

# A Multicenter Whole Slide Image Dataset for Classification and Out-of-Distribution Detection in Spindle Cell Cutaneous Neoplasms

Francisco Javier Sáez-Maldonado<sup>1,\*</sup>, Javier Mateos<sup>1</sup>, Pablo Meseguer<sup>4</sup>, Alejandro Golfe<sup>4</sup>, Miguel López-Pérez<sup>1</sup>, Sandra Morales<sup>4</sup>, Rocío del Amor<sup>4</sup>, Liria Terradez<sup>5</sup>, Aurelio Martín-Castro<sup>3</sup>, Juan Gómez-Valcárcel<sup>3</sup>, Paloma Talavera<sup>3</sup>, Pablo Morales-Álvarez<sup>2,7</sup>, Rafael Molina<sup>1</sup>, and Valery Naranjo<sup>4,6</sup>

<sup>1</sup>Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain.

<sup>2</sup>Department of Statistics and Operations Research, University of Granada, Granada, Spain.

<sup>3</sup>Provincial Inter-center Unit of Pathological Anatomy of Granada, Granada, Spain

<sup>4</sup>Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, HUMAN-tech, Universitat Politècnica de València, Valencia, Spain

<sup>5</sup>Pathology Department. Hospital Clínico Universitario de Valencia, Universidad de Valencia, Spain.

<sup>6</sup>Artikode Intelligence S.L, Valencia, Spain

<sup>7</sup>Institute of Mathematics(IMAG), University of Granada, Spain

\*Corresponding author. Email: [fjaviersaezm@ugr.es](mailto:fjaviersaezm@ugr.es)

## ABSTRACT

Cutaneous spindle cell (CSC) neoplasms are a notoriously challenging diagnostic group within the spectrum of skin malignancies. While specialized CSC neoplasm datasets like AI4SKIN have enabled the development of AI whole-slide image (WSI) classification methods, these methods are largely constrained by closed-set designs. However, in real-world clinical practice, biopsies often contain secondary tumors or rare entities unknown to the developed AI classifiers. When faced with such Out-of-Distribution (OoD) cases, automated systems can produce highly confident but incorrect predictions, posing a significant risk to patient safety and hindering the reliable deployment of AI in routine pathology. To address this limitation, we present ASSIST, a multicenter dataset of 410 WSIs comprising both spindle-cell tumors and a diverse set of metastatic and rare lesions explicitly included as OoD samples. ASSIST expands AI4SKIN by increasing sample diversity, reinforcing underrepresented categories, and introducing clinically meaningful OoD cases, thereby enabling the development of more robust and safety-oriented computational pathology models. We describe the dataset acquisition pipeline, annotation structure, and technical validation using multiple instance learning (MIL) models combined with a suite of OoD detection methods. ASSIST provides an essential benchmark for advancing open-world pathology and supports future research in dependable AI systems for dermatopathology.

## Background & Summary

Cutaneous spindle cell (CSC) lesions comprise a heterogeneous group of tumors that includes both benign and malignant processes with partially overlapping histological patterns. Among the most frequent spindle cell neoplasms are leiomyomas (lm), leiomyosarcomas (lms), dermatofibromas (df), dermatofibrosarcomas (dfs), spindle cell melanomas (scm), atypical fibroxanthomas (afx), and squamous cell carcinomas (sc). Morphological similarity among these entities can complicate their diagnosis; therefore, histopathological evaluation often requires the integration of morphological findings, immunohistochemical studies, molecular alterations and clinical data. These seven neoplastic categories were included in the previous AI4SKIN<sup>1,2</sup> and CR-AI4SKIN<sup>3</sup> studies to ensure a sufficiently meaningful number of cases per diagnosis for the identification of shared and distinctive morphological features. Other cutaneous spindle cell tumors recognized by the WHO of vascular, neural, melanocytic and fibroblastic lineages, as well as entities secondary to rearrangements, were excluded due to their lower frequency.

In everyday diagnostic practice, pathologists also face cases that do not correspond to common tumor entities. In such situations, it is necessary to consider alternative diagnoses such as cutaneous metastases, that is, secondary tumors whose primary origin lies elsewhere, as well as rare tumors or poorly differentiated neoplasms. Typically, primary cutaneous neoplasms are connected to the epidermis and/or with the epithelium structures of cutaneous adnexa from which they originate. These

