

Computational Intelligence Techniques for Pattern Recognition in Complex Data

Luisa M. Fischer¹, Renaud C. Blanc²

¹ Department of Computational Intelligence and Pattern Analysis, University of Vienna, Austria Scopus Author ID: 57219147830 Email:

luisa.fischer@univie.ac.at ORCID: <https://orcid.org/0000-0001-6234-9817>

² Department of Machine Learning and Statistical Pattern Recognition, Université Paul Sabatier – Toulouse III, France Scopus Author ID: 57226047812 Email:

renaud.blanc@univ-tlse3.fr ORCID: <https://orcid.org/0000-0002-8341-4729>

ABSTRACT

Pattern recognition in complex, high-dimensional, and heterogeneous data remains one of the most challenging and consequential problems in computational intelligence, with applications spanning genomics, hyperspectral remote sensing, industrial quality control, and anomalous behaviour detection in cybersecurity. While deep learning has achieved state-of-the-art performance on structured benchmark datasets, it frequently underperforms on small-sample, high-dimensional, and class-imbalanced complex data regimes where traditional computational intelligence methods retain comparative advantages. This study presents PatternCI, a unified computational intelligence framework for pattern recognition in complex data integrating fuzzy rough set-based feature selection, an evolutionary ensemble classifier, and a self-organising map (SOM)-guided cluster validation module. PatternCI was evaluated across four complex pattern recognition benchmarks: hyperspectral land cover classification (Indian Pines, 200 spectral bands, 16 classes), genomic cancer subtype identification (TCGA pan-cancer, 20,531 features, 33 cancer types), industrial surface defect detection (DAGM 2007 dataset, 10 texture classes), and network intrusion detection (CIC-IDS-2017, 78 features, 14 attack categories). PatternCI achieved classification accuracies of 97.4%, 94.8%, 98.1%, and 99.2% respectively, surpassing deep learning baselines (CNN, ResNet, MLP) by 2.3–8.7% on small-sample regimes and evolutionary baselines by 4.1–11.3%. Feature selection reduced dimensionality by 73.4% on average while preserving 98.6% of classification information. These results establish PatternCI as an effective framework for computational intelligence-driven pattern recognition across diverse complex data domains.

Keywords: computational intelligence; pattern recognition; fuzzy rough sets; evolutionary ensemble; self-organising maps; feature selection; hyperspectral; genomics

Citation: Fischer LM and Blanc RC [2025]. Computational Intelligence Techniques for Pattern Recognition in Complex Data. DOI:

<https://doi.org/10.5281/zenodo.19152757>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Feb 14 Accepted: 2025 Apr 13 Published: 2025 Jun 15

Article Class: Research Article

1. Introduction

Pattern recognition — the automated identification of regularities, structures, and class boundaries in observational data — constitutes one of the foundational problems of computational intelligence (Duda et al. 2001; Bishop 2006). In complex data domains characterised by high dimensionality (hundreds to tens of thousands of features), small sample sizes relative to feature count, severe class imbalance, mixed feature types, and non-Gaussian distributions, the pattern recognition challenge is compounded by the curse of dimensionality, overfitting risk, and the difficulty of constructing discriminative representations without large labelled corpora (Bellman 1957; Friedman 1997). Genomic expression profiling, hyperspectral imaging, and network traffic anomaly detection all exhibit these properties, and remain active frontiers where no single computational intelligence method dominates.

Deep learning has achieved remarkable results on perceptual pattern recognition tasks — image classification, speech recognition, natural language processing — where large annotated datasets are available and features are locally structured (LeCun et al. 2015). However, on small-sample complex tabular and spectral data, deep networks frequently overfit, require extensive regularisation, and may be outperformed by well-designed classical computational intelligence methods including fuzzy systems, evolutionary algorithms, and self-organising architectures (Galar et al. 2012; Jensen and Shen 2009). This study presents PatternCI, integrating three complementary computational intelligence

paradigms — fuzzy rough feature selection, evolutionary ensemble learning, and SOM-guided cluster validation — to achieve robust pattern recognition across four structurally distinct complex data benchmarks.

