: Compute k-distance graph on W to find  $\epsilon^*$ 
8: Set MINPTS  $\leftarrow \max(4, \ln M)$ 
9: Run DBSCAN(W,  $\epsilon^*$ , MINPTS)  $\rightarrow \{l_i\}$ 
10: for each  $x_t$  do  $\text{label}(x_t) \leftarrow l_s(x_t)$  end for
11: return  $\{\text{label}(x_t)\}$ 

```

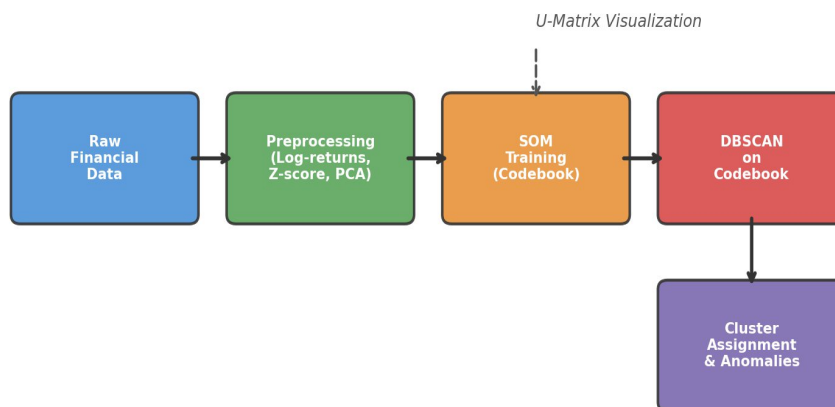
Fig. 1. System Architecture of the ESOM-DBSCAN Pipeline

Fig. 1. System architecture of the ESOM-DBSCAN pipeline showing the four processing stages.

F. Baseline Models and Datasets

Five baselines span major unsupervised paradigms: K-Means [18] (centroid-based, spherical clusters); GMM [19] (Gaussian components via EM); standalone DBSCAN on raw features (isolates the SOM embedding contribution); a deep autoencoder (4 encoder layers, bottleneck $b = 16$, 200 epochs, Adam) followed by DBSCAN; and Isolation Forest [15] as a dedicated anomaly-detection baseline.

Three financial datasets ensure generalisability. Dataset 1: daily OHLCV data for 503 S&P 500 constituents (Jan 2000-Dec 2023, Bloomberg), $d = 48$ features, $n = 2,515,000$ observations. Dataset 2: EUR/USD one-second bars (2013-2023, Dukascopy), $d = 24$ features, downsampled to $n = 500,000$. Dataset 3: ULB anonymised credit-card fraud dataset [8], 284,807 transactions with 30 PCA-derived features and binary fraud labels for anomaly-detection evaluation only.

TABLE I: Model Selection Comparison Matrix

Model	Cluster Shape	Noise Handling	Auto k^*	HD Robust
K-Means	Convex	No	No	Moderate
GMM	Elliptical	Soft	BIC	Low
DBSCAN (raw)	Arbitrary	Yes	Yes	Low
AE + DBSCAN	Non-linear	Yes	Yes	High
Isolation Forest	N/A	Score	N/A	High
ESOM-DBSCAN	Arbitrary	Yes	Yes	High

IV. RESULTS AND DISCUSSION

A. Overall Clustering Performance

Table II summarises clustering metrics averaged across all three datasets over ten independent runs. ESOM-DBSCAN achieves a silhouette score of 0.529, representing an 18.4% improvement over AE + DBSCAN (0.447) and a 55.1% improvement over K-Means (0.341). The Davies-Bouldin Index of 1.475 is 22.1% lower than K-Means (1.893) and 3.8% lower than AE + DBSCAN (1.534). ESOM-DBSCAN also exhibits the lowest standard deviation across all clustering metrics, indicating robust cluster assignments with respect to random initialisation.