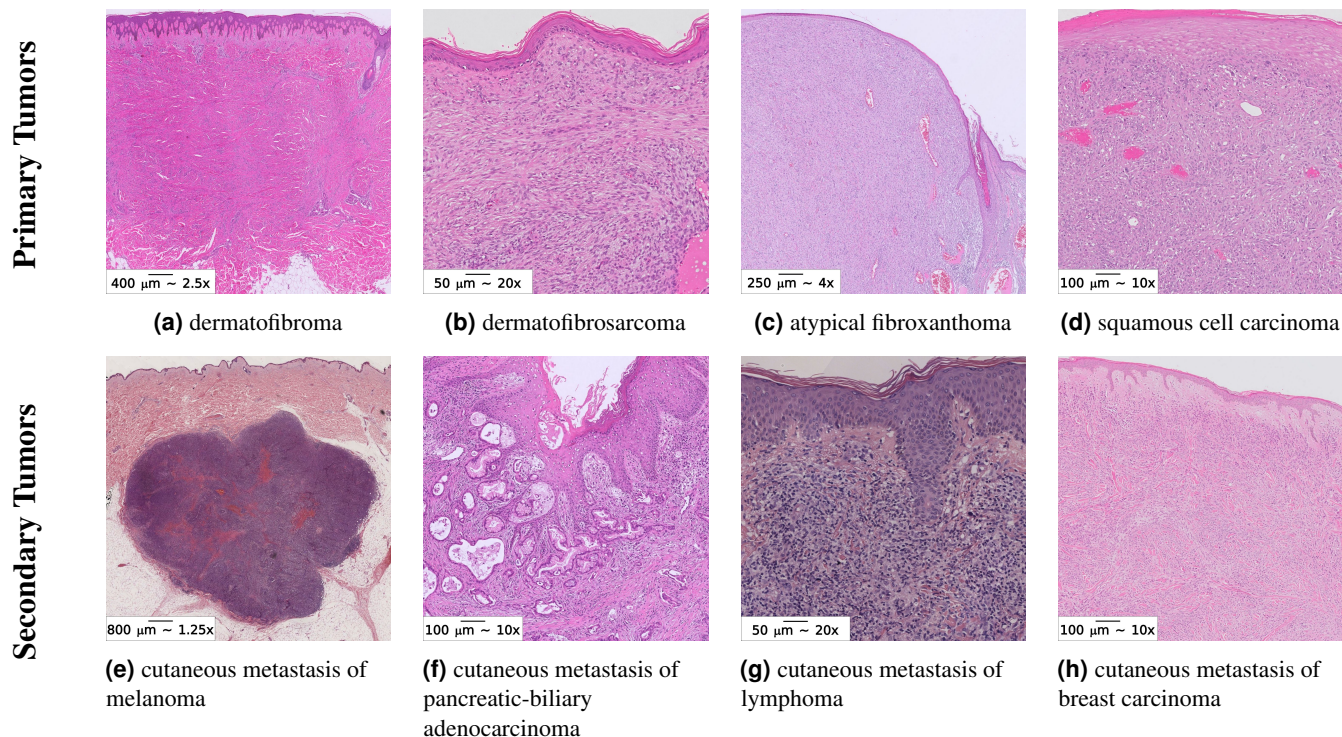


**Figure 1.** Representative areas of some common primary and secondary (OoD) neoplasms.

structures may exhibit early or in situ tumor changes, or show reactive alterations induced by the tumor. In addition, primary neoplasms tend to occupy the superficial or mid dermis in an organized manner. Secondary (metastatic) neoplasms, by contrast, usually present with a nodular or diffuse pattern in relatively intact skin, sparing the epidermis and involving deeper regions, including the hypodermis. Recognizing these cases is essential, as an incorrect diagnosis may lead to inappropriate therapeutic decisions and affect patient outcomes. Therefore, beyond recognizing known entities, the diagnostic process also requires identifying cases that do not correspond to any expected diagnosis. Figure 1 shows examples of primary neoplasms from various categories (dermatofibroma, dermatofibrosarcoma protuberans, atypical fibroxanthoma, and spindle cell squamous cell carcinoma) as well as examples of metastatic neoplasms originating from melanoma, pancreatic-biliary and breast carcinomas, and lymphoma.

The evolution of digital pathology and whole-slide imaging (WSI) has driven the progress of computer-aided diagnosis (CAD) systems in the high-resolution analysis of histopathological images. Several studies have demonstrated the potential of these approaches for automatic classification of tumors from WSIs<sup>4-7</sup>. However, most models are trained under a closed-set classification paradigm, where every WSI is forced into a predefined diagnostic category<sup>8</sup>. While this approach simplifies the computational task, it fails to account for the inherent complexity of real-world clinical practice.

As a pathologist in clinical practice, if a CAD system is expected to be deployed in a hospital, it needs to recognize entities that were not encountered during the training procedure and for which what has been learned so far does not hold<sup>9</sup>. In the machine learning community, such cases are referred to as Out-of-Distribution (OoD) samples, as their characteristics differ from the data distribution learned during model training. The inability to detect these cases may lead to inaccurate predictions made with high confidence, representing a significant risk for the clinical implementation of CAD systems. Therefore, detecting OoD samples is a key requirement for the development of robust and clinically safe diagnostic support systems.

In the context of CSC neoplasms, OoD cases may correspond to metastatic tumors, rare entities, or neoplasms not represented in the training datasets. Public datasets with scalable annotation strategies play a crucial role in enabling robust evaluation. The CR-AI4SKIN<sup>3</sup> and AI4SKIN<sup>1,2</sup> datasets, for example, provide multicenter WSIs of spindle-cell lesions along with expert and crowdsourced labels, supporting studies on label noise, inter-hospital variability, and weak supervision. However, they are designed using the closed-set assumption and, hence, they do not take into account any OoD cases.

In this work, we address the limitations of existing datasets, notably their limitation in accounting for OoD cases, by introducing the ASSIST dataset. This is a new multicenter dataset specifically designed to complement and extend AI4SKIN. The WSIs for ASSIST were obtained from two centers: the Provincial Inter-center Unit of Pathological Anatomy of Granada (Unidad Provincial Intercentros de Anatomía Patológica de la provincia de Granada, UPIGAP) and the Hospital University

Clinic of Valencia (Hospital Clínico Universitario de Valencia, HCUV). ASSIST not only increases the number of WSIs for underrepresented spindle-cell tumor categories but also incorporates a wide spectrum of OoD cases without further distinction, including common metastatic tumors (of pulmonary, breast, or colorectal origin, among others) and other skin neoplasms that pathologists can encounter in real clinical practice. By combining AI4SKIN and ASSIST, we create a substantially enriched dataset that reflects the heterogeneity and diagnostic challenges of spindle-cell cutaneous lesions, allowing for a robust evaluation of both classification and OoD detection methods. This paper presents the ASSIST dataset, details its construction and annotation pipeline, and provides a comprehensive technical validation using state-of-the-art foundation models, multiple instance learning classification methods, and OoD detection approaches.

## Methods

### Study approval

The Provincial Research Ethics Committee of Granada and the Ethic Committee of the HCUV approved studies no. SICEIA-2024-001086 and no. 2024/011, respectively, conducted within the framework of the ASSIST project. This project was funded by the Spanish Ministry of Science and Innovation under grant numbers PID2022-140189OB-C21 and PID2022-140189OB-C22. All research procedures adhered to the ethical principles outlined in the Declaration of Helsinki. Furthermore, the Research Ethics Committee of the Universitat Politècnica de València (UPV), acting as project coordinator, issued a favorable ethics report for the ASSIST project (reference P07\_24-07-2024).

The authors guarantee that all data included in this study was de-identified and processed in compliance with strict security and confidentiality standards to protect participant rights and welfare. All images were obtained from the hospitals archives and anonymized histological images, dissociated from clinical data, were always used. The requirement for informed consent was formally waived by the corresponding ethics committees for the retrospective use of pseudonymised archival material.

