DBSCAN Controlled Target Classification Using Self Mixing with a Semiconductor Laser

Başak Ersöz
Politecnico di Torino
Torino, Italy
basak.ersozyildirim@polito.it

Priyanka Chaudhary
Università di Bari Aldo Moro
Bari, Italy
priyanka.chaudhary@uniba.it

Paolo Bardella
Politecnico di Torino
Torino, Italy
paolo.bardella@polito.it

Maurizio Dabbicco
Università di Bari Aldo Moro
Bari, Italy
maurizio.dabbicco@uniba.it

Lorenzo Columbo
Politecnico di Torino
Torino, Italy
lorenzo.columbo@polito.it

Massimo Brambilla
Politecnico di Bari
Bari, Italy
massimo.brambilla@poliba.it

***Abstract*—We present an experimental optical feedback (OF) platform operating in the self mixing (SM) configuration, where target back reflection is reinjected into a semiconductor laser cavity to produce SM waveforms under speckle illumination generated by a nanofiber diffuser. The acquired dataset refers to six targets (letters U,B,A,R,I and a Blank rejection class) and was recorded under realistic alignment and feedback fluctuations, where regime inconsistent measurements can dominate waveform statistics.**

Each waveform is mapped to a compact f_6 descriptor built from global statistics of the signal and its discrete time derivative, standardized, embedded in 2-D via t-distributed stochastic neighbor embedding (t-SNE), and filtered using Density-Based Spatial Clustering of Applications with Noise (DBSCAN), regime consistent cores while rejecting sparse outliers. We quantify the impact of the DBSCAN operating point by sweeping (epsilon, minPts) and defining four filtered datasets (A,B,C,D) spanning a retention selectivity tradeoff. We evaluate standard classifiers (logistic regression (LR), linear support vector machines (SVMs), decision tree, random forest (RF)) under a fixed 70/15/15 protocol. RF yields the best result, with the most selective operating point achieving accuracy 0.8482 and macro F1 0.8469. Overall, (epsilon, minPts) acts as a practical control knob: stricter cores reduce U and Blank confusions and improve stability, while less tolerant filtering boosts retention but increases cross class confusions.

Overall, DBSCAN based core selection improves robustness by removing regime inconsistent waveforms, yielding cleaner separation between informative targets and Blank while preserving most of the dataset. In realistic sensing, alignment drift and feedback fluctuations make reliable rejection of non informative measurements as important as target classification.

***Index Terms*—Self mixing, optical feedback, t SNE, DBSCAN, machine learning, hyperparameter optimization**

I. Introduction

We study target classification and rejection from experimental SM signals measured in an OF interferometric setup. A semiconductor laser (830 nm) illuminates printed targets through a nanofiber based diffuser that generates multiple speckle patterns as the diffuser position is scanned. For each illumination, target back reflection is reinjected into the laser cavity in the SM configuration [1] [2], producing a measured SM waveform without an external interferometer.

Figure 1 summarizes the measurement chain used to generate speckle illumination and acquire the waveforms analyzed in this work. In practice, SM signals are highly sensitive to alignment, feedback strength, target reflectivity, and speckle decorrelation [3], so experimental acquisitions often contain regime inconsistent measurements. When such heterogeneous data are used for supervised learning, regime variability can dominate the statistics and reduce robustness. This problem is especially relevant for sensing pipelines where the system must both classify informative targets and reject non informative measurements, a role captured here by the Blank class.

