



Article

Interleaved Fusion Learning for Trustworthy AI: Improving Cross-Dataset Performance in Cervical Cancer Analysis

Carlos Martínez ^{1,2} , Laura Busto ^{1,2} , Olivia Zulaica ^{1,2} and César Veiga ^{1,2,*} ¹ Cardiology Research Group, Galicia Sur Health Research Institute (IIS Galicia Sur), 36312 Vigo, Spain; carlos.martinez@iisgaliciasur.es (C.M.); laura.busto@iisgaliciasur.es (L.B.); olivia.zulaica@iisgaliciasur.es (O.Z.)² AI Platform, Galicia Sur Health Research Institute (IIS Galicia Sur), 36312 Vigo, Spain

* Correspondence: cesar.veiga@iisgaliciasur.es

Abstract

This study introduces a novel Interleaved Fusion Learning (IFL) methodology leveraging transfer learning to generate a family of models optimized for specific datasets while maintaining superior generalization performance across others. The approach is demonstrated in cervical cancer screening, where cytology image datasets present challenges of heterogeneity and imbalance. By interleaving transfer steps across dataset partitions and regulating adaptation through a dynamic learning parameter, IFL promotes both domain-specific accuracy and cross-domain robustness. To evaluate its effectiveness, complementary metrics are used to capture not only predictive accuracy but also fairness in performance distribution across datasets. Results highlight the potential of IFL to deliver reliable and unbiased models in clinical decision support. Beyond cervical cytology, the methodology is designed to be scalable to other medical imaging tasks and, more broadly, to domains requiring equitable AI solutions across multiple heterogeneous datasets.

Keywords: transfer learning; data fusion; deep learning; medical imaging; cervical cancer screening; pattern recognition; AI fairness; healthcare bias mitigation; trustworthy AI



Academic Editor: Razavi-Far Roozbeh

Received: 1 September 2025

Revised: 5 October 2025

Accepted: 14 October 2025

Published: 23 October 2025

Citation: Martínez, C.; Busto, L.; Zulaica, O.; Veiga, C. Interleaved Fusion Learning for Trustworthy AI: Improving Cross-Dataset Performance in Cervical Cancer Analysis. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 128. <https://doi.org/10.3390/make7040128>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cervical cancer remains one of the leading health threats for women worldwide, especially in low- and middle-income countries, where it is the second-most prevalent cancer among women [1,2]. Globally, over 600,000 new cases and over 340,000 deaths were reported in 2020, of which about 90% occurred in resource-limited regions [3]. The high mortality in these areas largely stems from inadequate access to screening, human papillomavirus (HPV) vaccination, and timely medical intervention [4]. HPV infection is recognized as the primary cause of cervical cancer, yet barriers to implementing preventive measures persist due to economic and healthcare disparities [5]. Advances in diagnostic tools, including machine learning and artificial intelligence, show promise for enhancing early detection, which is critical given that early-stage intervention could reduce cervical cancer mortality by up to 60% [6]. More broadly, AI and machine learning (ML) applications are increasingly transforming healthcare by improving diagnostics, optimizing treatments, and enhancing patient outcomes [7,8].

The integration of Deep Learning (DL) in cervical cancer diagnostics has advanced significantly in recent years, with applications that enhance automated image analysis and assist in early detection [9,10]. DL techniques now focus extensively on cervical cell classification, leveraging large datasets to train models, although publicly available data

remains limited [11]. For this reason, transfer learning has been widely employed to address data scarcity by adapting pre-trained models from large, general datasets [12]. This method has shown particular promise in cervical cell detection tasks, allowing networks to achieve higher accuracy despite constrained data [13]. As DL continues to evolve, its impact on cervical cancer detection grows, offering scalable solutions that aid in screening and reduce the diagnostic burden on pathologists.

Ensuring trustworthiness in AI models is essential, especially in healthcare, where biases in training data or model structures can disproportionately impact certain demographic groups. Recent research emphasizes the need for fairness-focused metrics and strategies to mitigate these biases, which are present in diverse applications such as environmental security and semantic segmentation, and now in clinical predictions as well [14–16]. Despite advances in algorithmic fairness, the evaluation and reduction in biases within clinical models remains challenging due to the limited accessibility of representative health data and the complexity of high-dimensional medical information [17]. In healthcare, these biases can lead to disparities in diagnostic outcomes, reinforcing existing inequities. Addressing these issues requires a comprehensive approach involving data quality enhancement, explicitly fair algorithms, and interdisciplinary collaboration to ensure ethical, equitable AI deployment in clinical settings [18].

Transfer learning has emerged as a powerful technique to mitigate the effects of data scarcity and to address biases in medical image analysis by leveraging pre-trained models from tasks with varying dataset sizes, regardless of their relative scale [19,20]. In medical applications, transfer learning enables models to inherit feature representations from large, well-curated datasets, which can be fine-tuned to perform effectively even with limited domain-specific data [21,22]. This is particularly useful for complex tasks like cervical cell detection, where annotated data is often limited. Studies show that using domain-specific transfer learning can significantly improve model accuracy and reduce bias by aligning features more closely with the target medical context [13,22]. This approach not only enhances model robustness but also minimizes the need for extensive manual labeling, making it a viable solution to achieve fairer and more accurate predictions in clinical settings. Building on these foundations, this study introduces a novel dataset fusion methodology that further addresses the challenge of balancing domain-specific optimization with cross-domain adaptability.

The purpose of this work is to develop a methodology for creating more robust and Trustworthy Artificial Intelligence (TAI) models for health applications that mitigate bias effects and address the challenges described above. This objective will be achieved by developing a novel methodology, Interleaved Fusion Learning, that can be applied to a family of models, each specialized for a specific dataset, while also benefiting from shared knowledge across all datasets. By allowing models to share knowledge while maintaining specialization for their respective datasets, this methodology facilitates a synergistic fusion of diverse data sources, integrating heterogeneous data to enhance predictive performance across a broader range of data. The proposed methodology will be evaluated using cervical cancer datasets to demonstrate its potential in improving screening solutions.

The organization of this paper is outlined as follows: initially, Section 2 provides an overview of the materials and techniques essential for conducting this research. Subsequently, Section 3 provides an explanation of the proposed approach for the development and evaluation of the interleaved fusion learned models. In Section 4, we detail our implementation of the methodology and present a selection of the data obtained through our algorithm. Finally, in Sections 5 and 6, we engage in an analysis of the data and seek to draw some definitive conclusions.

2. Materials and Methods

2.1. Open Cervical Cancer Datasets

To develop robust models for cervical cell segmentation and classification, large and well-annotated datasets are essential. Among the most comprehensive publicly available resources, the APACC (Annotated PAP cell images and smear slices for Cell Classification) dataset and the CRIC (Center for Recognition and Inspection of Cells) Cervix collection stand out for their extensive cell annotations and segmentation data, which make them highly suitable for DL applications [23,24]. Both datasets offer the detailed, large-scale data necessary for effective model training and evaluation in the context of cervical cancer screening, supporting a range of tasks from detection to classification. Given their quality and scope, APACC dataset and CRIC Cervix collection provide an invaluable foundation for advancing automated cytological analysis.

In contrast, other commonly used datasets in cervical cell research, such as SIPaKMeD, Herlev and Mendeley, are less suited to our objectives. SIPaKMeD and Herlev datasets focus on isolated cells and are primarily used for image-based classification rather than full-image detection and segmentation, limiting their application for models requiring comprehensive smear data [25–27]. On the other hand, the Mendeley dataset includes images with pointed-out cells but lacks full cell labeling, which restricts its effectiveness for segmentation-focused tasks [27,28].