### Input Data

ASSIST dataset extends the AI4SKIN<sup>2</sup> image dataset by including more images of some classes, especially underrepresented categories, and, most importantly, images of secondary tumors. In order to import AI4SKIN dataset, the “WSI.7z” file, containing the digitalized slides grouped in folders according to the subtype, should be downloaded from Figshare (<https://figshare.com/s/74a12f4b5c6473496b82>) and decompressed in the same folder that ASSIST dataset (see data records section). The class of each WSI is available in the column GT of spreadsheet “AI4SkIN\_Database\_v2.xlsx” in Figshare. Additionally, the ASSIST dataset also includes this information in spreadsheet “ASSIST\_Full\_Dataset.xlsx”. Note that AI4SKIN used HUSC as center name for the Hospital Universitario San Cecilio of Granada, which is part of UPIGAP.

### ASSIST Slide Acquisition

UPIGAP identified the cases included in this study from the database of the pathology information system, Anapath, which is used in the Department of Pathology for the management of histopathological studies. Case selection was performed using the “Scientific Analysis” tool integrated into the system, which allows structured searches based on different database fields. To identify potentially relevant cases, specific search criteria for skin metastasis were applied. Recent cases corresponding to the neoplasms represented in the AI4SKIN dataset were also identified. Note that these slides were digitized after the release of AI4SKIN and, therefore, were not included in the dataset. The addition of a larger and more comprehensive number of categories, considering rare diagnoses as well as emerging entities, is envisioned as a future and challenging endeavor. The cases retrieved using this strategy constituted the initial set of images to be included in the analysis. HCUV used an in-house software tool, PATWin, to identify the cases to be considered in this study. Among other functionalities, the tool enables the search of cases according to the patient diagnosis or data availability. The set of candidates was obtained by selecting the cases with a diagnosis of a skin metastasis with at least one biopsy sample available. Given the diagnostic complexity of cutaneous spindle-cell neoplasms, case annotation was established through expert dermatopathology review rather than direct extraction from pathology records. All cases were evaluated on hematoxylin-eosin sections, and diagnoses with unequivocal morphology in the appropriate clinical context were accepted on morphological grounds. Cases with overlapping morphology or limited specificity underwent a stepwise diagnostic work-up including targeted immunohistochemistry and, in selected cases, molecular studies. Immunohistochemical panels were tailored to the differential diagnosis and typically included markers such as CD34, factor XIIIa, S100, SOX10, Melan-A, HMB45, SMA, desmin, h-caldesmon, pancytokeratins, p63/p40, CD31, ERG, STAT6, and H3K27me3, as appropriate. Molecular testing was reserved for unresolved or entity-specific cases, including COL1A1::PDGFB fusion testing when indicated. For dataset curation, diagnoses were reviewed by two experienced dermatopathologists from two different hospitals. Difficult or borderline cases were jointly reassessed using morphology together with ancillary and clinical information, and consensus was required for inclusion. Cases that remained insufficiently resolved were excluded.

At UPIGAP, slides were digitized using the Philips Ultra Fast Scanner and the Philips Pathology Scanner SG300, which acquire images at 40× magnification, yielding a spatial resolution of 0.25 µm per pixel. The whole-slide images were saved in TIFF format via the Philips Image Management System (IMS). At HCUV, digitization was performed using the Ventana iScan HT scanner (Roche), equipped with a 40× objective (0.227 µm/pixel), with WSIs exported directly in TIFF format. Thus, in both centers, the slides were consistently digitized at a maximum magnification of 40× although they retain scanner native spatial resolutions. No digital resampling or interpolation to a uniform microns-per-pixel (MPP) value was applied, and all WSIs were preserved at their original resolution. Also, no color profile or stain normalization was applied during preprocessing to preserve inter-scanner variability.

Quality assurance through visual inspection by trained personnel ensured that all WSI met minimum thresholds for focus sharpness, color stability, and absence of scanning artifacts, thereby maintaining uniformity across the whole dataset. During dataset curation, additional checks were carried out, and slides with insufficient quality were excluded.

To ensure complete anonymization and consistency among sources, all image files were de-identified at both hospitals following strict institutional protocols and embedded metadata was removed. Then, we standardize the dataset to provide uniform file characteristics by recompressing each image into TIFF format, using JPEG compression with a quality factor of 80, while preserving their pyramidal structure, ensuring compatibility with open-source libraries such as OpenSlide.

**Table 1.** ASSIST dataset distribution. CSC neoplasms contained in the dataset (OoD: secondary lesions; df: dermatofibromas; dfs: dermatofibrosarcomas; afx: atypical fibroxanthomas; scc: squamous cell carcinoma.)

|        | OoD | df | dfs | afx | scc | Total | Unique patients |
|--------|-----|----|-----|-----|-----|-------|-----------------|
| HCUV   | 238 | –  | –   | –   | –   | 238   | 224             |
| UPIGAP | 77  | 31 | 24  | 24  | 16  | 172   | 172             |
| Total  | 315 | 31 | 24  | 24  | 16  | 410   | 396             |

**Table 2.** ASSIST+AI4SKin dataset distribution. CSC neoplasms contained in the dataset (OoD: secondary lesions; lm: leiomyomas; lms: leiomyosarcomas; df: dermatofibromas; dfs: dermatofibrosarcomas; scm: spindle cell melanomas; afx: atypical fibroxanthomas; scc: squamous cell carcinoma.)

|             | OoD | In Distribution |     |     |     |     |     |     | Total |
|-------------|-----|-----------------|-----|-----|-----|-----|-----|-----|-------|
|             |     | lm              | lms | df  | dfs | scm | afx | scc |       |
| HCUV        | 238 | 31              | 23  | 102 | 21  | 48  | 44  | 15  | 522   |
| UPIGAP      | 77  | 73              | 23  | 124 | 60  | 74  | 82  | 16  | 519   |
| Total/class | 315 | 104             | 46  | 226 | 81  | 122 | 126 | 31  |       |
| Total       | 315 | 736             |     |     |     |     |     |     | 1051  |

Each case was associated with its corresponding histopathological diagnosis according to the final pathologist decision. The ASSIST dataset includes a total of 410 WSIs from 396 different patients (note that a patient may have multiple images of lesions in different locations, but they always originate from the same primary neoplasm) including different types of spindle cell cutaneous neoplasms as well as 315 cases of metastatic tumors. The dataset distribution is presented in Table 1. Additional cohort-level statistics about the OoD cases, including the anatomical site and morphology information (when available), is presented in Table 3. When integrating the ASSIST and AI4SKIN datasets, we obtain a substantially enriched and comprehensive resource for the study of spindle cell cutaneous neoplasms. The combined dataset, which distribution is presented in Table 2, increases the diversity and representativeness of clinically relevant entities in the seven targeted tumor categories and provides support for robust out-of-distribution (OOD) detection.