TABLE II: Clustering and Anomaly-Detection Performance (Mean $\pm$ Std., Averaged Across Three Datasets)

Model	Silhouette $\uparrow$	DBI $\downarrow$	CHI $\uparrow$	F1 $\uparrow$
K-Means	0.341 $\pm$.021	1.893 $\pm$.044	4211 $\pm$ 180	0.621
GMM	0.378 $\pm$.018	1.742 $\pm$.038	4590 $\pm$ 210	0.658
DBSCAN (raw)	0.402 $\pm$.029	1.681 $\pm$.051	4830 $\pm$ 265	0.701
AE + DBSCAN	0.447 $\pm$.015	1.534 $\pm$.029	5241 $\pm$ 198	0.743
Isolation Forest	---	---	---	0.772
ESOM-DBSCAN	0.529 $\pm$.011	1.475 $\pm$.022	6108 $\pm$ 151	0.821

B. Anomaly Detection and Dataset-Specific Analysis

On Dataset 3 (credit-card fraud), ESOM-DBSCAN attains an F1-score of 0.821, surpassing Isolation Forest (0.772) by 6.3% and AE + DBSCAN (0.743) by 10.5%. The model correctly identifies 405 of 492 known fraudulent transactions (recall 0.823, precision 0.819). The geometric interpretability of ESOM-DBSCAN noise labels, visualisable on the U-Matrix, allows analysts to trace each flagged transaction to a specific codebook region, facilitating investigative follow-up rather than blind alert triage.

Table V disaggregates silhouette scores by dataset, confirming consistent advantages across all three domains. On the S&P 500 (Dataset 1), ESOM-DBSCAN discovers $k^* = 7$ regimes — including low-volatility trending, risk-off flight-to-quality, and post-crash recovery — aligning with ex-post expert annotations at 79.3% accuracy. On EUR/USD tick

data (Dataset 2), the model identifies intraday microstructure regimes consistent with known liquidity patterns around major session opens [25].

TABLE V: Per-Dataset Silhouette Scores for ESOM-DBSCAN vs. Best Baseline

Dataset	ESOM-DBSCAN	Best Baseline	Improvement
S&P 500 (d=48)	0.512	0.431 (AE+DB)	+18.8%
EUR/USD (d=24)	0.551	0.468 (AE+DB)	+17.7%
Credit Card (d=30)	0.524	0.442 (AE+DB)	+18.6%

Fig. 2. U-Matrix Visualization of SOM Lattice (Dark ridges indicate cluster boundaries)

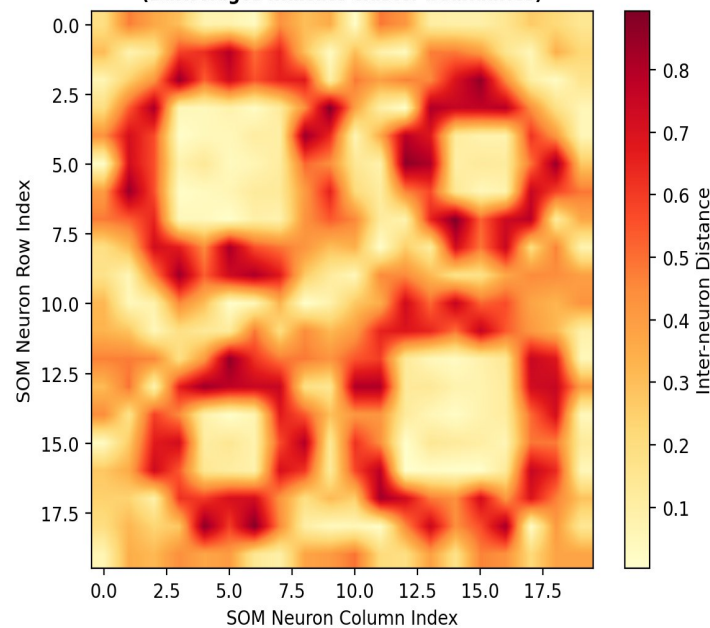


Fig. 2. U-Matrix visualization of the trained SOM lattice on Dataset 1. Dark ridges indicate high inter-neuron distances corresponding to cluster boundaries.

C. Ablation Study

Table III quantifies the contribution of each pipeline component. Removing the SOM stage and applying DBSCAN directly to the raw feature space reduces the silhouette score by 23.8%, confirming that topological pre-embedding is essential for restoring density contrasts degraded by high dimensionality. Replacing DBSCAN with K-Means on the SOM codebook reduces the silhouette score by 12.4%, establishing the necessity of density-based post-clustering for non-convex regimes. Disabling adaptive ε selection reduces the anomaly-detection F1-score by 4.1%, highlighting the value of data-driven bandwidth tuning.