2.1.1. CRIC Cervix Collection

The CRIC Cervix collection is a robust dataset, specifically curated to support automated analysis and detection in cervical cytology. Created as part of the CRIC initiative, this dataset includes 400 high-resolution RGB images (1376×1020 pixels), each containing manually classified cells. With a total of over 11,000 annotated cells, the CRIC Cervix collection offers the high-quality labels needed for machine learning models focused on cytopathological tasks [24].

The CRIC Cervix collection uses the Bethesda System, which is the standardized terminology most widely adopted worldwide for cervical cytopathology, ensuring uniformity and reproducibility across laboratories and pathologists [29]. This dataset classifies cells into six categories based on Bethesda nomenclature: (1) negative for intraepithelial lesion or malignancy (NILM), (2) atypical squamous cells of undetermined significance, possibly non-neoplastic (ASC-US), (3) low-grade squamous intraepithelial lesion (LSIL), (4) atypical squamous cells that cannot exclude a high-grade lesion (ASC-H), (5) high-grade squamous intraepithelial lesion (HSIL), and (6) squamous cell carcinoma (SCC). To streamline model development and enhance classification homogeneity with the APACC database, we have unified these categories into a binary classification system. The NILM category is maintained as is, while all other categories are combined and labeled as “Positive”.

The reference images for each category are shown in Figure 1, with subfigures (a) through (f) illustrating representative cells from each class.

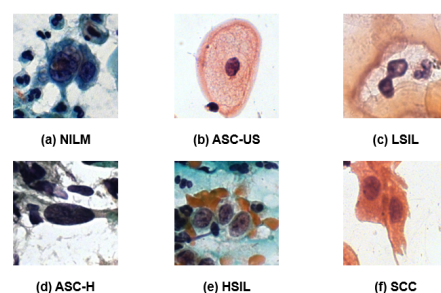


Figure 1. CRIC dataset: 6 types of cells. The images are 100×100 pixel squares extracted from the original images, centered on the diagnostic position indicated by the dataset labels.

2.1.2. APACC Dataset

The APACC dataset is one of the most recent and comprehensive publicly available resources for cervical cell analysis. This dataset includes 103,675 annotated cell images, extracted from 107 whole Pap smears, and divided into over 21,000 sub-regions to support finer analysis [23]. These sub-regions are RGB images with a resolution of 1984×1984 pixels.

The APACC dataset categorizes cells into four classes: healthy (normal), unhealthy (abnormal), rubbish (not valid), and bothcells (a mixture of healthy and unhealthy cells). These classes loosely map onto Bethesda categories, where “healthy” corresponds to the NILM category, “unhealthy” represents cells from the broader Epithelial cell abnormality category (though without subdivisions like ASC, LSIL, or HSIL), “rubbish” aligns with Unsatisfactory for evaluation, and “bothcells” also falls within Epithelial cell abnormality as it includes malignant cells intermingled with normal cells. For our analysis, we have simplified the dataset by applying the same binary classification approach as in the CRIC database: NILM cells remain as “Negative,” while both “unhealthy” and “bothcells” are consolidated under a “Positive” label. The “rubbish” class has been excluded to ensure data relevance and consistency.

Representative examples of each original class from APACC are shown in Figure 2, with subfigures (a) through (d) illustrating these types.

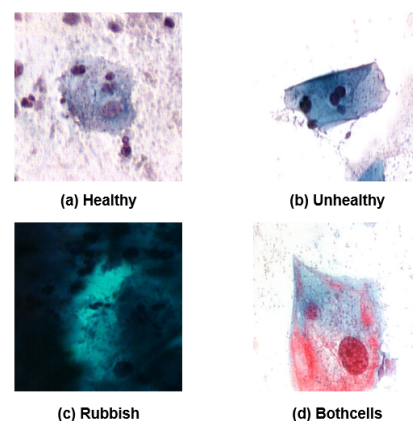


Figure 2. APACC dataset: 4 types of cells. The images are 100×100 pixel squares extracted from the original images, centered on the diagnostic position indicated by the dataset labels.

2.2. Deep Learning Architectures for Object Recognition: YOLOv8

YOLOv8 (You Only Look Once, Version 8), developed by Ultralytics, is one of the most advanced and efficient architectures for real-time object detection in medical imaging [30,31]. This DL model combines high accuracy with speed, making it ideal for applications requiring rapid identification of abnormalities, such as cervical cancer screening. With its optimized structure, YOLOv8 is particularly effective in identifying abnormal cells within complex, high-resolution images, which is crucial for early detection of precancerous and malignant lesions.

Its flexibility also allows it to be adapted to clinical environments where both diagnostic precision and processing speed are essential. Supporting both detection and classification tasks, YOLOv8 enhances the efficiency of cytological analysis, contributing to faster and more reliable early cancer screening workflows.

YOLOv8 provides several adjustable parameters that allow fine-tuning for specific tasks and datasets. The initial learning rate controls the speed of weight updates during training, while the optimizer manages how the model minimizes the loss function. Early stopping prevents overfitting by halting training when validation metrics cease to improve. Data augmentation enhances generalization by applying random transformations to the

training data. The batch size defines the number of samples processed before updating model weights, and the image size determines the resolution at which images are resized for input, balancing accuracy and computational efficiency.

2.3. Transfer Learning

Transfer learning is a DL approach that enables models to leverage knowledge acquired from one task (source task) to improve performance on a related task (target task), especially useful when the target domain has limited labeled data [32].

Formally, a domain is defined as $D = \{\mathcal{X}, \mathcal{Y}, d(\cdot)\}$, where $\mathcal{X}$ is a feature space, $\mathcal{Y}$ is a label space, and the function $d : \mathcal{X} \rightarrow \mathcal{Y}$ ensures that each feature $x \in \mathcal{X}$ has a corresponding label $y \in \mathcal{Y}$, such that $d(x) = y$. Given a specific domain D , we define a task $T = \{X, \mathcal{Y}\}$, where $X \subset \mathcal{X}$ is a subset of the feature space. For a specific task T , any function $f(\cdot)$ learned based on the relationships determined by the image of $d(X)$ is referred to as a predictive function.

In transfer learning, the objective is to enhance a predictive function $f_T(\cdot)$ in a target domain $D_T = \{\mathcal{X}_T, \mathcal{Y}_T, d_T(\cdot)\}$ with a corresponding target task $T_T = \{X_T, \mathcal{Y}_T\}$, by leveraging knowledge from a source predictive function $f_S(\cdot)$ learned from a source domain $D_S = \{\mathcal{X}_S, \mathcal{Y}_S, d_S(\cdot)\}$ and a source task $T_S = \{X_S, \mathcal{Y}_S\}$. This process is valid under the assumption that either $D_S \neq D_T$ or $T_S \neq T_T$.

The condition $D_S \neq D_T$ implies that the feature spaces, label spaces, or mapping functions differ between the source and target domains, i.e., $\mathcal{X}_S \neq \mathcal{X}_T$, $\mathcal{Y}_S \neq \mathcal{Y}_T$, or $d_S(\cdot) \neq d_T(\cdot)$. Conversely, $T_S \neq T_T$ implies that the tasks differ, either in the subset of the feature space $X_S \neq X_T$ or in the label space $\mathcal{Y}_S \neq \mathcal{Y}_T$.

Enhancing Transfer Learning

Two primary strategies for improving transfer learning outcomes are fine-tuning and weight initialization. Fine-tuning involves initializing a new model with pre-trained weights from the source domain and adapting specific layers to target domain data. This can involve adjusting all layers, or selectively fine-tuning only the last layers tailored to the target task. Weight initialization, by contrast, freezes certain pre-trained layers to retain general feature representations while adapting to the new domain through the remaining layers [19].