## Data Records

The complete ASSIST<sup>10</sup> dataset is publicly available via BioImage Archive at (the temporary URL) <https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BIAD3206?key=5d1829c4-d18c-49fd-9b47-c0dd758c94fd>. The dataset comprises all the digitized WSIs organized into folders according to their corresponding histopathological subtype and two spreadsheets: one named “ASSIST\_Dataset.xlsx”, with the histopathological diagnosis (class) for each image in the

**Table 3.** Cohort level statistics of the OoD cases (secondary tumors) presented in ASSIST. (a) Distribution of cases by anatomical site. (b) Distribution of cases by morphology.

| Anatomical Site            | HCUV | UPIGAP | Total | Morphology                 | HCUV | UPIGAP | Total |
|----------------------------|------|--------|-------|----------------------------|------|--------|-------|
| Breast, NOS                | 108  | 63     | 171   | Adenocarcinoma             | 99   | 2      | 101   |
| Female reproductive system | 1    | 2      | 3     | Carcinoma                  | 37   | 65     | 102   |
| Gastrointestinal           | 19   | 0      | 19    | Melanoma                   | 38   | 7      | 45    |
| Hemato. and Ret. systems   | 2    | 0      | 2     | Neuroendocrine carcinoma   | 0    | 1      | 1     |
| Lymphoma                   | 25   | 0      | 25    | Pleomorphic leiomyosarcoma | 0    | 2      | 2     |
| Oral cavity                | 0    | 1      | 1     | Unknown                    | 64   | 0      | 64    |
| Pancreas, NOS              | 3    | 0      | 3     |                            |      |        |       |
| Pancreatic-biliary         | 0    | 2      | 2     |                            |      |        |       |
| Pulmonary                  | 23   | 1      | 24    |                            |      |        |       |
| Renal                      | 5    | 1      | 6     |                            |      |        |       |
| Skin, NOS                  | 52   | 7      | 59    |                            |      |        |       |

(a) Distribution of OoD cases by anatomical site.

(b) Distribution of OoD cases by morphology.

dataset, and another named “ASSIST\_Full\_Dataset.xlsx”, including the class of the images of both AI4SKIN and ASSIST datasets. For each sample, we include the class, the name, and the center from which the image is acquired. For OoD samples, the anatomical site and morphology of the lesion, if available, are included. For those cases where the anatomical site has been identified but does not possess specific features that allow it to be classified into a more precise, narrow subtype, NOS (no otherwise specified) is added to the anatomical site.

### Image naming convention

Following AI4SKIN, TIF files in ASSIST dataset follow the naming scheme `ASSIST_HOSPITAL_NUMBER_SAMPLE.tif`, where

- `HOSPITAL` is either `UPIGAP` or `HCUV`, depending on the center where the image was digitized,
- `NUMBER` is a unique random identifier assigned to the WSI.
- `SAMPLE` is A for the initial biopsy. If an additional biopsies are available for the same patient, they will be labeled as B, C, and so on.

## Technical Validation

To validate the dataset presented in this work, we assess its relevance for computational pathology tasks by training Multiple Instance Learning (MIL) classifiers and quantifying their ability to differentiate the different spindle cell neoplasms. Then, we evaluate out-of-distribution detection task by computing several OoD scores to distinguish the spindle cell lesions from the OoD cases included in the dataset. These experiments establish a benchmark for CSC subtype classification and OoD detection on the ASSIST dataset. The complete methodology, illustrated in the proposed framework in Figure 2, includes WSI pre-processing, feature extraction with foundation models, and MIL model training for classification and OoD detection tasks.

### Data pre-processing and feature extraction

We use the software tools in the TRIDENT repository<sup>11</sup> to preprocess the WSIs. The tissue in each image, detected using HEST<sup>12</sup>, is divided into  $512 \times 512$  non-overlapping patches at a  $20\times$  magnification level. Then, we utilize five different histopathological foundation models, which are trained to produce semantically-rich patch embeddings, to obtain a representation of each patch. More precisely, we use CONCHv1.5<sup>13</sup>, VIRCHOW<sup>14</sup>, UNI and UNI2<sup>15</sup>, and PHIKON<sup>16</sup>. The patches were resized using bilinear interpolation to match the input dimensions required by each foundation model, enabling extraction of patch embeddings. All the patch embeddings in a bag are used as input for the MIL models.

### WSI Classification and OOD Detection

To classify the slides, we use MIL methods, which consider slides as bags of instances that, in this case, are the patches of the image. MIL models aggregate the information from the patches in each slide to create a single bag representation. Depending on how this bag representation is obtained, MIL models are divided into two categories: those treating instances independently

