STOOD-X: Detección Explicable de Muestras Fuera de Distribución*

Iván Sevillano-García[1], Julián Luengo[1], and Francisco Herrera[1]

Departamento de Ciencias de la Computación e Inteligencia Artificial, Instituto Andaluz de Investigación en Ciencia de Datos e Inteligencia Computacional (DaSCI), Universidad de Granada, Granada, España

Abstract. En este trabajo presentamos STOOD-X, una metodología de dos etapas que combina una prueba estadística no paramétrica para la detección de muestras fuera de distribución (OOD) con explicabilidad centrada en el usuario. La primera etapa utiliza distancias en el espacio de características y la prueba de Wilcoxon-Mann-Whitney para identificar muestras OOD. La segunda etapa genera explicaciones visuales conceptuales mediante vecinos más cercanos y Concept-based Relevance Propagation (CRP), alineadas con el paradigma BLUE XAI. Los experimentos en ImageNet con ViT-B/16 demuestran rendimiento competitivo (81.95% AUROC cercano, 92.20% lejano).

Keywords: Inteligencia Artificial Explicable · Aprendizaje Profundo · Fuera de Distribución

1 Introducción

La detección de muestras fuera de distribución (OOD, por sus siglas en inglés) se ha consolidado como un desafío fundamental en el aprendizaje automático [1], consistente en diferenciar entre muestras dentro de distribución (ID) y fuera de distribución. En aplicaciones críticas para la seguridad, la ausencia de algoritmos OOD puede llevar a modelos de inteligencia artificial (IA) a tomar decisiones incorrectas en lugar de diferir al juicio humano.

Los algoritmos OOD se categorizan en basados en entrenamiento y post hoc. Los primeros modifican el entrenamiento para aumentar la separabilidad entre muestras ID y OOD [2,3], mientras que los post hoc se aplican a modelos ya entrenados, utilizando clasificación, gradientes o distancias en el espacio de características. Mientras algunos métodos paramétricos asumen normalidad [4], otros como KNN [5] no requieren tales suposiciones. Recientes trabajos también han explorado combinaciones de enfoques paramétricos y no paramétricos [6].

A pesar de estos avances, muchos métodos OOD carecen de explicabilidad, dificultando comprender cómo y por qué una muestra se clasifica como OOD.

* Este trabajo resume la publicación original: Sevillano-García, I., Luengo, J., Herrera, F.: STOOD-X: Explainable out-of-distribution detection via nonparametric statistical testing on large-scale datasets. *Pattern Recognition*, 113254 (2026). https://doi.org/10.1016/j.patcog.2026.113254

Para abordar estos desafíos, introducimos STOOD-X (Statistical Test for OOD detection enhanced with eXplainability), una metodología de dos etapas que combina detección robusta con explicaciones comprensibles. El diagrama de flujo completo de la metodología se encuentra en el trabajo original [7].

2 Metodología STOOD-X

STOOD-X es una metodología de dos etapas que utiliza análisis de espacio de características y pruebas estadísticas para la detección OOD, complementada con explicaciones conceptuales alineadas con el paradigma BLUE XAI. La metodología ejecuta dos etapas secuenciales e interdependientes:

1. En la primera etapa, el modelo original procesa las características y las mapea al espacio de características aprendido. Dentro de este espacio, se analizan las distancias a muestras previamente observadas y se realiza una prueba de hipótesis estadística para determinar si la nueva muestra está suficientemente cerca de las muestras ID validadas.
2. La segunda etapa proporciona información adicional sobre el proceso de toma de decisiones presentando muestras cercanas junto con las características más importantes compartidas entre la nueva muestra y sus contrapartes más cercanas.

La detección se basa en un análisis no paramétrico de las distancias en el espacio de características mediante la prueba de Wilcoxon-Mann-Whitney, que compara las distancias de la muestra a sus vecinos más cercanos con las distancias que estos vecinos tienen a los vecinos de su misma clase. La formulación matemática detallada se encuentra en el trabajo original [7].

Para justificar las decisiones del método, la segunda etapa se basa en los siguientes pasos:

1. **Búsqueda de vecinos más cercanos (NNs):** Seleccionamos las muestras de entrenamiento con características más cercanas en el espacio de características. Dado que la puntuación de STOOD-X se basa en distancias a muestras de entrenamiento, las muestras más cercanas deben tener visualizaciones XAI similares si son clasificadas como ID.
2. **Marco de Visualización de Características:** Basado en Concept-based Relevance Propagation (CRP) [8], las visualizaciones muestran la localización de características más importantes en las muestras de entrenamiento y la nueva muestra, identificando conceptos compartidos.

Este enfoque está alineado con BLUE XAI, permitiendo la validación por parte del usuario de si las características aprendidas por el modelo son consistentes con muestras validadas.