Let θ_S represent the learned parameters of the source function $f_S(\cdot)$. In fine-tuning, we initialize the target model $f_T(\cdot)$ with parameters $\theta_T = \theta_S$ and proceed to optimize a subset or the entirety of θ_T using labeled data from D_T to better fit T_T . On the other hand, weight initialization can be expressed by partitioning θ_S into two parameter sets, θ_S^{frozen} and $\theta_S^{\text{tunable}}$, such that $\theta_T = \{\theta_S^{\text{frozen}}, \theta_S^{\text{tunable}}\}$, where only $\theta_S^{\text{tunable}}$ is optimized with the target domain data, preserving the general representations from θ_S^{frozen} for use in D_T .

3. Method

In this section we present Interleaved Fusion Learning (IFL), a new methodology designed to enhance model robustness and bias mitigation. The core concept behind this approach is to develop a sequence of models that effectively perform in their source datasets while adding essential knowledge from other datasets in the sequence. To determine whether this objective is achieved, a comprehensive evaluation framework will be established to assess and compare results across an arbitrary number n of datasets.

In Figure 3, we present a schematic representation of the methodology pipeline. The diagram illustrates the IFL process: initially, each dataset $\mathcal{D}_i$ is employed to obtain a specific model f_i trained only on this dataset. Furthermore, a global model f is trained on the combined dataset $\mathcal{D} = \bigcup_{i=1}^n \mathcal{D}_i$. The specific models f_i then undergo the IFL pipeline,

to produce a set of final models $\{f_{T_i}\}_{i=1}^n$, each model f_{T_i} specialized in $\mathcal{D}_i$. Each final model f_{T_i} is evaluated against the global model f on the dataset $\mathcal{D}_i$, and against its corresponding initial model f_i on the remaining datasets $\mathcal{D}_j$, with $j \neq i$.

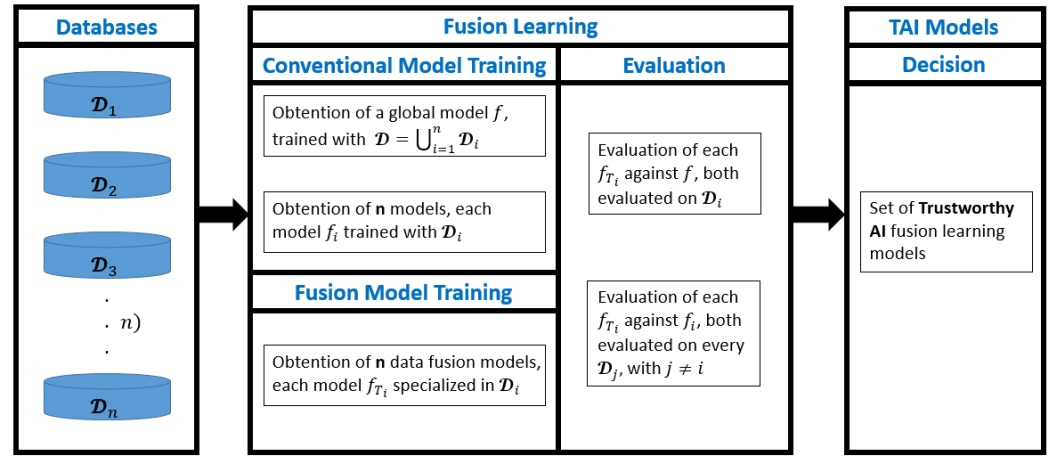


Figure 3. Schematic representation of the methodology pipeline.

To optimize performance across datasets while ensuring robustness and fairness, we start by formally defining the problem, which serves as the foundation for the proposed methodology.

3.1. Problem Definition

Building on the concepts introduced in Section 2.3, transfer learning offers several key advantages for developing TAI systems. By leveraging transfer learning, models can be designed with enhanced robustness and fairness, mitigating biases inherent in training data while preserving performance specific to the target domain.

One approach to enhancing trustworthiness is through integrating a primary, potentially biased dataset D_B , which represents the model's main objective, with a secondary, less biased or unbiased dataset D_U . Transfer learning between these datasets can adjust parameters in a way that preserves key knowledge from D_B while mitigating bias through fine-tuning with D_U . Formally, let θ_B and θ_U denote the learned parameters from D_B and D_U , respectively. Let $\theta_T = \{\theta_B^{\text{frozen}}, \theta_B^{\text{tunable}}\}$ be a partition of θ_B , and we can then optimize θ_T such that only $\theta_B^{\text{tunable}}$ is fine-tuned using samples from D_U , adding robustness against bias while retaining essential information from D_B .

The principles outlined above set the foundation for applying sequential transfer learning, where each dataset contributes to refining the model parameters progressively.

3.2. Sequential Transfer Learning

The methodology we propose defines a domain $D = \{\mathcal{X}, \mathcal{Y}, d(\cdot)\}$ for each dataset, where $\mathcal{X}$ is the set of features in the elements of the dataset, $\mathcal{Y}$ is the set of their possible labels, and d is the optimal function that correctly classifies the features with their corresponding labels. We also define the task $T = \{X, \mathcal{Y}\}$, where X represents the labeled features of the dataset, i.e., the $x \in \mathcal{X}$ for which the image of $d(x)$ is known. Specifically, each dataset can be represented as $\mathcal{D}_i = \{D_i, T_i\}$, where $D_i = \{\mathcal{X}_i, \mathcal{Y}_i, d_i(\cdot)\}$ defines the domain with feature space $\mathcal{X}_i$, label space $\mathcal{Y}_i$, and mapping function $d_i(\cdot)$, and $T_i = \{X_i, \mathcal{Y}_i\}$ represents the unique task associated with the labeled subset $X_i \subset \mathcal{X}_i$ and label space $\mathcal{Y}_i$. Each trained model f_i on a dataset $\mathcal{D}_i$ functions as a predictive function $f_i(\cdot)$, and its learned parameters θ_i correspond to the weights of $f_i(\cdot)$.

The process is repeated sequentially for every $i \in \{1, \dots, n\}$, where n is the total number of available datasets. To begin, we define a baseline model $f_i(\cdot)$, trained from scratch using only the source dataset $\mathcal{D}_i$, resulting in initial weights θ_i . We then construct a sequence by choosing a subsequent index j defined by

$$j = 1 + ((i + k - 1) \bmod n), \quad (1)$$

where $k \in \{1, \dots, n - 1\}$ represents the step within the sequence. For each j in this sequence, a model $f_{ij}(\cdot)$ is obtained by training a model in the dataset $\mathcal{D}_i \cup \mathcal{D}_j$, initializing it with the weights from training on the previous dataset in the sequence, denoted as $\theta_{j(\text{prev})}$, where

$$j(\text{prev}) = 1 + ((j + k - 2) \bmod n). \quad (2)$$

This sequence continues iteratively through each dataset until we reach $j = i$, obtaining a final model $f_{T_i}(\cdot)$ that has been adapted across all datasets in $\{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_n\}$. After completing this sequence for a given i , the process is repeated for each $i \in \{1, \dots, n\}$, resulting in a set of n models $\{f_{T_i}(\cdot)\}_{i=1}^n$, each initialized with a different dataset. A schematic representation of this method is shown in Figure 4.