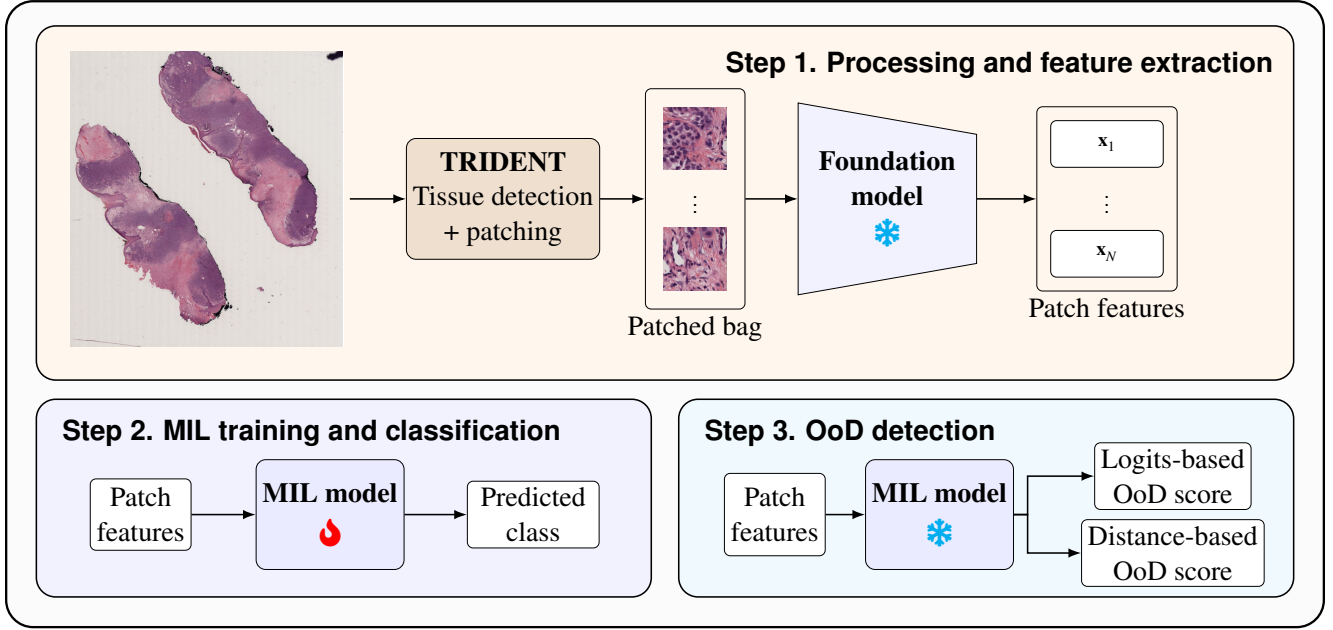


**Figure 2.** Technical validation pipeline for the ASSIST dataset. The workflow is organized into three separate stages: processing and feature extraction, MIL training and classification, and OoD detection.

and those considering inter-instance relationships. For approaches that treat instances independently, we employ ABMIL<sup>17</sup> and VAEABMIL<sup>18</sup>. For models that account for inter-instance dependencies, we use SMABMIL<sup>19</sup> and TRANSMIL<sup>20</sup>. Finally, the resulting bag representations are mapped to diagnostic labels using a linear classifier.

To perform the OoD detection task, we leverage the trained MIL models to compute OoD scores. These scores rely either on the model’s final output (classifier-based) or on an intermediate feature representation (distance-based) to derive a proxy measure of how likely a new test sample is originated from the same data distribution as the training set. In this work, we use the model’s output to compute the following scores: ENERGY<sup>21</sup>, ENTROPY<sup>22</sup>, MAXLOGIT<sup>23</sup>, ODIN<sup>24</sup>, and REACT<sup>25</sup>. Additionally, we use the bag representation produced by the MIL models to compute the KNN<sup>26</sup> and MAHALANOBIS<sup>27</sup> scores.

|       |        | OoD | lm | lms | df | dfs | scm | afx | scc |
|-------|--------|-----|----|-----|----|-----|-----|-----|-----|
| Train | HCUV   | -   | 23 | 18  | 87 | 18  | 36  | 36  | 11  |
|       | UPIGAP | -   | 59 | 18  | 98 | 48  | 60  | 64  | 13  |
| Test  | HCUV   | -   | 8  | 5   | 15 | 3   | 12  | 8   | 4   |
|       | UPIGAP | -   | 14 | 5   | 26 | 12  | 14  | 18  | 3   |
| OoD   | HCUV   | 238 | -  | -   | -  | -   | -   | -   | -   |
|       | UPIGAP | 77  | -  | -   | -  | -   | -   | -   | -   |

**Table 4.** Number of WSIs per class and hospital in the train, test, and OoD subsets.

We divide the ASSIST+AI4SKIN dataset into training and testing sets (see Table 4), using 80% and 20% of the in-distribution data, respectively, while reserving the OoD samples entirely for evaluation. We ensure that the class distribution is stratified and that WSIs from the same patient are kept within the same set.

To train the MIL models, we split the training set into five folds of equal size and perform five runs, each using four folds for training and one fold for validation. All MIL models are implemented the TORCHMIL library<sup>28</sup>. For each foundation model, the slide representation dimensionality matches that of the corresponding extracted patch embeddings (see the exact dimensions in the TRIDENT repository)<sup>11</sup>. We train the models to optimize the binary cross entropy loss for up to 20 epochs using a batch size of 1 and the Adam optimizer with a learning rate of  $10^{-4}$ . During training, the validation set is used to apply early stopping if the validation loss does not improve for 5 consecutive epochs. Finally, to compute the test metrics, we evaluate the testing set using the model weights that achieved the lowest validation loss during training.

| Feat. Ext. | Model                  | Accuracy                            | AUROC                               | F1-Score                            |
|------------|------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
| CONCH      | ABMIL <sup>17</sup>    | $0.853 \pm 0.034$                   | $0.972 \pm 0.005$                   | $0.743 \pm 0.094$                   |
|            | VAEABMIL <sup>18</sup> | $0.854 \pm 0.010$                   | $0.969 \pm 0.006$                   | <b><math>0.804 \pm 0.011</math></b> |
|            | TRANSMIL <sup>20</sup> | <b><math>0.865 \pm 0.030</math></b> | <b><math>0.975 \pm 0.006</math></b> | $0.787 \pm 0.089$                   |
|            | SMABMIL <sup>19</sup>  | $0.842 \pm 0.029$                   | $0.965 \pm 0.009$                   | $0.723 \pm 0.077$                   |
| UNI        | ABMIL <sup>17</sup>    | $0.846 \pm 0.018$                   | $0.962 \pm 0.010$                   | $0.723 \pm 0.045$                   |
|            | VAEABMIL <sup>18</sup> | $0.853 \pm 0.014$                   | <b><math>0.976 \pm 0.003</math></b> | $0.755 \pm 0.031$                   |
|            | TRANSMIL <sup>20</sup> | <b><math>0.869 \pm 0.006</math></b> | $0.963 \pm 0.010$                   | <b><math>0.776 \pm 0.014</math></b> |
|            | SMABMIL <sup>19</sup>  | $0.850 \pm 0.008$                   | $0.965 \pm 0.009$                   | $0.734 \pm 0.029$                   |
| UNI2       | ABMIL <sup>17</sup>    | $0.850 \pm 0.010$                   | <b><math>0.970 \pm 0.005</math></b> | $0.713 \pm 0.011$                   |
|            | VAEABMIL <sup>18</sup> | <b><math>0.865 \pm 0.007</math></b> | $0.957 \pm 0.004$                   | <b><math>0.795 \pm 0.025</math></b> |
|            | TRANSMIL <sup>20</sup> | $0.850 \pm 0.019$                   | $0.964 \pm 0.011$                   | $0.742 \pm 0.031$                   |
|            | SMABMIL <sup>19</sup>  | $0.860 \pm 0.011$                   | $0.965 \pm 0.007$                   | $0.744 \pm 0.036$                   |
| VIRCHOW    | ABMIL <sup>17</sup>    | $0.845 \pm 0.019$                   | $0.955 \pm 0.013$                   | $0.717 \pm 0.035$                   |
|            | VAEABMIL <sup>18</sup> | <b><math>0.849 \pm 0.010</math></b> | <b><math>0.969 \pm 0.003</math></b> | <b><math>0.744 \pm 0.035</math></b> |
|            | TRANSMIL <sup>20</sup> | $0.839 \pm 0.016$                   | $0.956 \pm 0.014$                   | $0.704 \pm 0.042$                   |
|            | SMABMIL <sup>19</sup>  | $0.841 \pm 0.013$                   | $0.957 \pm 0.010$                   | $0.708 \pm 0.010$                   |
| PHIKON     | ABMIL <sup>17</sup>    | $0.819 \pm 0.020$                   | $0.937 \pm 0.018$                   | $0.683 \pm 0.027$                   |
|            | VAEABMIL <sup>18</sup> | <b><math>0.838 \pm 0.016</math></b> | <b><math>0.964 \pm 0.005</math></b> | <b><math>0.717 \pm 0.018</math></b> |
|            | TRANSMIL <sup>20</sup> | $0.827 \pm 0.026$                   | $0.954 \pm 0.006$                   | $0.697 \pm 0.052$                   |
|            | SMABMIL <sup>19</sup>  | $0.834 \pm 0.018$                   | $0.948 \pm 0.011$                   | $0.711 \pm 0.046$                   |