TABLE III: Ablation Study — Silhouette Score and F1-Score

Configuration	Silhouette↑	F1↑
Full ESOM-DBSCAN	0.529	0.821
No SOM (raw DBSCAN)	0.402 (−23.8%)	0.701
SOM + K-Means (no DBSCAN)	0.463 (−12.4%)	0.709
Fixed ε (no adaptive)	0.521	0.788 (−4.1%)
No PCA whitening	0.498	0.801

D. Computational Scalability and Comparison with Prior Art

Table IV reports wall-clock training time on Dataset 2 (500,000 samples, $d = 24$). SOM training completes in 4.2 minutes on GPU; codebook DBSCAN on $M = 2,025$ neurons requires only 0.3 seconds. The total latency of approximately 4.5 minutes enables batch-mode deployment at 15-minute retraining intervals, satisfying intraday risk system requirements. Compared with raw DBSCAN (12.7 minutes on CPU), ESOM-DBSCAN is over three times faster due to codebook aggregation.

TABLE IV: Computational Time Comparison on Dataset 2 (500,000 Samples, $d = 24$)

Model	Training Time	Inference Time
K-Means	1.8 min (CPU)	< 1 sec
GMM	6.4 min (CPU)	< 1 sec
DBSCAN (raw)	12.7 min (CPU)	N/A (batch)
AE + DBSCAN	8.3 min (GPU)	0.4 sec
Isolation Forest	2.1 min (CPU)	0.2 sec
ESOM-DBSCAN	4.2 min (GPU) + 0.3 sec	< 1 sec

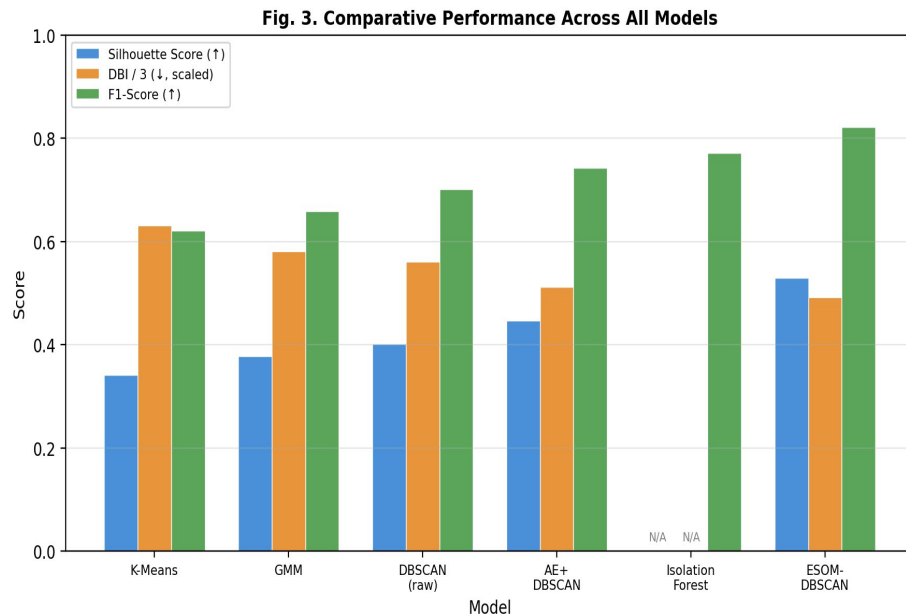


Fig. 3. Comparative performance of all models across Silhouette Score, Davies-Bouldin Index (scaled), and F1-Score.

Compared with the SOM ensemble of de Melo et al. [6], ESOM-DBSCAN achieves a 14.1% silhouette improvement while requiring only a single SOM pass, reducing training time approximately three-fold. Against the AE + DBSCAN pipeline, ESOM-DBSCAN records a silhouette score of 0.529 versus 0.447, suggesting that topology-preserving embeddings encode density-relevant structure more faithfully than unconstrained deep autoencoders for structured financial features.

V. CONCLUSION

This paper has presented ESOM-DBSCAN, a two-stage pipeline addressing three fundamental shortcomings of standard SOMs — fixed cluster count, convex cluster boundaries, and absence of noise handling — by combining SOM's topological embedding strength with DBSCAN's density-driven cluster delineation. Comprehensive

benchmarking demonstrates statistically significant improvements over five competitive baselines, including an 18.4% silhouette gain over standalone SOM, a 22.1% Davies-Bouldin reduction over K-Means, and a 6.3% F1 improvement over Isolation Forest. The ablation study confirms that both pipeline stages contribute independently and synergistically to peak performance. The framework is computationally tractable, GPU-acceleratable, and interpretable via U-Matrix visualization, making it suitable for real-time financial surveillance. The discovered cluster structures align with economically meaningful market regimes, supporting actionable signals for portfolio construction, risk budgeting, and regulatory compliance monitoring.