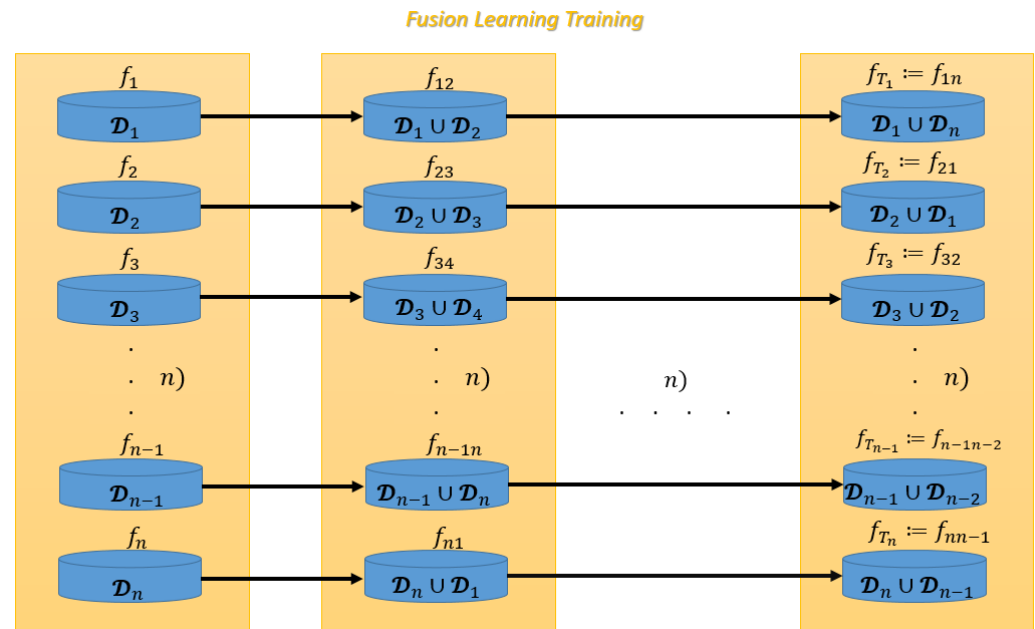


Figure 4. Scheme of the process for obtaining each f_{T_i} .

Let $f(\cdot)$ represent a model trained from scratch on the combined dataset $\mathcal{D} = \bigcup_{i=1}^n \mathcal{D}_i$. Our goal for each transfer-learned model $f_{T_i}(\cdot)$ is to outperform $f(\cdot)$ when evaluated on its initial dataset $\mathcal{D}_i$. Additionally, each $f_{T_i}(\cdot)$ should yield better results on any other dataset $\mathcal{D}_j$ (where $j \neq i$) compared to the baseline model $f_i(\cdot)$, which is trained solely on $\mathcal{D}_i$. This criterion aims to demonstrate the advantage of transfer learning in enhancing performance consistency and fairness across diverse domains.

Learning Rate Selection for Transfer Learning

For each new dataset $\mathcal{D}_j$, this methodology aims to set a learning rate that reflects the model's performance on $\mathcal{D}_j$. A high learning rate can reduce bias but may cause the model

to lose specificity for the source datasets. To balance this trade-off, the new learning rate for transfer learning from $\mathcal{D}_{j(\text{prev})}$ to $\mathcal{D}_j$ is determined as

$$\eta_j = \min \left(\frac{\sqrt{\lambda}}{|X_j|} \sum_{x_j \in X_j} \mathcal{A}(f_{ij(\text{prev})}(x_t)), \lambda \right), \quad (3)$$

where

$$\mathcal{A}(f(x_t)) = \begin{cases} 1 & \text{if } f_{ij(\text{prev})}(x_t) \neq y_t, \\ 0 & \text{if } f_{ij(\text{prev})}(x_t) = y_t, \end{cases} \quad (4)$$

with $\lambda \in [0, 1]$ controlling the maximum learning rate, and y_t representing the label corresponding each $x_t \in X_j$ in the dataset $\mathcal{D}_j$.

The choice of this learning rate definition is motivated by several factors. Primarily, we aim for $\eta_j \in [0, 1]$, as is typical in transfer learning. This holds because the sum $\sum_{x_t \in X_j} \mathcal{A}(f_{ij(\text{prev})}(x_t))$ is an integer in $[0, |X_j|]$, so dividing by $|X_j|$ normalizes it to $[0, 1]$. Multiplying by $\sqrt{\lambda}$, where $\lambda \in [0, 1]$, ensures

$$\sqrt{\lambda} \cdot \frac{\sum_{x_t \in X_j} \mathcal{A}(f_{ij(\text{prev})}(x_t))}{|X_j|} \in [0, 1]. \quad (5)$$

Since η_j is the minimum of this product and λ , both in $[0, 1]$, it follows that $\eta_j \in [0, 1]$. The parameter λ controls the maximum learning rate, setting an upper bound of $\eta_j \leq \lambda$. By using $\sqrt{\lambda}$ in the other term of the minimum function, the learning rate is modulated by λ even when the error measure is low. This setup allows flexibility to adjust λ based on specific model requirements, enabling either a higher or lower learning rate as needed.

In Figure 5, we present a graph to illustrate the possible values of η_j based on the error rate on the new dataset, $\sum_{x_t \in X_j} \mathcal{A}(f_{ij(\text{prev})}(x_t)) / |X_j|$, and the chosen value of λ .

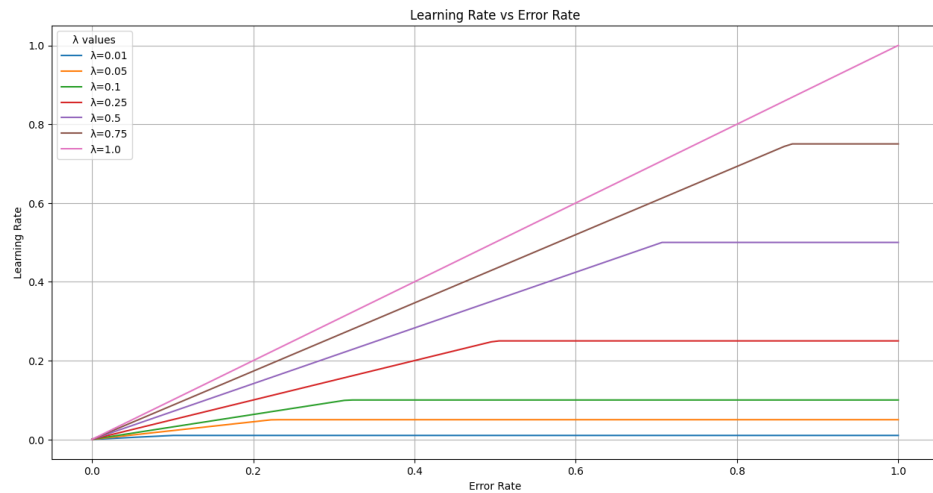


Figure 5. Learning rate η_j as a function of error rate on the new dataset, $\sum_{x_t \in X_j} \mathcal{A}(f_{ji(\text{prev})}(x_t)) / |X_j|$, and λ values.

3.3. Evaluation

Once f_{T_i} is trained, it is evaluated through two comparisons. First, f_{T_i} is compared to the model f on the initial dataset $\mathcal{D}_i$. Second, f_{T_i} is compared to f_i on datasets different from $\mathcal{D}_i$. Figure 6 illustrates these comparisons.