**Table 5.** Accuracy, AUROC and F1-Score classification results in the ASSIST test set, averaged across five train/validation splits. We report mean and standard deviation of the results. For each foundation model, the highest value in each metric is highlighted bold while the second best is underlined.

## Validation Results

The classification performance of the four MIL models is evaluated on the test set using standard metrics: accuracy, area under the ROC curve (AUROC), and F1-score. The results are shown in Table 5. Regarding the accuracy and AUROC metrics, VAEABMIL achieves the highest performance in most cases. Note that, while accuracy and AUROC remain consistently high across all four models, the F1-Score reveals a more significant performance gap between them.

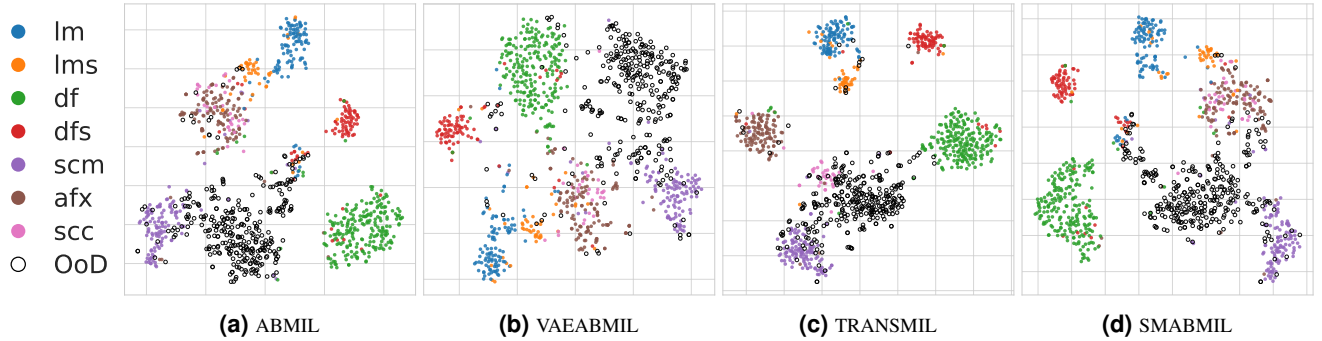


**Figure 3.** TSNE plots of the slide representations obtained by the MIL models.

Table 6 presents the results on the OoD detection task, which aims to identify slides containing secondary tumors as OoD samples. We observe that, on average, OoD scores calculated using the intermediate bag representations of the MIL models (such as KNN and Mahalanobis) achieve a higher AUROC than those derived from the models' final classification outputs. More precisely, the best and second best results are obtained by the scores Mahalanobis and KNN, respectively, using the CONCH foundation model on the bag representation of VAEABMIL. Notably, using the foundation models UNI2, VIRCHOW and PHIKON, the best OoD detector is KNN on the bag representation produced by ABMIL. In Figure 3, we show the TSNE representation of the slide-level embeddings produced by the MIL models, using the CONCH feature extractor. The figure shows that the four models separate the classes and leave OoD samples close to each other, which supports the good results of the KNN and Mahalanobis scores.