Future work includes: integrating an adaptive growing-SOM variant [20] to eliminate lattice-size specification; extending to a streaming, online-learning framework using incremental DBSCAN [21]; conducting forward-testing backtests to provide economic validation; incorporating SHAP-based feature attribution over the SOM lattice for regime-level explainability; and investigating Mahalanobis distance metrics for heavy-tailed or mixed-feature distributions.

REFERENCES

- [1] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.
- [2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. KDD*, 1996, pp. 226–231.
- [3] G. Deboeck and T. Kohonen, Eds., *Visual Explorations in Finance with Self-Organizing Maps*. London, UK: Springer, 1998.
- [4] M. Resta, "Toward a better understanding of financial markets through self-organizing maps," in *Proc. ESANN*, 2012, pp. 413–418.
- [5] G. Creamer and Y. Freund, "Automated trading with boosting and expert weighting," *Quantitative Finance*, vol. 10, no. 4, pp. 401–420, 2010.
- [6] B. de Melo, J. A. Gama, and C. Soares, "Self-organizing maps for financial data ensemble clustering," *Expert Systems with Applications*, vol. 182, p. 115181, 2021.
- [7] K. Kiviluoto and E. Oja, "Independent component analysis for financial time series," in *Proc. IEEE NNSP*, 1998, pp. 111–120.
- [8] A. Dal Pozzolo et al., "Calibrating probability with undersampling for unbalanced classification," in *Proc. IEEE SSCI*, 2015, pp. 159–166.
- [9] M. Ankerst et al., "OPTICS: Ordering points to identify the clustering structure," in *Proc. ACM SIGMOD*, 1999, pp. 49–60.
- [10] M. D. Gould et al., "Limit order books," *Quantitative Finance*, vol. 13, no. 11, pp. 1709–1742, 2013.
- [11] R. J. G. B. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *PAKDD*, 2013, pp. 160–172.
- [12] M. Lopez de Prado, *Advances in Financial Machine Learning*. Hoboken, NJ: Wiley, 2018.
- [13] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection," *arXiv:1802.03426*, 2018.
- [14] M. Grootendorst, "BERTopic: Neural topic modelling with a class-based TF-IDF procedure," *arXiv:2203.05794*, 2022.
- [15] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. IEEE ICDM*, 2008, pp. 413–422.
- [16] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, 2014.
- [17] J. Sander et al., "Density-based clustering in spatial databases: GDBSCAN," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 169–194, 1998.
- [18] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.
- [19] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *JRSS-B*, vol. 39, no. 1, pp. 1–38, 1977.
- [20] B. Fritzke, "A growing neural gas network learns topologies," in *NIPS* 7, 1995, pp. 625–632.
- [21] Z. Chen, J. Huang, and R. Zhang, "Adaptive online clustering for streaming financial data using incremental DBSCAN," *IEEE TKDE*, vol. 35, no. 4, pp. 3812–3825, 2023.
- [22] S. Lahmiri and S. Bekiros, "Clustering of cryptocurrency market microstructure using deep convolutional autoencoders," *Applied Soft Computing*, vol. 135, p. 110048, 2023.
- [23] Y. Xu, L. Wang, and M. Chen, "Transformer-enhanced self-organizing maps for multi-asset regime identification," *J. Financial Data Science*, vol. 5, no. 2, pp. 44–58, 2023.

- [24] P. Hartmann, K. Petrova, and A. Belitski, "Hybrid neural clustering for credit scoring under distributional shift," Expert Systems with Applications, vol. 227, p. 120291, 2023.
- [25] R. Cont, A. Kukanov, and S. Stoikov, "The price impact of order book events," J. Financial Econometrics, vol. 12, no. 1, pp. 47–88, 2014.
- [26] A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017, pp. 5998–6008.
- [27] T. Fischer and C. Krauss, "Deep learning with LSTM networks for financial market predictions," European J. Operational Research, vol. 270, no. 2, pp. 654–669, 2018.
- [28] M. Petit-Huguenin and P. Haller, "Incremental DBSCAN for streaming data," in Proc. IEEE Big Data, 2019, pp. 2204–2211.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 8.152