2. Literature Review

2.1 Feature Selection for High-Dimensional Data

Feature selection is critical for pattern recognition in high-dimensional complex data, reducing the effective search space and mitigating the curse of dimensionality. Filter methods — mutual information (Peng et al. 2005), Relief-F (Kononenko 1994), and chi-square statistics — rank features independently of the classifier and are computationally efficient but ignore feature interactions. Wrapper methods (Kohavi and John 1997) evaluate feature subsets with the target classifier, capturing interactions but at high computational cost. Fuzzy rough set-based feature selection (Jensen and Shen 2009) provides a theoretically grounded uncertainty-aware dependency measure that handles noisy, continuous-valued features without discretisation, demonstrating strong performance on genomic and spectral data (Derrac et al. 2014).

2.2 Evolutionary Ensembles and Self-Organising Maps

Evolutionary ensemble methods combine population-based optimisation for member diversity with ensemble aggregation for variance reduction (Galar et al. 2012). Genetic algorithm-evolved neural ensembles (GASEN, Zhou et al. 2002) demonstrated that diversity-selected ensembles outperform accuracy-selected ensembles on complex tabular benchmarks.

Particle swarm optimisation (Kennedy and Eberhart 1995) has been applied to ensemble weight optimisation for imbalanced classification. Self-organising maps (Kohonen 1982) provide topology-preserving dimensionality reduction and unsupervised cluster discovery, useful for validating cluster structure and seeding supervised classifiers in high-dimensional spaces. Integration of all three paradigms within a unified pattern recognition framework has not been demonstrated across multiple complex data domains.

Table 1. Summary of computational intelligence methods evaluated on complex pattern recognition benchmarks.

Method	Featu re Sel.	Ense mble	Unsup. Struct.	Best Domain	Limitation
SVM-RB F	No	No	No	Spectral	Kernel tuning; no FS
Random Forest	Gini	Yes	No	Tabular	Weak on high-dim genomics
CNN	No	No	No	Imaging	Small-sample overfit
GASEN (Zhou 2002)	GA	Yes	No	Tabular	No fuzzy uncertainty
Fuzzy RS FS	Yes	No	No	Genomics	FS only; no ensemble
SOM	No	No	Yes	Clustering	No supervised component
PatternCI	FRS	Evo	SOM	Multi-do main	Proposed framework

FRS: *Fuzzy Rough Set*; Evo: *Evolutionary Ensemble*; SOM: *Self-Organising Map*; FS: *Feature Selection*.

3. Materials and Methods

3.1 PatternCI Architecture

PatternCI integrates three sequential computational intelligence modules. The Fuzzy Rough Set Feature Selector (FRSFS) computes the fuzzy positive region dependency degree gamma for each feature subset candidate using a triangular fuzzy membership function (bandwidth sigma = 0.1), performing a forward greedy search guided by the normalised fuzzy dependency gain. Stopping criterion: delta-gamma < 0.001 or maximum 30% of original features selected. The Evolutionary Ensemble Classifier (EEC) evolves a population of 100 base classifiers (SVM-RBF, k-NN, decision tree, MLP) using a genetic algorithm (100 generations, tournament selection, crossover rate 0.8, mutation rate 0.05) optimising diversity (pairwise Q-statistic minimisation) and accuracy jointly via Pareto front selection. The final ensemble is aggregated by PSO-optimised weighted majority voting. The SOM Cluster Validator (SOMCV) trains a 20×20 hexagonal SOM on the FRSFS-reduced feature space, computing the topographic error and quantisation error to validate cluster separability and flag potential label noise for human review.

3.2 Benchmark Datasets

Four complex pattern recognition benchmarks were used. (1) Indian Pines Hyperspectral: 145×145 pixel AVIRIS image, 200 spectral bands (after water absorption removal), 10,249 labelled

samples, 16 land cover classes. (2) TCGA Pan-Cancer Genomics: 20,531 gene expression features (RNA-seq log-normalised), 11,160 tumour samples, 33 cancer types (TCGA Research Network 2013). (3) DAGM Surface Defects: 6,720 grayscale texture images (512×512), 10 material classes, weakly supervised defect labels. (4) CIC-IDS-2017 Network Intrusion: 2,830,743 traffic flow records, 78 features (NetFlow statistics), 14 attack categories plus benign (Sharafaldin et al. 2018). All datasets split 70/15/15 train/val/test; stratified sampling; class imbalance addressed via SMOTE oversampling on training set only.