| Feat. Ext. | Model                  | KNN                  | MAHALANOBIS          | ENERGY               | ENTROPY       | MAXLOGIT      | ODIN                 | REACT         |
|------------|------------------------|----------------------|----------------------|----------------------|---------------|---------------|----------------------|---------------|
| CONCH      | ABMIL <sup>17</sup>    | 0.949 ± 0.003        | 0.924 ± 0.007        | 0.748 ± 0.100        | 0.744 ± 0.043 | 0.754 ± 0.057 | 0.755 ± 0.054        | 0.755 ± 0.056 |
|            | VAEABMIL <sup>18</sup> | <u>0.959 ± 0.005</u> | <b>0.960 ± 0.003</b> | 0.306 ± 0.046        | 0.705 ± 0.022 | 0.594 ± 0.051 | 0.645 ± 0.043        | 0.603 ± 0.045 |
|            | TRANSMIL <sup>20</sup> | 0.886 ± 0.028        | 0.872 ± 0.031        | 0.744 ± 0.046        | 0.788 ± 0.024 | 0.810 ± 0.017 | 0.822 ± 0.022        | 0.806 ± 0.017 |
|            | SMABMIL <sup>19</sup>  | 0.950 ± 0.005        | 0.914 ± 0.007        | 0.664 ± 0.070        | 0.767 ± 0.048 | 0.769 ± 0.046 | 0.777 ± 0.046        | 0.760 ± 0.048 |
| UNI        | ABMIL <sup>17</sup>    | 0.885 ± 0.021        | 0.898 ± 0.004        | 0.890 ± 0.048        | 0.847 ± 0.042 | 0.867 ± 0.048 | 0.865 ± 0.049        | 0.870 ± 0.047 |
|            | VAEABMIL <sup>18</sup> | 0.884 ± 0.021        | 0.872 ± 0.013        | 0.613 ± 0.046        | 0.863 ± 0.009 | 0.867 ± 0.014 | 0.876 ± 0.015        | 0.870 ± 0.012 |
|            | TRANSMIL <sup>20</sup> | 0.884 ± 0.016        | 0.871 ± 0.022        | <b>0.926 ± 0.010</b> | 0.881 ± 0.015 | 0.906 ± 0.012 | 0.903 ± 0.012        | 0.912 ± 0.011 |
|            | SMABMIL <sup>19</sup>  | 0.898 ± 0.018        | 0.889 ± 0.011        | 0.868 ± 0.021        | 0.868 ± 0.010 | 0.888 ± 0.008 | 0.888 ± 0.010        | 0.891 ± 0.009 |
| UNI2       | ABMIL <sup>17</sup>    | <b>0.922 ± 0.013</b> | 0.896 ± 0.009        | 0.829 ± 0.028        | 0.897 ± 0.010 | 0.905 ± 0.008 | <u>0.909 ± 0.008</u> | 0.904 ± 0.009 |
|            | VAEABMIL <sup>18</sup> | 0.867 ± 0.025        | 0.823 ± 0.019        | 0.595 ± 0.018        | 0.830 ± 0.007 | 0.852 ± 0.009 | 0.849 ± 0.011        | 0.855 ± 0.012 |
|            | TRANSMIL <sup>20</sup> | 0.896 ± 0.073        | 0.906 ± 0.030        | 0.906 ± 0.054        | 0.863 ± 0.051 | 0.891 ± 0.048 | 0.888 ± 0.045        | 0.895 ± 0.050 |
|            | SMABMIL <sup>19</sup>  | 0.862 ± 0.035        | 0.859 ± 0.019        | 0.686 ± 0.039        | 0.881 ± 0.016 | 0.889 ± 0.016 | 0.893 ± 0.016        | 0.889 ± 0.016 |
| VIRCHOW    | ABMIL <sup>17</sup>    | <b>0.908 ± 0.011</b> | 0.876 ± 0.008        | 0.737 ± 0.036        | 0.735 ± 0.034 | 0.755 ± 0.033 | 0.757 ± 0.032        | 0.751 ± 0.036 |
|            | VAEABMIL <sup>18</sup> | 0.901 ± 0.016        | 0.888 ± 0.013        | 0.588 ± 0.046        | 0.703 ± 0.020 | 0.723 ± 0.023 | 0.729 ± 0.019        | 0.737 ± 0.025 |
|            | TRANSMIL <sup>20</sup> | 0.877 ± 0.010        | 0.845 ± 0.016        | 0.722 ± 0.052        | 0.733 ± 0.055 | 0.748 ± 0.054 | 0.756 ± 0.053        | 0.745 ± 0.054 |
|            | SMABMIL <sup>19</sup>  | <u>0.907 ± 0.007</u> | 0.865 ± 0.011        | 0.725 ± 0.032        | 0.720 ± 0.023 | 0.727 ± 0.028 | 0.727 ± 0.028        | 0.721 ± 0.032 |
| PHIKON     | ABMIL <sup>17</sup>    | <b>0.912 ± 0.003</b> | 0.887 ± 0.009        | 0.705 ± 0.055        | 0.768 ± 0.035 | 0.771 ± 0.039 | 0.778 ± 0.039        | 0.776 ± 0.042 |
|            | VAEABMIL <sup>18</sup> | 0.871 ± 0.023        | 0.875 ± 0.018        | 0.723 ± 0.057        | 0.764 ± 0.021 | 0.789 ± 0.020 | 0.770 ± 0.009        | 0.796 ± 0.019 |
|            | TRANSMIL <sup>20</sup> | 0.882 ± 0.019        | 0.868 ± 0.013        | 0.718 ± 0.078        | 0.783 ± 0.028 | 0.800 ± 0.039 | 0.807 ± 0.038        | 0.799 ± 0.039 |
|            | SMABMIL <sup>19</sup>  | <u>0.906 ± 0.005</u> | 0.869 ± 0.007        | 0.675 ± 0.047        | 0.759 ± 0.018 | 0.759 ± 0.024 | 0.767 ± 0.023        | 0.757 ± 0.025 |

**Table 6.** Secondary Tumor OoD detection results (AUROC), averaged across five train/validation splits. We report mean and standard deviation results. For each foundation model, the highest value in each metric is highlighted bold while the second best is underlined.

Ultimately, these findings indicate that current approaches leave margin for future exploration and improvement in the OoD detection of secondary tumors. In this benchmark, OoD detection is evaluated at slide level, where each method produces a single OoD score for a complete WSI. Therefore, the benchmark assesses the ability to distinguish spindle-cell neoplasms from secondary tumors and other diagnostically distinct entities at slide level, which corresponds to semantic novelty, rather than to localize small OoD regions within a slide. Local tissue heterogeneity, tumor proportion, background tissue, and localized OoD regions may influence the final score through the MIL aggregation mechanism. Investigating the sensitivity of MIL-based OoD detection methods to such factors remains an important direction for future research. We note that several clinically relevant spindle-cell differentials in routine dermatopathology, such as pleomorphic dermal sarcoma (PDS), Kaposi sarcoma, angiosarcoma, and other rare entities, are not explicitly represented as distinct categories in the current dataset and are implicitly included within the OoD group. As a result, the benchmark primarily evaluates the rejection of diagnostically distinct entities rather than providing direct support for their classification.

## Code availability

The code used to process the dataset and perform the technical validation is implemented in Python and is available in GitHub (<https://github.com/vipgugr/ASSIST-technical-validation>). The foundational models used to extract the patch embeddings can be publicly accessed for research purposes from TRIDENT<sup>11</sup> and Hugging Face. TORCHMIL<sup>28</sup> is used to implement the MIL models, and the OoD scores are implemented using PyTorch.