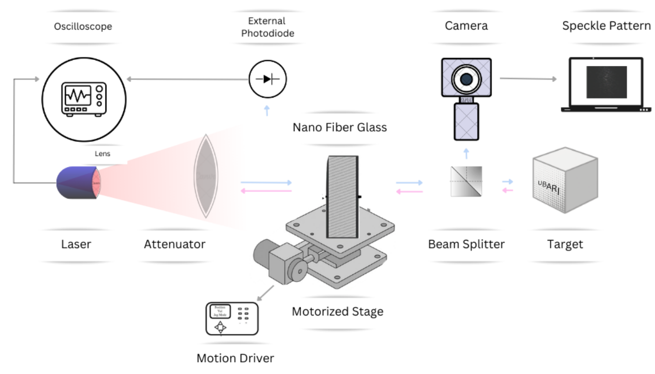
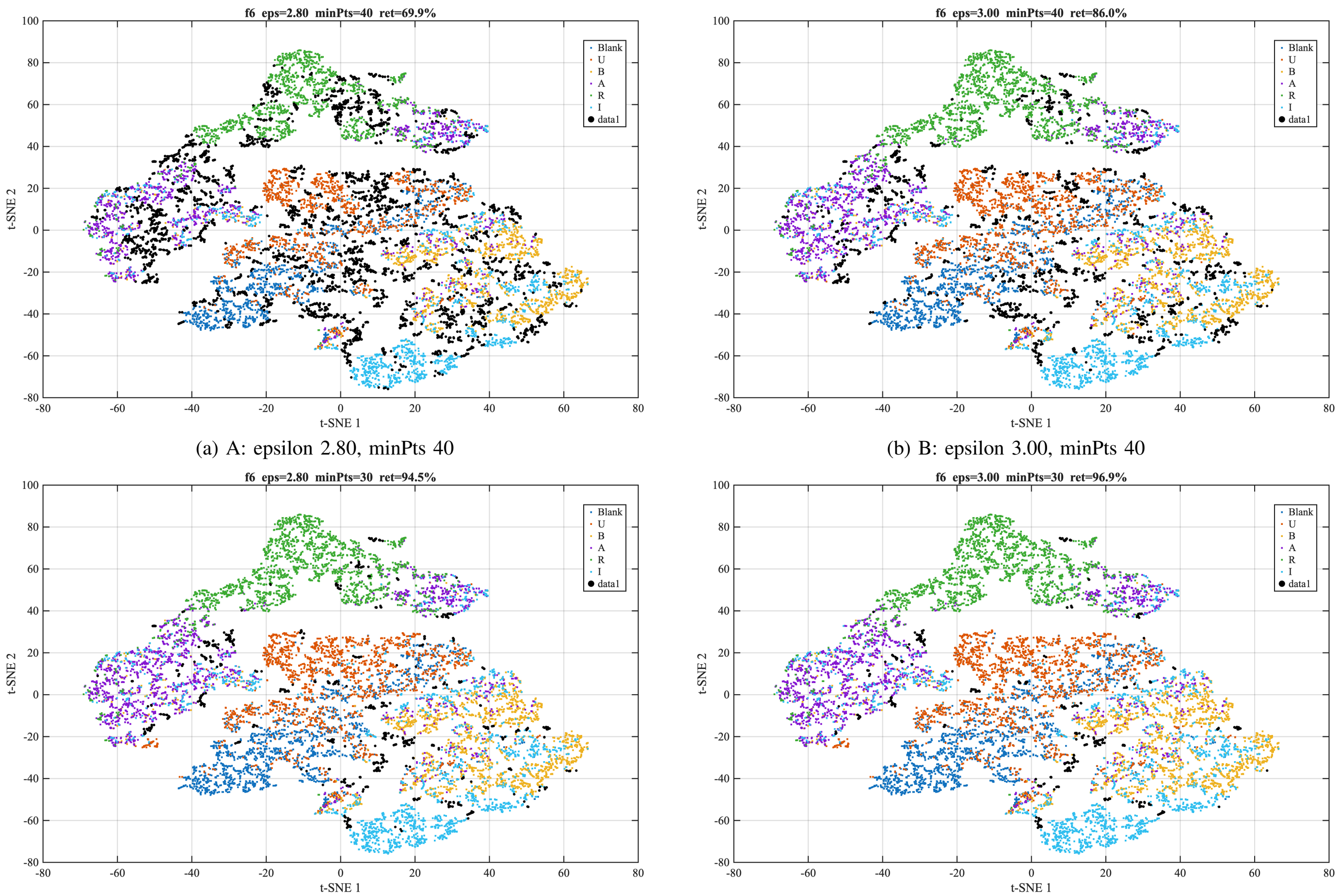


Fig. 1. Experimental optical feedback measurement chain. The attenuated laser output is split to illuminate the target through the nanofiber diffuser on a motorized stage (speckle scanning) and provide monitoring with an external photodiode. Light back reflected from the target is reinjected into the laser cavity (SM), and the resulting waveform is acquired by the oscilloscope. A CMOS camera records the speckle pattern.

