

Article

A Dual-Track Specialist Feature Fusion and Meta-Learning Stacking Ensemble for Cervical Transformation Zone Classification in Colposcopy

Edgar Fabián Rivera-Guzmán ^{1,2,*}, Vladimir Espartaco Robles-Bykbaev ^{1,*} , Bernardo J. Vega-Crespo ³ 
and Veronique Verhoeven ⁴ 

¹ GI-IATa, UNESCO Chair on Support Technologies for Educational Inclusion, Universidad Politécnica Salesiana, Cuenca 010103, Ecuador

² Carrera de Agroindustrias, Universidad Estatal de Bolívar, Guaranda 020150, Ecuador

³ Facultad de Ciencias Médicas, Universidad de Cuenca, Cuenca 010203, Ecuador;
bernardo.vegac@ucuenca.edu.ec

⁴ Family Medicine and Population Health, University of Antwerp, 2610 Antwerp, Belgium;
veronique.verhoeven@uantwerpen.be

* Correspondence: eriverag2@est.up.s.edu.ec (E.F.R.-G.); vrobles@ups.edu.ec (V.E.R.-B.)

Abstract

Cervical cancer remains a major global public health challenge, and the accurate classification of cervical transformation zones (TZs) constitutes a critical step in early detection and clinical decision-making. However, distinguishing between Type 2 and Type 3 transformation zones remains particularly challenging due to their high morphological similarity and the inherent interobserver variability associated with colposcopic assessment. In this study, we propose a novel Dual-Track Specialist Feature Fusion and Meta-Learning Stacking Ensemble architecture for the automated classification of cervical transformation zones using the Intel & MobileODT Cervical Cancer Screening dataset. The proposed framework integrates a global feature extractor based on ResNet50 (Gatekeeper) with a visual specialist based on InceptionResNetV2, trained exclusively on the most diagnostically ambiguous cases (Type 2 and Type 3). The extracted features are fused and processed through a multi-level stacking scheme composed of Multilayer Perceptron (MLP), Support Vector Machine (SVM), Gradient Boosting (GB), XGBoost, and LightGBM classifiers at the base level, followed by an XGBoost meta-learner and a clinically guided probability calibration strategy designed to maximize diagnostic sensitivity. Experimental results demonstrate a peak overall accuracy of 91.22%, substantially outperforming the baseline ResNet50 model (70%). Furthermore, the proposed system achieved Recall values of 0.90, 0.90, and 0.94 for Type 1, Type 2, and Type 3 transformation zones, respectively, highlighting its ability to accurately identify diagnostically challenging cases. Ablation studies, Grad-CAM visualizations, and external-image validation experiments confirm that the proposed architecture improves discrimination between ambiguous categories, learns clinically meaningful representations, and maintains strong generalization capability across heterogeneous scenarios. These findings demonstrate the potential of visual specialization and calibrated meta-learning strategies for the development of artificial intelligence-assisted colposcopic decision-support systems.



Academic Editor: Paolo Bellavista

Received: 5 June 2026

Revised: 14 July 2026

Accepted: 14 July 2026

Published: 16 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: cervical cancer; meta-learning; colposcopy; cervical transformation zones; deep learning; stacking ensemble

1. Introduction

Cervical cancer remains one of the leading causes of cancer-related morbidity and mortality among women worldwide, with a disproportionately high burden in low- and middle-income countries, particularly across Latin America [1]; early detection through organized screening programs represents the most effective strategy for reducing both incidence and mortality rates. In this context, visual inspection with acetic acid (VIA) and colposcopy continues to serve as an essential diagnostic tool, especially in resource-constrained healthcare settings [2]. Accurate identification of the cervical transformation zone (TZ) is a critical step in the colposcopic assessment process, as the majority of cervical neoplasias originate within this anatomical region. According to the International Federation for Cervical Pathology and Colposcopy (IFCPC), transformation zones are classified into three categories: Type 1, entirely visible on the ectocervix; Type 2, partially visible with endocervical extension; and Type 3, not fully visible due to its location within the endocervical canal [3]. While Type 1 transformation zones are generally straightforward to identify, distinguishing between Type 2 and Type 3 remains a clinically challenging task. The boundary between these categories is often characterized by subtle variations in vascularization, acetowhite epithelium, and glandular patterns, leading to considerable interobserver variability and increased diagnostic uncertainty. Consequently, clinically significant false-negative cases may occur, particularly in Type 3 transformation zones, potentially delaying timely intervention and treatment [4].