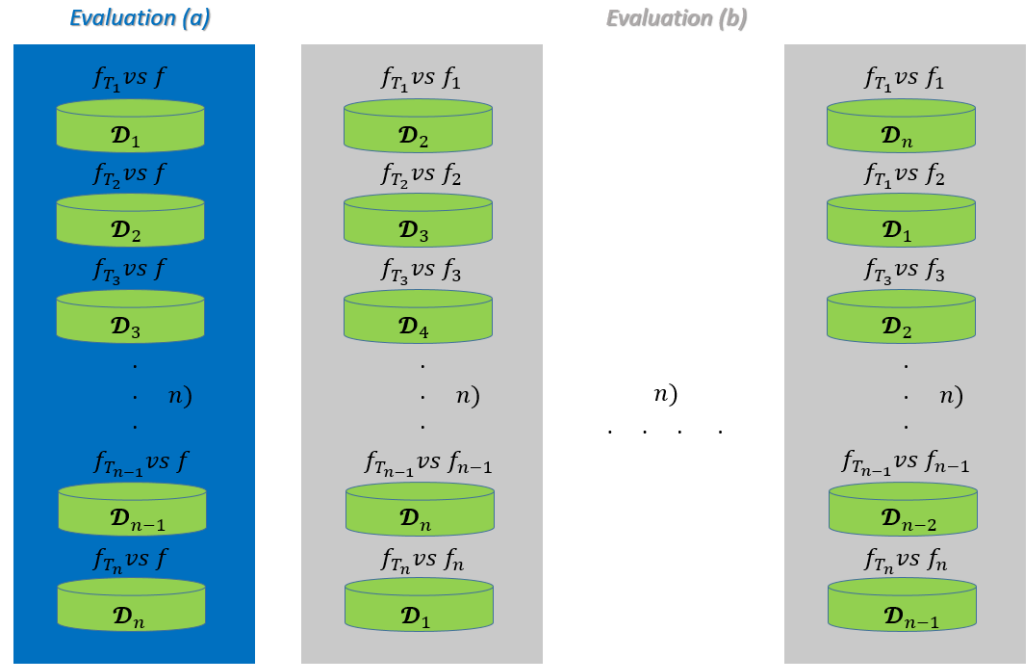


Figure 6. Scheme of the process for evaluating each f_{T_i} . Evaluation (a) compares f_{T_i} with the global model f on the dataset $\mathcal{D}_i$. Evaluation (b) compares f_{T_i} with the initial model f_i on the remaining datasets $\mathcal{D}_j$, where $j \neq i$.

To comprehensively evaluate the performance of our methodology, we designed an evaluation protocol that focuses on three key aspects: (1) measuring cross-domain generalization through a structured performance matrix, (2) analyzing performance deviations to assess consistency and robustness, and (3) benchmarking IFL models against baseline and global models to validate their effectiveness.

3.3.1. Cross-Domain Performance Matrix

To systematically evaluate the IFL models, we propose a cross-domain performance matrix M , where each entry $M_{ij}(\text{Metric})$ represents the performance of model $f_{T_i}(\cdot)$ when evaluated on dataset $\mathcal{D}_j$ using a chosen evaluation metric. The matrix is constructed as follows:

$$M_{ij}(\text{Metric}) = \text{Metric}(f_{T_i}(\mathcal{D}_j)), \quad (6)$$

where Metric is a user-specified evaluation metric, such as Accuracy, F1-score, or AUC-ROC. The diagonal entries $M_{ii}(\text{Metric})$ quantify the model's performance on its source dataset using the selected metric, while off-diagonal entries $M_{ij}(\text{Metric})$ (for $i \neq j$) measure its generalization to other datasets.

3.3.2. Performance Deviation Analysis

We define two aggregate metric vectors to summarize cross-domain performance using the elements of the Cross-Domain Performance Matrix ($M_{ij}(\text{Metric})$):

- **Cross-Domain Generalization Score (CDGS):** The vector whose elements are the average performance of $f_{T_i}(\cdot)$ on all datasets other than its source, computed as

$$\text{CDGS}_i(\text{Metric}) = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n M_{ij}(\text{Metric}), \quad (7)$$

where Metric determines the range of $CDGS_i(\text{Metric})$. For normalized metrics such as accuracy or F1-score, $CDGS_i(\text{Metric}) \in [0, 1]$. Higher values indicate better generalization across datasets.

- Performance Variance (PV): The vector whose elements are the variance of each $M_{ij}(\text{Metric})$ across all datasets, reflecting the consistency of the model's performance:

$$PV_i(\text{Metric}) = \frac{1}{n} \sum_{j=1}^n (M_{ij}(\text{Metric}) - \bar{M}_i(\text{Metric}))^2, \quad (8)$$

where $\bar{M}_i(\text{Metric}) = \frac{1}{n} \sum_{j=1}^n M_{ij}(\text{Metric})$ is the mean performance of $f_{T_i}(\cdot)$ across all datasets. Lower PV values indicate more consistent performance, while higher values suggest variability or bias. Again, for normalized metrics, we have $M_{ij}(\text{Metric}) \in [0, 1]$, and consequently $PV_i(\text{Metric}) \in [0, 1]$, with smaller values indicating greater consistency across datasets in this case.

3.3.3. Benchmark Comparison Matrix

To validate the effectiveness of our methodology, we construct a benchmark comparison matrix B , which directly mirrors the structure of the cross-domain performance matrix M . The entries of B are defined as follows:

$$B_{ij}(\text{Metric}) = \begin{cases} \text{Metric}(f(\mathcal{D}_i)), & \text{if } i = j, \\ \text{Metric}(f_i(\mathcal{D}_j)), & \text{if } i \neq j. \end{cases} \quad (9)$$

The diagonal entries $B_{ii}(\text{Metric})$ represent the performance of the global model $f(\cdot)$ on dataset $\mathcal{D}_i$, while the off-diagonal entries $B_{ij}(\text{Metric})$ (for $i \neq j$) capture the performance of the baseline model $f_i(\cdot)$ on the other datasets $\mathcal{D}_j$.

By directly comparing the cross-domain performance matrix M with the benchmark comparison matrix B , we can evaluate the improvements introduced by the IFL models $f_{T_i}(\cdot)$. Specifically, this comparison enables the following:

- Assessment of Cross-Domain Generalization: Comparing $M_{ij}(\text{Metric})$ against $B_{ij}(\text{Metric})$ (for $i \neq j$) reveals whether the IFL model $f_{T_i}(\cdot)$ generalizes better to other datasets $\mathcal{D}_j$ than the baseline model $f_i(\cdot)$.
- Evaluation of Source Dataset Performance: Comparing $M_{ii}(\text{Metric})$ with $B_{ii}(\text{Metric})$ indicates whether the IFL model $f_{T_i}(\cdot)$ outperforms the global model $f(\cdot)$ on its original dataset $\mathcal{D}_i$.

This benchmarking approach allows for a straightforward comparison, as each element in both matrices represents a single value of the chosen metric. This simplicity ensures that performance differences can be directly interpreted while preserving the properties of the utilized metric, such as normalization or scale consistency.

4. Results

In this section, we present the results obtained from the application of the proposed IFL methodology, as described in Section 3. The implementation of the IFL methodology was applied with the APACC and CRIC Cervix datasets defined in Section 2.1. In order to increase the dimensionality of the experiment, each dataset was first divided into two halves, resulting in four distinct datasets to facilitate multidimensional analysis: $\mathcal{D}_{1a}$, $\mathcal{D}_{1b}$, $\mathcal{D}_{2a}$, and $\mathcal{D}_{2b}$, without losing generality. Each of these datasets was further partitioned into training (80%) and validation (20%) subsets, with an equal split to ensure a balanced distribution of “Negative” and “Positive” classes across all partitions. The YOLOv8 architecture, as detailed in Section 2.2, was employed as the base DL model.

4.1. Implementation Details

We begin by providing details about the code implementation specific to our methodology. All code was specifically implemented for this study and executed on an NVIDIA A100 GPU to ensure efficient training and inference.

4.1.1. Training Cost and Deployment Feasibility

Each baseline YOLOv8 model required approximately 4–5 GPU-hours on the NVIDIA A100. Transfer steps were faster, taking around 1–2 GPU-hours each, resulting in a total training cost close to 30 GPU-hours for the full set of experiments. With partial freezing, approximately 35% of the parameters were updated at each step, while the remaining layers were kept frozen to preserve previously acquired features. These figures provide a reference for the computational budget and the feasibility of applying the proposed methodology in practice.