As a proof of principle benchmark, we consider a six class dataset comprising five UBARI letters (U,B,A,R,I) and a Blank rejection class. Blank is practically necessary to reject non informative measurements, but it also increases task difficulty because its waveforms can overlap with true letters under imperfect optical feedback conditions, notably U versus Blank. In this work, we treat DBSCAN as a selectivity control knob that trades retention for regime consistency and class separability under imperfect experimental conditions.



(a) A: epsilon 2.80, minPts 40

(b) B: epsilon 3.00, minPts 40

(c) C: epsilon 2.80, minPts 30

(d) D: epsilon 3.00, minPts 30

Fig. 2. t SNE of standardized f6 descriptors with DBSCAN filtering for operating points A to D. Colored markers indicate targets; black markers are rejected as noise.

II. Experimental Setup and Dataset

A. Optical feedback acquisition and point by point scanning

The experiment uses a semiconductor laser operating at 830 nm in SM mode. A nanofiber diffuser is moved on a motorized stage to generate a sequence of distinct speckle illuminations. For each stage position, the speckle pattern illuminates the printed target, and the back reflected light is reinjected into the laser cavity. In parallel, a CMOS camera records the speckle pattern corresponding to the same diffuser position. The acquisition is performed point by point along the scan, intentionally exposing the measurement process to small alignment drift and feedback fluctuations, which leads to mixed regimes in the collected waveforms.

B. UBARI and Blank dataset

The dataset contains 12000 labeled waveforms, balanced across the six targets (2000 samples per class): Blank, U, B, A, R, and I. Blank samples correspond to non letter regions or non informative measurements. All reported results use the same fixed stratified 70/15/15 split into training, validation, and test sets.

C. Compact descriptors and standardization

Each waveform is mapped to a time independent descriptor using global statistics computed on the raw signal and on its discrete time derivative. We use three statistics: root mean square, variance, and peak to peak amplitude. Computing the same statistics on the derivative emphasizes rapid fluctuations and complements amplitude dominated information from the raw signal. The resulting six dimensional descriptor is denoted f_6.

Before distance based analysis and training linear models, each feature dimension is standardized using z score normalization computed on the training set only. The training mean and standard deviation are used to normalize training, validation, and test features.

III. Manifold Embedding and Density Core Regime Definition

Standardized f6 descriptors are embedded into two dimensions using t-SNE [4]. DBSCAN [5] is then applied in the embedded space with hyperparameters epsilon and minPts. Dense regions are treated as regime consistent cores, while sparse samples are rejected as noise.

TABLE I
DBSCAN OPERATING POINTS AND RETENTION FOR THE F6 MANIFOLD. COLUMNS: EPSILON, MINPTS, KEPT SAMPLES, AND RETENTION RATIO RHO.

Dataset	epsilon	minPts	Kept	rho
A	2.80	40	8388	0.699
B	3.00	40	10314	0.860
C	2.80	30	11335	0.945
D	3.00	30	11633	0.969

TABLE II
EFFECT OF DBSCAN OPERATING POINT ON PERFORMANCE. LOGISTIC REGRESSION AND RANDOM FOREST ARE REPORTED. SPLIT IS (NTRAIN/NVAL/NTEST).

		LR		RF	
Dataset	Split	Acc	Macro F1	Acc	Macro F1
A	5872/1258/1258	0.7154	0.6915	0.8482	0.8469
B	7220/1547/1547	0.7207	0.7055	0.8345	0.8341
C	7935/1700/1700	0.6812	0.6660	0.8171	0.8166
D	8144/1745/1744	0.7070	0.6946	0.8171	0.8172

TABLE III
BLANK REJECTION SUMMARY (RF, DATASET A). LETTERS ARE POOLED INTO A SINGLE TARGET CLASS. ROWS ARE TRUE; COLUMNS ARE PREDICTED.

	Pred Blank	Pred Target
True Blank	172	33
True Target	35	1018

TABLE IV
CONFUSION MATRIX (COUNTS) OF RANDOM FOREST ON DATASET A AFTER DBSCAN CORE SELECTION. ROWS: TRUE; COLUMNS: PREDICTED.