## References

- Meseguer, P. *et al.* A Fusocellular Skin Dataset with Whole Slide Images for Deep Learning Models, [10.6084/m9.figshare.27118035.v2](https://doi.org/10.6084/m9.figshare.27118035.v2) (2025).
- del Amor, R. *et al.* A fusocellular skin dataset with whole slide images for deep learning models. *Sci. Data* **12**, [10.1038/s41597-025-05108-3](https://doi.org/10.1038/s41597-025-05108-3) (2025).
- del Amor, R. *et al.* Annotation protocol and crowdsourcing multiple instance learning classification of skin histological images: The CR-AI4SKIN dataset. *Artif. Intell. Medicine* **145**, 102686 (2023).
- Al Thelaya, K. *et al.* Applications of discriminative and deep learning feature extraction methods for whole slide image analysis: A survey. *J. Pathol. Informatics* **14**, 100335, [10.1016/j.jpi.2023.100335](https://doi.org/10.1016/j.jpi.2023.100335) (2023).
- Lee, M. Recent advancements in deep learning using whole slide imaging for cancer prognosis. *Bioengineering* **10**, 897, [10.3390/bioengineering10080897](https://doi.org/10.3390/bioengineering10080897) (2023).

6. Berman, A. G., Orchard, W. R., Gehrung, M. & Markowetz, F. SliDL: A toolbox for processing whole-slide images in deep learning. *PLOS ONE* **18**, e0289499, [10.1371/journal.pone.0289499](https://doi.org/10.1371/journal.pone.0289499) (2023).
7. Geijs, D. J. *et al.* Detection and subtyping of basal cell carcinoma in whole-slide histopathology using weakly-supervised learning. *Med. Image Analysis* **93**, 103063 (2024).
8. Zhong, L., Liao, X., Zhang, S., Zhang, S. & Wang, G. Openpath: Open-set active learning for pathology image classification via pre-trained vision-language models. In *Medical Image Computing and Computer Assisted Intervention*, 481–490 (2026).
9. Zhang, D. *et al.* Stadis: Stability distance to detecting out-of-distribution data in computational pathology. *Med. Image Analysis* **106**, 103774, <https://doi.org/10.1016/j.media.2025.103774> (2025).
10. Sáez-Maldonado, F. J. *et al.* A multicenter whole slide image dataset for classification and out-of-distribution detection in spindle cell cutaneous neoplasms. <https://www.ebi.ac.uk/biostudies/bioimages/studies/S-BIAD3206>, [10.6019/S-BIAD3206](https://doi.org/10.6019/S-BIAD3206) (2026).
11. Zhang, A., Jaume, G., Vaidya, A., Ding, T. & Mahmood, F. Accelerating data processing and benchmarking of AI models for pathology. *arXiv preprint arXiv:2502.06750* (2025).
12. Jaume, G. *et al.* Hest-1k: A dataset for spatial transcriptomics and histology image analysis. In *Advances in Neural Information Processing Systems*, vol. 37, 53798–53833, [10.52202/079017-1704](https://doi.org/10.52202/079017-1704) (2024).
13. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nat. Medicine* **30**, 863–874 (2024).
14. Vorontsov, E. *et al.* A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Medicine* **30**, 2924–2935, [10.1038/s41591-024-03141-0](https://doi.org/10.1038/s41591-024-03141-0) (2024).
15. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Medicine* (2024).
16. Filiot, A. *et al.* Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv* [10.1101/2023.07.21.23292757](https://doi.org/10.1101/2023.07.21.23292757) (2023). <https://www.medrxiv.org/content/early/2023/07/26/2023.07.21.23292757.full.pdf>.
17. Ilse, M., Tomczak, J. & Welling, M. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, 2127–2136 (PMLR, 2018).
18. Sáez-Maldonado, F. J. *et al.* Using variational autoencoders for out of distribution detection in histological multiple instance learning. *IEEE Access* **13**, 133351–133369, [10.1109/ACCESS.2025.3593420](https://doi.org/10.1109/ACCESS.2025.3593420) (2025).
19. Castro-Macías, F. M., Morales-Álvarez, P., Wu, Y., Molina, R. & Katsaggelos, A. K. Sm: enhanced localization in multiple instance learning for medical imaging classification. In *Advances in Neural Information Processing Systems*, vol. 37, 77494–77524 (2024).
20. Shao, Z. *et al.* Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, vol. 34, 2136–2147 (2021).
21. Liu, W., Wang, X., Owens, J. & Li, Y. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, vol. 33, 21464–21475 (2020).
22. Hendrycks, D. & Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations* (2017).
23. Basart, S., Mantas, M., Mohammadreza, M., Jacob, S. & Dawn, S. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning* (2022).
24. Liang, S., Li, Y. & Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations* (2018).
25. Sun, Y., Guo, C. & Li, Y. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, vol. 34, 144–157 (2021).
26. Sun, Y., Ming, Y., Zhu, X. & Li, Y. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 20827–20840 (PMLR, 2022).
27. Lee, K., Lee, K., Lee, H. & Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, vol. 31 (2018).
28. Castro-Macías, F. M., Sáez-Maldonado, F. J., Morales-Álvarez, P. & Molina, R. Torchmil: A pytorch-based library for deep multiple instance learning. *Neurocomputing* **680**, 133286, <https://doi.org/10.1016/j.neucom.2026.133286> (2026).

## Author contributions statement

Conceptualization: FJ.S-M., J.M., S.M., L.T., A.M-C., J.G-V., R.M. and V.N. Data Curation: FJ.S-M., J.M., P.Me., A.G., S.M., L.T., A.M-C., J.G-V. and P.T. Formal Analysis: JF.S-M. and M.L-P. Funding Acquisition: P.Mo., R.M. and V.N. Investigation: JF.S-M., J.M., L.T., A.M-C., J.G-V. and P.T. Methodology: JF.S-M., J.M., S.M., P.Mo., R.M. and V.N. Software: FJ.S-M. and J.M. Supervision: P.Mo., R.M. and V.N. Writing – original draft : FJ.S-M., J.M. and A.M-C. Writing – review & editing: M.L-P., P.Me., A.G., S.M., R.A., L.T., J.G-V., P.Mo., R.M. and V.N. All authors have read and agreed to the published version of the manuscript.

## Acknowledgements and funding

This work was partially supported Grants PID2022-140189OB-C21 and PID2022-140189OB-C22 (ASSIST) funded by MICIU/AEI/10.13039/501100011033 and by “ERDF A way of making Europe”, by the European Union and partially by GVA through project CIPROM/2022/20.