4.1.2. Dataset Preparation and Binary Label Mapping

The datasets were organized in the YOLO format, where each sample $x_t \in X_j$ was an annotated cell with associated bounding boxes and labels $y_t \in \mathcal{Y}$. For consistency, the class mappings were defined as

$$\mathcal{Y} = \{\text{"Negative"} : 0, \text{"Positive"} : 1\}.$$

Both datasets ($\mathcal{D}_j$) were processed to align with the binary classification framework:

- CRIC Cervix: The six Bethesda categories (Figure 1) were mapped to two labels: "Negative" ($y = 0$) and "Positive" ($y = 1$) as outlined in Section 2.1.
- APACC: The original four classes (Figure 2) were similarly consolidated, with "rubbish" excluded from analysis, resulting in "Negative" ($y = 0$) and "Positive" ($y = 1$) labels as outlined in Section 2.1.

4.1.3. Model Initialization and Sequential Transfer

The sequential transfer learning methodology was implemented by first training four distinct models, each initialized using one of the four datasets ($\mathcal{D}_{1a}$, $\mathcal{D}_{1b}$, $\mathcal{D}_{2a}$, $\mathcal{D}_{2b}$) as the starting point. The YOLOv8 model f_j for each dataset $\mathcal{D}_j$ was trained from scratch using the training subset of the dataset. The model was trained for 300 epochs using the hyperparameters detailed in Section 2.2. For each dataset, the training was conducted using

- Initial learning rate (η_0): 0.001.
- Optimizer: Adam.
- Early stopping: Patience of 50 epochs.
- Data augmentation: Enabled.
- Batch size: 16.
- Image size: 640×640 pixels.

After training the initial models f_{1a} , f_{1b} , f_{2a} , and f_{2b} , sequential transfer learning was applied to each, reusing the trained weights as initialization for subsequent fine-tuning on the remaining three datasets. During this process, the initial layers were manually frozen to retain previously acquired features. The learning rate η_j for fine-tuning was calculated dynamically as described in Equation (3), being

- $\lambda = 0.001$: Maximum learning rate.
- $|X_j^{\text{val}}|$: Cardinality of the validation set.
- IoU threshold: 0.5, ensuring prediction–ground truth alignment.
- $\mathcal{A}(f(x_t))$: Indicates prediction errors, as defined in Section Learning Rate Selection for Transfer Learning.

This resulted in a total of 12 transfer learning processes, each involving

1. Validation of the current model $f_{ij_{\text{prev}}}$ on X_j^{val} to calculate η_j .
2. Fine-tuning the model f_{ij} on the training subset X_j^{train} of the new dataset $\mathcal{D}_i$ using the dynamically adjusted η_j .

At the end of each fine-tuning iteration, the model performance was evaluated on X_j^{val} to assess the impact of the IFL process.

4.2. Evaluation of Performance Matrices

In this section, we focus on evaluating the results summarized in two matrices previously defined in Section 3.3: the Cross-Domain Performance Matrix M , which captures the performance of the IFL models, and the Benchmark Comparison Matrix B , which serves as a baseline for comparison.

4.2.1. Cross-Domain Performance Matrix

The cross-domain performance matrix M summarizes the evaluation of each model $f_{T_i}(\cdot)$ on all datasets $\mathcal{D}_j$ using accuracy. Results are presented in Table 1.

Table 1. Cross-Domain Performance Matrix M for F1-score.

Model	$\mathcal{D}_{1a}$	$\mathcal{D}_{1b}$	$\mathcal{D}_{2a}$	$\mathcal{D}_{2b}$
$f_{T_{1a}}$	0.52	0.59	0.15	0.10
$f_{T_{1b}}$	0.54	0.61	0.14	0.09
$f_{T_{2a}}$	0.50	0.55	0.28	0.19
$f_{T_{2b}}$	0.49	0.56	0.30	0.20

Diagonal entries (M_{ii}) correspond to evaluations of models on their source datasets, while off-diagonal entries reflect performance on other datasets. This table highlights the cross-domain generalization behavior of the IFL models when measured with F1-score.

4.2.2. Benchmark Comparison Matrix

As stated in Section 3.3, the benchmark comparison matrix B mirrors the structure of M , facilitating a direct comparison. Each entry in B represents the performance of the baseline or global model using the same F1-score metric. Results are shown in Table 2.

Table 2. Benchmark Comparison Matrix B for F1-score.

$\mathcal{D}_{1a}$	$\mathcal{D}_{1b}$	$\mathcal{D}_{2a}$	$\mathcal{D}_{2b}$
0.55	0.60	0	0
0.53	0.62	0	0.02
0	0	0.26	0.18
0	0	0.21	0.17

Diagonal entries (B_{ii}) capture the performance of the global model f on each dataset $\mathcal{D}_i$, while off-diagonal entries reflect the baseline models f_i evaluated on other datasets $\mathcal{D}_j$ ($j \neq i$). As the matrix contains results of two different models in each row, the column models. In order to evaluate the IFL models performance, we compare the elements of matrix B with their corresponding entries in matrix M . For any $i, j \in \{1, \dots, n\}$, if $M_{ij} > B_{ij}$, it indicates that the IFL model outperforms its corresponding baseline model on the given dataset.

4.3. Generalization and Consistency Analysis

The generalization and consistency of the models are analyzed using Cross-Domain Generalization Score (CDGS) and Performance Variance (PV) for the Accuracy, as defined in Equations (7) and (8), respectively:

$$\text{CDGS} = [0.2567, 0.2333, 0.4200, 0.4700], \quad \text{PV} = [0.0621, 0.0711, 0.0284, 0.0245].$$

CDGS reflects the average performance of a model on datasets other than its source, while PV measures the variability in performance across datasets. Blank cells indicate entries where indices coincide, as they are not relevant for cross-domain evaluation. These tables allow direct comparison between IFL models (f_{T_i}) and their respective baselines (f_i).

5. Discussion

The results of this study offer valuable insights into the application of the novel IFL method, specifically in the context of cervical cytology image analysis. The methodology presented is versatile and holds significant potential for broader applications across various domains. However, several important considerations arise when interpreting these findings.

In our experiment, 2 out of the 16 models do not improve their specific benchmark models. Specifically, the model $f_{T_{1a}}$ did not surpass the global model f on dataset $\mathcal{D}_{1a}$, and the model $f_{T_{1b}}$ has the same accuracy as the global model f on dataset $\mathcal{D}_{1b}$, even though the rest of the models either matched or exceeded the anticipated performance. This discrepancy highlights the challenges in ensuring consistent improvement across all models and datasets.

The APACC and CRIC Cervix datasets, while extensive and well-annotated, are the only publicly available datasets that meet the requirements for this study, particularly the detailed labeling and high resolution needed for both detection and classification tasks. Splitting these datasets allows for multidimensional analysis and IFL experimentation, but it may also reduce the diversity and variability within each subset. Future work could address this limitation by incorporating additional datasets, either from the domain of cervical cancer or from other medical imaging fields, to test the generalizability of the methodology across entirely different domains. This expansion could further validate the robustness of the proposed approach and explore its applicability to other pathologies. We also have to note that the selected order of the experiment could have varied. We chose the assignments $\{\mathcal{D}_{1a} = \mathcal{D}_1, \mathcal{D}_{1b} = \mathcal{D}_2, \mathcal{D}_{2a} = \mathcal{D}_3, \mathcal{D}_{2b} = \mathcal{D}_4\}$, but there are $4! = 24$ possible permutations, and we can choose $3! = 6$ different permutations with different cycle order. Each permutation could potentially yield specific results due to differences in training order influencing the observed outcomes. Future research could explore the impact of such permutations systematically. This analysis could provide additional insights into the stability and optimality of the training process for transfer learning and multidimensional experiments.