	Blank	U	B	A	R	I
Blank	172	30	0	0	0	3
U	34	159	1	0	0	2
B	1	0	187	19	0	17
A	0	0	16	167	12	7
R	0	0	0	12	208	0
I	0	7	18	12	0	174

IV. HYPERPARAMETER STUDY, PROTOCOL, AND METRICS

We quantify the effect of the DBSCAN operating point by defining four filtered datasets (A,B,C,D): A (2.80,40), B (3.00,40), C (2.80,30), and D (3.00,30). Decreasing minPts and increasing epsilon boost retention.

For objective validation of regime consistency, we evaluate standard supervised classifiers under a fixed stratified 70/15/15 split with a fixed seed. We report accuracy, macro F1, and retention ratio rho.

V. RESULTS AND CONCLUSION

In this work, with standart classifiers, RF consistently achieves the highest performance across operating points A-D (Table II). The best overall result is obtained for Dataset A (most conservative filtering), with $\mathrm{Acc} = 0.8482$ and $\text{macro-F1} = 0.8469$ (Table II). The pooled Blank rejection view in Table III summarizes the same outcome as a two class decision (Blank vs. Target), which matches practical sensing pipelines where non informative measurements (background, weak feedback, unstable regimes) must be rejected. In addition, the corresponding RF confusion matrix in Table IV shows that the dominant residual errors occur between U and Blank, while the remaining classes are largely preserved, confirming that the rejection boundary (decision boundary that separates "Blank" (reject) from "Target" (accept as a letter) is the most sensitive failure mode under imperfect optical feedback.

These results also validate the methodological choices made in the proposed pipeline. First, the compact f_6 descriptor is appropriate for regime analysis because it captures both amplitude driven statistics (linked to feedback level) and derivative driven dynamics (linked to fast waveform variations), yielding a feature space where regime inconsistent samples separate as sparse regions that DBSCAN can reject. Second, we report macro-F1 because it weights all classes equally and therefore reflects rejection reliability in the presence of class specific difficulty, especially the Blank class. Third, confusion analyzes are necessary because a single scalar score cannot reveal whether errors arise from target to target confusions or from failures of the Blank reject decision, which is the critical behavior in sensing applications.

We proposed a density core regime selection method for experimental SM signals based on standardized f_6 descriptors, t-SNE embedding, and DBSCAN filtering. A four operating point hyperparameter study (A–D) shows that filtering selectivity changes retention and class confusability in a six class UBARI+Blank task. Conservative filtering (Dataset A) yields the best RF performance; for this Dataset, regime consistent cores improve robustness under imperfect experimental conditions. Table I quantifies how the DBSCAN operating point controls the retained fraction ρ: moving from A to D increases retention but also admits progressively less consistent measurements, which is reflected by the decrease in macro-F1 for more tolerant settings (Table II). Overall, the results support using $(\varepsilon, \mathrm{minPts})$ as a practical knob to tune the stability coverage tradeoff, with Blank related confusions providing the most sensitive indicator of regime quality in this dataset.

ACKNOWLEDGMENT

This study was carried out within the MIRABILIS project (CUPM. D53D23002780006), funded by European Union Next Generation EU within the PRIN 2022 program. This manuscript reflects only the authors views and opinions and the Ministry cannot be considered responsible for them.

REFERENCES

[1] D. M. Kane and K. A. Shore, *Unlocking Dynamical Diversity: Optical Feedback Effects on Semiconductor Lasers*. Hoboken, NJ: Wiley-IEEE Press, 2005.

[2] T. Fukui, Y. Kohno, R. Tang, Y. Nakano, and T. Tanemura, "Single-pixel imaging using multimode fiber and silicon photonic phased array," *J. Lightwave Technol.*, vol. 39, no. 3, pp. 839–844, Feb 2021. [Online]. Available: https://opg.optica.org/jlt/abstract.cfm?URI=jlt-39-3-839

[3] B. Ersöz, P. Chaudhary, P. Bardella, M. Dabbicco, L. Columbo, and M. Brambilla, "Inaccuracy amplified: Compressed sensing under experimental misalignment," in *2025 International Conference on Numerical Simulation of Optoelectronic Devices (NUSOD)*, 2025, pp. 51–52.

[4] S. Gare, S. Chel, M. Kuruba, S. Jana, and L. Giri, "Dimension reduction and clustering of single cell calcium spiking: Comparison of t-sne and umap," in *2021 National Conference on Communications (NCC)*, 2021, pp. 1–6.

[5] D. Deng, "Dbscan clustering algorithm based on density," in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, 2020, pp. 949–953.