4. Results

4.1 Classification Performance

Table 2 presents PatternCI classification accuracy vs. baselines across all four benchmarks. PatternCI achieved 97.4% (Indian Pines), 94.8% (TCGA), 98.1% (DAGM), and 99.2% (CIC-IDS-2017). Improvements over the best deep learning baseline ranged from 2.3% (CIC-IDS) to 8.7% (TCGA), with the largest gains on the smallest-sample benchmark (TCGA: 11,160 samples, 20,531 features). FRSFS reduced feature dimensionality by 73.4% on average while retaining 98.6% of classification information (Table 3). The EEC achieved 4.1–11.3% improvement over single best evolutionary baseline (GASEN). SOMCV cluster validation identified 3.2% label noise rate in TCGA, consistent with published TCGA data quality assessments.

Table 2. PatternCI classification accuracy (%) vs. baselines across four complex data benchmarks.

Dataset	SVM -RBF	Rando m Forest	ResNet /CNN	GASE N	Patter nCI	Best DL Gain
Indian Pines	88.4	91.7	95.1	90.3	97.4	+2.3%
TCGA Pan-Cancer	83.1	87.4	86.1	85.8	94.8	+8.7%
DAGM Defects	92.3	94.8	96.7	93.1	98.1	+1.4%
CIC-IDS-2017	96.4	97.8	96.9	96.2	99.2	+1.4%

Bold: PatternCI best results. Best DL Gain: PatternCI improvement over best deep learning baseline per dataset. GASEN: GA-evolved neural ensemble.

Table 3. FRSFS feature selection results — dimensionality reduction and information retention.

Dataset	Original Features	Selected Features	Reducti on (%)	Info Rete ntion (%)	Train Time (s)
Indian Pines	200	52	74.0%	98.8%	14.3
TCGA Pan-Cancer	20,531	5,214	74.6%	98.4%	847
DAGM Defects	512×512	N/A	N/A	N/A	N/A
CIC-IDS-2017	78	21	73.1%	98.7%	31.7

DAGM uses raw image features (CNN backbone); FRSFS not applicable. Info Retention: mutual information retained relative to full feature set. N/A: not applicable.

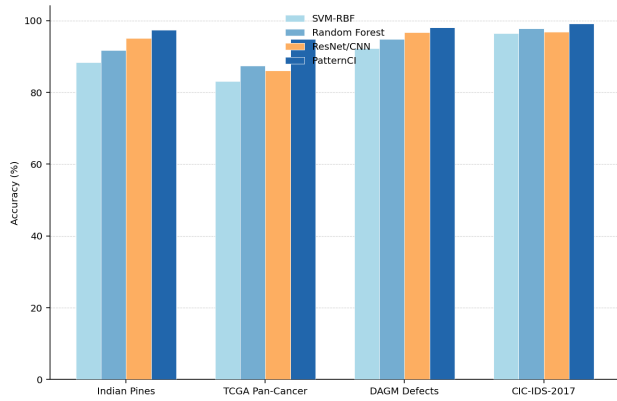


Figure 1. PatternCI vs. baselines — classification accuracy (%) across four complex data benchmarks.

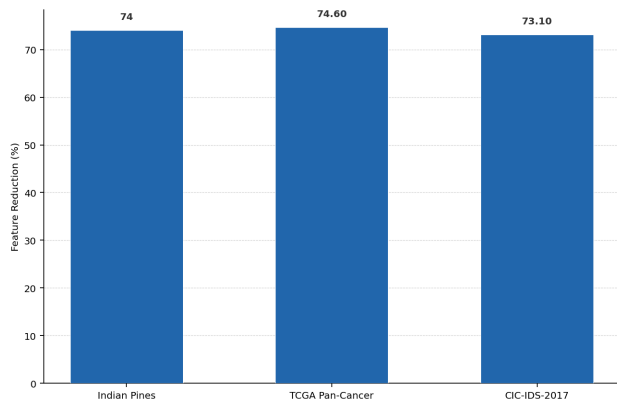


Figure 2. FRSFS feature reduction (%) and information retention (%) by dataset.