Another critical factor in the methodology is the choice of the parameter λ , which governs the maximum learning rate during transfer learning. As shown in Section Learning Rate Selection for Transfer Learning, λ directly influences the adaptability of the model to new datasets. A higher λ allows for faster adaptation but risks overfitting to the target dataset, potentially causing the model to lose important information from its source domain. Conversely, a lower λ enforces more conservative updates, which can preserve knowledge from the source dataset but might limit the model's ability to effectively learn features specific to the new domain. Optimizing λ for each transfer step is therefore crucial, and future studies could explore adaptive or data-driven methods for determining this

parameter to achieve an optimal balance between source retention and target adaptation. As illustrated in Figure 5, the learning rate cap λ mainly scales the error-dependent update. Preliminary checks showed that values in the range $[10^{-4}, 10^{-2}]$ produced near-identical accuracy, with fluctuations comparable to the variance observed across random seeds. While this indicates that λ has limited influence under the present conditions, we acknowledge its potential relevance in larger or more heterogeneous settings and identify it as a promising direction for future work. In addition, Equation (3) offers a more stable and interpretable alternative to validation loss or gradient-based schedulers, which can fluctuate under class imbalance or noisy labels.

Preliminary checks indicated that variations in interleaving, learning rate schedule (fixed vs. dynamic), dataset order, and freezing policy produced only marginal changes, generally within the variance observed across random seeds. While this suggests that these design factors are not decisive in our current datasets, we acknowledge that they may become more influential in larger or more heterogeneous settings. For this reason, we did not include a full ablation grid in the main results, but we explicitly note this as a limitation and highlight it as a relevant direction for future research.

YOLOv8 was selected as it integrates detection and classification in a single pipeline and represents the state of the art in many recent medical imaging works, making it especially suitable for cytology where abnormal cells must be first localized. Alternative backbones such as ViT, Swin, or ConvNeXt achieve strong results on isolated cell crops, but do not address whole-slide detection. Importantly, our Interleaved Fusion Learning framework is architecture-agnostic: the sequential transfer and dynamic learning rate adaptation can be applied to any network supporting fine-tuning, and could in future be explored with CNN or transformer backbones.

For the evaluation, in Section 3.3, by explicitly incorporating Metric into the definitions of $M_{ij}(\text{Metric})$, $\text{CDGS}_i(\text{Metric})$, and $\text{PV}_i(\text{Metric})$, this framework remains adaptable to different evaluation needs. For example, if fairness is a key concern, one might prioritize metrics such as the F1-score or balanced accuracy, as they are particularly suited to scenarios with class imbalances or where equitable performance across classes is critical. We acknowledge that the absolute accuracies remain modest. However, the focus here is not on achieving state-of-the-art single-dataset results, but on showing consistent improvements over global and local baselines, indicating that residual knowledge can indeed be preserved and transferred. Direct comparison with alternative approaches was intentionally avoided, as their objectives and setups differ substantially. Depending on the chosen metric, these methods could appear either superior or inferior, leading to potentially misleading conclusions. Instead, the most meaningful baselines for assessing IFL are the global and local training strategies across datasets, which directly reflect its goal of preserving and transferring residual knowledge between domains.

The structured methodology relies on curated binary class mappings and dynamic learning rate adjustments. While these choices have been effective for the current datasets, they may require modifications for datasets with more complex label distributions or larger imbalances. Investigating methods to handle multi-class problems or severe label imbalances within this IFL framework could expand its utility and improve its fairness in real-world applications.

In Section 2.3, a new notation for domains in transfer learning was introduced to address inconsistencies identified in previous literature. Although these studies provide initial domain definitions, their practical application often involves inconsistent alterations, resulting in discrepancies between the theoretical framework and its implementation. The revised notation seeks to promote a consistent and coherent use of domain definitions, improving the clarity and reproducibility of transfer learning methodologies.

Although the present study employed binary mappings to simplify the evaluation, it is important to acknowledge that many medical applications, including cervical cytology, are inherently multi-class in nature, with diagnostic categories such as NILM, ASC-US, LSIL, HSIL, and SCC. Extending the IFL framework to these settings is not trivial, as class overlap, hierarchical relationships, and unequal misclassification costs become critical factors. Furthermore, severe class imbalance remains a pervasive challenge in medical data, where minority categories, although clinically decisive, are often underrepresented. While balanced accuracy and F1-score partially mitigate these effects, future work should explore the integration of imbalance-aware strategies such as cost-sensitive losses, resampling techniques, or hierarchical classification schemes within the IFL pipeline. Addressing these two aspects—multi-class complexity and class imbalance—is essential to ensure that performance improvements are equitably distributed across categories.

In conclusion, while the IFL methodology demonstrates clear advantages in improving performance and robustness, the study also underscores the need for broader dataset diversity, careful parameter selection, and enhanced interpretability. These factors should guide future research to maximize the impact and applicability of IFL in medical imaging and beyond.

6. Conclusions

The results, summarized in Section 4, highlight key advantages of the IFL models (f_{T_i}) over both the baseline models (f_i) trained solely on their respective source datasets and the combined dataset model ($f(\cdot)$) trained on the entire dataset $\mathcal{D}$.

This work not only advances the field by proposing a novel IFL methodology but also provides a comprehensive set of practical tools to support future research. These tools include methods for the informed and systematic selection of learning rates, offering a structured optimization process based on our proposed dynamic learning rate for transfer steps. In the method, we provide a novel evaluation stage designed to accommodate diverse performance metrics, such as fairness, robustness, or domain-specific criteria. In addition, we address longstanding inconsistencies in transfer learning notation by presenting an enhanced and standardized framework. Collectively, these contributions establish a solid foundation for further developments in IFL and transfer learning methodologies.

In the studied cervix datasets, the IFL models exhibit consistent improvements in cross-domain generalization, as evidenced by values presented in Table 1. These models generally achieve higher accuracy on datasets other than their sources compared to the baseline models, demonstrating their ability to retain critical knowledge from the source dataset while adapting to new domains. The low PV values further support the robustness of the IFL models, indicating a relatively consistent performance across datasets.

These findings underscore the potential of IFL to enhance the robustness and fairness of DL models in complex multi-domain settings. Future work will focus on expanding the methodology to incorporate additional metrics, such as fairness measures, and exploring its applicability to other medical imaging domains. Furthermore, integrating explainability techniques could provide additional insights into the decision-making processes of the IFL models, facilitating their adoption in clinical practice.

Author Contributions: Conceptualization, C.M., L.B., O.Z. and C.V.; methodology, C.M. and C.V.; software, C.M.; validation, C.M. and C.V.; formal analysis, C.M. and C.V.; investigation, C.M., L.B., O.Z. and C.V.; resources, C.V.; data curation, C.M.; writing—original draft preparation, C.M., L.B., O.Z. and C.V.; writing—review and editing, C.M., L.B., O.Z. and C.V.; visualization, C.M. and C.V.; supervision, C.V.; project administration, C.V.; funding acquisition, C.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work was co-funded by the Spanish Ministry of Science and Innovation (grant number PID2022-138936OB-C32, project COGNISANCE). It was also partially funded and supported by the Spanish Ministry for Digital Transformation and the Civil Service (grant number TSI-100121-2024-35, project BRILLIANT).

Informed Consent Statement: All patients are de-identified.