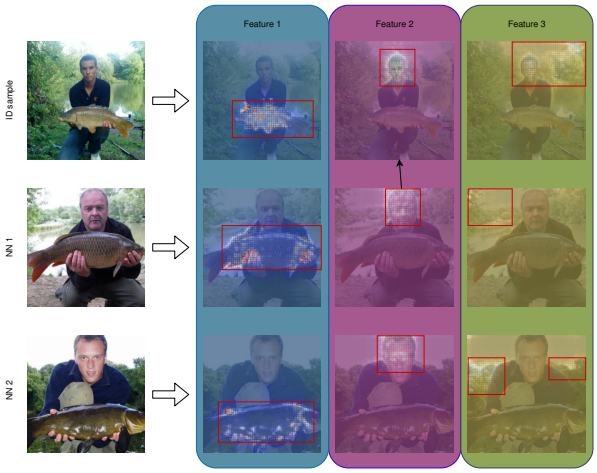


Fig. 1. Visualización explicativa de una muestra ID (tinca, puntuación 52%). Se muestran los vecinos más cercanos y la localización de las características más importantes compartidas.

3 Resultados Experimentales

La evaluación de STOOD-X se realizó utilizando OpenOOD [9] sobre sus conjuntos predefinidos. La Tabla 1 muestra la comparativa en ImageNet con ViT-B/16 de los modelos del estado del arte de ese conjunto.

	RMDS	CombOOD	MDS	VIM	**STOOD-X**
Cercano	80.09	79.60	79.04	77.03	**81.95**
Lejano	92.60	92.65	92.60	**92.84**	92.20

Table 1. AUROC (%) en ImageNet con ViT-B/16 para los mejores métodos en escenario cercano.

STOOD-X alcanza 81.95% AUROC en escenarios cercanos y 92.20% en lejanos sobre ImageNet con ViT-B/16, situándose como estado del arte en la configuración cercana. En la muestra ID, el 76% de la decisión se explica con solo 3 características frente al 12% en la muestra OOD, confirmando que la explicabilidad actúa como indicador de confianza adicional.

4 Conclusiones

STOOD-X combina análisis estadístico no paramétrico con explicabilidad centrada en el usuario: detección OOD mediante test estadístico y visualizaciones

conceptuales alineadas con BLUE XAI. Los resultados en ImageNet con ViT-B/16 muestran rendimiento competitivo (81.95% cercano, 92.20% lejano), incluso situandose como estado del arte en ciertas configuraciones, y donde la explicabilidad actúa como indicador de confianza (76% vs 12% de decisión explicada con 3 características en ID vs OOD).

Agradecimientos. Esta investigación es parte del Proyecto TSI-100927-2023-1, financiado por Plan de Recuperación, Transformación y Resiliencia Europeo, de la Unión Europea-Next Generation EU a través del Ministerio para la Transformación Digital y de la Función Pública y está parcialmente financiada por el Ministerio de Ciencia y Tecnología de España en el marco del proyecto PID2023-150070NB-I00, financiado por MICIU/AEI/10.13039/501100011033 y por FEDER, UE.

References

1. Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., Cui, P.: Towards out-of-distribution generalization: A survey. arXiv preprint arXiv:2108.13624 (2021)
2. Tack, J., Mo, S., Jeong, J., Shin, J.: Csi: Novelty detection via contrastive learning on distributionally shifted instances. Advances in neural information processing systems **33**, 11839–11852 (2020)
3. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. arXiv preprint arXiv:1802.04865 (2018)
4. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in Neural Information Processing Systems **31** (2018)
5. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 20827–20840. Proceedings of Machine Learning Research (17–23 Jul 2022)
6. Rajasekaran, M., Sajol, M.S.I., Berglind, F., Mukhopadhyay, S., Das, K.: Combood: A semiparametric approach for detecting out-of-distribution data for image classification. In: Proceedings of the 2024 Society for Industrial and Applied Mathematics(SIAM) International Conference on Data Mining (SDM). pp. 643–651. Society for Industrial and Applied Mathematics (2024)
7. Sevillano-García, I., Luengo, J., Herrera, F.: Stood-x: Explainable out-of-distribution detection via nonparametric statistical testing on large-scale datasets. Pattern Recognition **177**, 113254 (2026). https://doi.org/https://doi.org/10.1016/j.patcog.2026.113254, https://www.sciencedirect.com/science/article/pii/S0031320326002190
8. Achtibat, R., Dreyer, M., Eisenbraun, I., Bosse, S., Wiegand, T., Samek, W., Lapuschkin, S.: From attribution maps to human-understandable explanations through concept relevance propagation. Nature Machine Intelligence **5**(9), 1006–1019 (2023). https://doi.org/10.1038/s42256-023-00711-8
9. Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Li, Y., Liu, Z., Chen, Y., Li, H.: Openood v1.5: Enhanced benchmark for out-of-distribution detection. Journal of Data-centric Machine Learning Research (2024), https://openreview.net/forum?id=cnnTnJQigs, dataset Certification