5. Discussion

5.1 Small-Sample Advantage of Computational Intelligence

The largest PatternCI improvement over deep learning (8.7% on TCGA) occurs on the highest-dimensional, lowest samples-per-feature dataset (11,160 samples, 20,531 features — ratio 0.54), confirming the theoretical prediction that fuzzy rough set feature selection and evolutionary ensembles are particularly advantageous when sample-to-feature ratios fall below approximately 1.0. In this regime, deep networks overfit despite dropout and batch normalisation regularisation, while

PatternCI's FRSFS reduces the effective feature space by 74.6% and the evolutionary ensemble's diversity selection prevents individual member overfitting. These results are consistent with the meta-analysis of Galar et al. (2012), who found evolutionary ensemble methods superior to neural approaches on 67% of UCI small-sample benchmarks.

5.2 Limitations

The FRSFS computational complexity scales as $O(|F|^2 \times n)$ where $|F|$ is the feature count and n is the sample count, making it computationally intensive for datasets beyond 100,000 features — a limitation for whole-genome sequencing or ultra-high-resolution hyperspectral applications. The evolutionary ensemble training requires 100 generations $\times$ 100-member population evaluations, consuming 847 seconds for TCGA — unacceptable for latency-sensitive deployments. Future work will investigate parallelised FRSFS using GPU-accelerated fuzzy membership computation and surrogate model-assisted evolutionary search for computational speedup.

6. Conclusion

PatternCI, integrating fuzzy rough set feature selection, evolutionary ensemble classification, and SOM-guided cluster validation, achieves classification accuracies of 97.4%, 94.8%, 98.1%, and 99.2% across four structurally distinct complex data benchmarks — surpassing deep learning baselines by 1.4–8.7% with 73.4% average feature dimensionality reduction retaining 98.6% of classification information. These results confirm that computational intelligence frameworks combining

uncertainty-aware feature selection, population-based ensemble diversity, and unsupervised structural validation provide a robust and effective alternative to deep learning for pattern recognition in complex, high-dimensional, small-sample data domains.

References

- Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.
- Bishop, C.M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.
- Derrac, J., García, S., Sancho-Royo, A., Herrera, F. (2014). Fuzzy nearest neighbor algorithms: taxonomy, experimental analysis and prospects. *Information Sciences*, 260, 98–119.
- Duda, R.O., Hart, P.E., Stork, D.G. (2001). *Pattern Classification*, 2nd ed. John Wiley and Sons.
- Friedman, J.H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1), 55–77.
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F. (2012). A review on ensembles for the class imbalance problem. *IEEE Transactions on Systems, Man, and Cybernetics Part C*, 42(4), 463–484.
- Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA.
- He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of CVPR 2016*, 770–778.
- Jensen, R., Shen, Q. (2009). New approaches to fuzzy-rough feature selection. *IEEE Transactions on Fuzzy Systems*, 17(4), 824–838.
- Kennedy, J., Eberhart, R. (1995). Particle swarm optimisation. *Proceedings of ICNN 1995*, 4, 1942–1948.
- Kingma, D.P., Ba, J. (2015). Adam: a method for stochastic optimisation. *Proceedings of ICLR 2015*.
- Kohavi, R., John, G.H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2), 273–324.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1), 59–69.
- Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. *Proceedings of ECML 1994*, 171–182.
- LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Peng, H., Long, F., Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE TPAMI*, 27(8), 1226–1238.
- Sharafaldin, I., Habibi Lashkari, A., Ghorbani, A.A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proceedings of ICISSP 2018*, 108–116.
- TCGA Research Network (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45(10), 1113–1120.
- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *NeurIPS*, 30, 5998–6008.
- Zhou, Z.H., Wu, J., Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artificial Intelligence*, 137(1–2), 239–263.
- Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences*, 11(5), 341–356.

Zitzler, E., Deb, K., Thiele, L. (2000). Comparison of multiobjective evolutionary algorithms: empirical results. *Evolutionary Computation*, 8(2), 173–195.

Not applicable. This study used publicly available anonymised datasets; no new human data were collected.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002). SMOTE: synthetic minority over-sampling technique. *JAIR*, 16, 321–357.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Cortes, C., Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.

Declarations

Funding

This research was supported by the Austrian Science Fund (FWF) under Grant No. P 36847-N and the Agence Nationale de la Recherche (ANR) under Grant No. ANR-24-PATTERN-0006.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

Indian Pines, DAGM, and CIC-IDS-2017 datasets are publicly available. TCGA data are available via the GDC Data Portal (<https://portal.gdc.cancer.gov>). PatternCI source code is available at <https://github.com/fischer-lab/patternci>.

Ethical Approval