Data Availability Statement: The code used in this study is available from the corresponding author upon reasonable request. The datasets employed, APPAC and CRIC, are publicly available at <https://cricdatabase.com.br/> (accessed on 1 September 2025) and https://appac.utu.fi/?page_id=42 (accessed on 1 September 2025), respectively.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HPV	Human Papillomavirus
DL	Deep Learning
ML	Machine Learning
TAI	Trustworthy Artificial Intelligence
APACC	Annotated PAp Cell Images and Smear Slices for Cell Classification
CRIC	Center for Recognition and Inspection of Cells
NILM	Negative for Intraepithelial Lesion or Malignancy
ASC-US	Atypical Squamous Cells of Undetermined Significance
LSIL	Low-Grade Squamous Intraepithelial Lesion
ASC-H	Atypical Squamous Cells—Cannot Exclude High-Grade Lesion
HSIL	High-Grade Squamous Intraepithelial Lesion
SCC	Squamous Cell Carcinoma
IFL	Interleaved Fusion Learning
CDGS	Cross-Domain Generalization Score
PV	Performance Variance
YOLOv8	You Only Look Once, Version 8

References

1. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [\[CrossRef\]](#)
2. Sung, H.; Ferlay, J.; Siegel, R.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [\[CrossRef\]](#)
3. World Health Organization. *WHO Guidelines for the Use of Thermal Ablation for Cervical Pre-Cancer Lesions*; World Health Organization: Geneva, Switzerland, 2022.
4. Chaturvedi, A. Epidemiology and clinical aspects of HPV in head and neck cancers. *Head Neck Pathol.* **2012**, *6*, 16–24. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Saslow, D.; Solomon, D.; Lawson, H.; Killackey, M.; Kulasingam, S.; Cain, J.; Garcia, F.; Moriarty, A.; Waxman, A.; Wilbur, D.; et al. American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. *Am. J. Clin. Pathol.* **2012**, *137*, 516–542. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Hu, Z.; Tang, J.; Wang, Z.; Zhang, K.; Zhang, L.; Sun, Q. Deep learning for image-based cancer detection and diagnosis-A survey. *Pattern Recognit.* **2018**, *83*, 134–149. [\[CrossRef\]](#)
7. Cao, L.; Yang, J.; Rong, Z.; Li, L.; Xia, B.; You, C.; Lou, G.; Jiang, L.; Du, C.; Meng, H.; et al. A novel attention-guided convolutional network for the detection of abnormal cervical cells in cervical cancer screening. *Med. Image Anal.* **2021**, *73*, 102197. [\[CrossRef\]](#)
8. González-Nóvoa, J.A.; Busto, L.; Campanioni, S.; Martínez, C.; Fariña, J.; Rodríguez-Andina, J.J.; Juan-Salvadores, P.; Jiménez, V.; Íñiguez, A.; Veiga, C. Advancing cuffless arterial blood pressure estimation: A patient-specific optimized approach reducing computational requirements. *Future Gener. Comput. Syst.* **2025**, *166*, 107689. [\[CrossRef\]](#)
9. Piccialli, F.; Somma, V.D.; Giampaolo, F.; Cuomo, S.; Fortino, G. A survey on deep learning in medicine: Why, how and when? *Inf. Fusion* **2021**, *66*, 111–137. [\[CrossRef\]](#)

10. Kanavati, F.; Hirose, N.; Ishii, T.; Fukuda, A.; Ichihara, S.; Tsuneki, M. A Deep Learning Model for Cervical Cancer Screening on Liquid-Based Cytology Specimens in Whole Slide Images. *Cancers* **2022**, *14*, 1159. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Zhang, L.; Lu, L.; Nogues, I.; Summers, R.; Liu, S.; Yao, J. DeepPap: Deep convolutional networks for cervical cell classification. *IEEE J. Biomed. Health Inform.* **2017**, *21*, 1633–1643. [\[CrossRef\]](#)
12. Han, H.; Li, M.; Wu, X.; Yang, H.; Qiao, J. Filter transfer learning algorithm for nonlinear systems modeling with heterogeneous features. *Expert Syst. Appl.* **2025**, *260*, 125445. [\[CrossRef\]](#)
13. Xu, C.; Li, M.; Li, G.; Zhang, Y.; Sun, C.; Bai, N. Cervical Cell/Clumps Detection in Cytology Images Using Transfer Learning. *Diagnostics* **2022**, *12*, 2477. [\[CrossRef\]](#)
14. Wang, Z.; Voiculescu, I. Dealing with Unreliable Annotations: A Noise-Robust Network for Semantic Segmentation through A Transformer-Improved Encoder and Convolution Decoder. *Appl. Sci.* **2023**, *13*, 7966. [\[CrossRef\]](#)
15. Szymoniak, S.; Depta, F.; Karbowiak, Ł.; Kubanek, M. Trustworthy Artificial Intelligence Methods for Users' Physical and Environmental Security: A Comprehensive Review. *Appl. Sci.* **2023**, *13*, 12068. [\[CrossRef\]](#)
16. Li, F.; Wu, P.; Ong, H.H.; Peterson, J.F.; Wei, W.Q.; Zhao, J. Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. *J. Biomed. Inform.* **2023**, *138*, 104294. [\[CrossRef\]](#)
17. Rajkomar, A.; Hardt, M.; Howell, M.; Corrado, G.; Chin, M. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* **2018**, *169*, 866–872. [\[CrossRef\]](#)
18. Ferrara, E. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci* **2024**, *6*, 3. [\[CrossRef\]](#)
19. Terzi, D.S.; Azginoglu, N. In-Domain Transfer Learning Strategy for Tumor Detection on Brain MRI. *Diagnostics* **2023**, *13*, 2110. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Syu, J.H.; Fojcik, M.; Cupek, R.; Lin, J.C.W. HTTPS: Heterogeneous Transfer learning for split Prediction System evaluated on healthcare data. *Inf. Fusion* **2025**, *113*, 102617. [\[CrossRef\]](#)
21. Kim, H.E.; Cosa-Linan, A.; Santhanam, N.; Jannesari, M.; Maros, M.E.; Ganslandt, T. Transfer learning for medical image classification: A literature review. *BMC Med. Imaging* **2022**, *22*, 69. [\[CrossRef\]](#)
22. Alzubaidi, L.; Al-Amidie, M.; Al-Asadi, A.; Humaidi, A.J.; Al-Shamma, O.; Fadhel, M.A.; Zhang, J.; Santamaría, J.; Duan, Y. Novel Transfer Learning Approach for Medical Imaging with Limited Labeled Data. *Cancers* **2021**, *13*, 1590. [\[CrossRef\]](#)
23. Kupas, D.; Hajdu, A.; Kovacs, I.; Hargitai, Z.; Szombathy, Z.; Harangi, B. Annotated Pap cell images and smear slices for cell classification. *Sci. Data* **2024**, *11*, 743. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Rezende, M.T.; Silva, R.; Bernardo, F.d.O.; Tobias, A.H.G.; Oliveira, P.H.C.; Machado, T.M.; Costa, C.S.; Medeiros, F.N.S.; Ushizima, D.M.; Carneiro, C.M.; et al. Cric searchable image database as a public platform for conventional pap smear cytology data. *Sci. Data* **2021**, *8*, 151. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Plissiti, M.E.; Dimitrakopoulos, P.; Sfikas, G.; Nikou, C.; Krikoni, O.; Charchanti, A. Sipakmed: A New Dataset for Feature and Image Based Classification of Normal and Pathological Cervical Cells in Pap Smear Images. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 3144–3148. [\[CrossRef\]](#)
26. Jantzen, J.; Norup, J.; Dounias, G.; Bjerregaard, B. Pap-smear benchmark data for pattern classification. In Proceedings of the Nature Inspired Smart Information Systems NiSIS, Albufeira, Portugal, 1 January 2005; pp. 1–9.
27. Fang, M.; Liao, B.; Lei, X.; Wu, F.X. A systematic review on deep learning based methods for cervical cell image analysis. *Neurocomputing* **2024**, *610*, 128630. [\[CrossRef\]](#)
28. Hussain, E.; Mahanta, L.B.; Borah, H.; Das, C.R. Liquid based-cytology Pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions. *Data Brief* **2020**, *30*, 105589. [\[CrossRef\]](#)
29. Nayar, R.; Wilbur, D.C. *The Bethesda System for Reporting Cervical Cytology: Definitions, Criteria, and Explanatory Notes*; Springer: Berlin/Heidelberg, Germany, 2015.
30. Ultralytics. YOLO by Ultralytics. Version 8. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 15 August 2025).
31. Yaseen, M. What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector. *arXiv* **2024**, arXiv:2408.15857.
32. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 9. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.