Artificial Intelligence (AI) has emerged as a promising tool to support the automated analysis of colposcopic images [5]. However, current deep learning approaches often apply generic computer vision strategies without sufficient adaptation to the specific cognitive and visual demands of clinical colposcopy [6]. Consequently, these models tend to prioritize dominant macroscopic shapes while overlooking the subtle, clinically critical microtextures—such as fine mosaic patterns, punctation, atypical vessels, and irregular acetowhite borders—that gynecologists rely on for accurate diagnosis. This lack of fine-grained morphological sensitivity is particularly detrimental in diagnostically challenging cases, such as resolving the boundary ambiguity between Type 2 and Type 3 transformation zones. In clinical practice, accurately distinguishing between these two zones is critical, as a Type 3 classification dictates a direct change in surgical management, often necessitating endocervical sampling. Because standard AI architectures fail to mirror this hierarchical diagnostic reasoning, they struggle to model uncertainty within these ambiguous regions, ultimately limiting their robustness and real-world clinical applicability [7].

To address these limitations, this study proposes a new Dual-Track Specialist Feature Fusion and Meta-Learning Stacking Ensemble framework for automated cervical transformation zone classification. The proposed architecture decomposes the diagnostic task into two complementary pathways: (i) a global classifier responsible for learning the overall distribution of transformation zone categories, and (ii) a specialist branch specifically designed to isolate and model diagnostically ambiguous cases, particularly those located within the decision boundary between Type 2 and Type 3 transformation zones. By explicitly focusing on difficult cases, the framework constructs a refined discriminative subspace capable of improving class separation and reducing inter-class confusion in clinically critical scenarios.

In addition, a clinically oriented probability calibration strategy is introduced to adapt decision thresholds and prioritize Recall for high-risk categories, particularly Type 3 transformation zones. This design choice acknowledges that, in medical decision-support systems, minimizing false negatives is often more important than maximizing overall accuracy. The proposed framework further integrates deep representations extracted from complementary convolutional architectures and combines them through a multi-

level stacking ensemble and a second-level meta-learning classifier, thereby enhancing robustness, generalization capability, and diagnostic reliability.

The main contributions of this work can be summarized as follows:

- **Dual-Track Specialist Ensemble Architecture:** A multi-level stacking framework is proposed that incorporates a dedicated specialist module for the explicit separation of diagnostically ambiguous classes, achieving validation accuracies ranging from 0.89 to 0.912 and outperforming conventional single-network approaches.
- **Mathematical Isolation of Difficult Cases:** A targeted learning mechanism is introduced to explicitly model the uncertainty region between Type 2 and Type 3 transformation zones, significantly improving classification performance under high diagnostic ambiguity.
- **Risk-Aware Decision Threshold Calibration:** A class-specific probabilistic calibration strategy is implemented to prioritize Recall in clinically critical categories, thereby reducing the likelihood of false-negative predictions.

The remainder of this paper is organized as follows. Section 2 reviews the related literature and state-of-the-art approaches. Section 3 describes the proposed methodology, including image preprocessing, specialist feature extraction, and ensemble design. Section 4 presents the experimental results and performance evaluation. Section 5 discusses the clinical and computational implications of the findings. Finally, Section 6 concludes the study and outlines future research directions.

2. Related Work

The application of deep learning to cervical cancer screening and colposcopic image analysis has evolved considerably over the last decade, driven by the need for accurate, low-cost, and scalable diagnostic solutions in resource-constrained environments. Recent advances have enabled the deployment of lightweight convolutional neural networks on embedded and mobile platforms, facilitating point-of-care screening and telemedicine applications [8]. Existing studies can be broadly categorized into three major research directions: (i) image quality assessment and segmentation, (ii) transfer learning using conventional convolutional neural networks, and (iii) advanced ensemble and transformer-based architectures.

2.1. Image Quality Assessment and Segmentation Approaches

Image quality remains a critical factor in automated cervical image analysis, as blur, illumination variability, specular reflections, and anatomical occlusions can significantly affect classification performance. Consequently, several studies have focused on filtering and preprocessing images prior to diagnostic analysis. In [9], an Automated Visual Evaluation (AVE) framework was proposed using a deep ensemble composed of RetinaNet, Deep SVDD, and a custom CNN to distinguish acceptable from inadequate cervical images, achieving an accuracy of 91.6%. Similarly, the framework presented in [10] combined YOLOv5-based region-of-interest detection with U-Net segmentation of the columnar epithelium, reporting a validation accuracy of 0.92.

For fine-grained anatomical segmentation, ref. [11] evaluated Mask R-CNN and MaskX R-CNN architectures across multiple cervical imaging datasets, obtaining Dice coefficients as high as 0.947. In contrast, ref. [12] developed a radiomics-based colposcopic analysis framework combining U-Net segmentation and SVM classification, achieving a classification accuracy of only 58%, highlighting the performance degradation often observed when handcrafted features are applied in real-world clinical environments.

Although these approaches improve image quality and anatomical localization, they frequently rely on computationally intensive preprocessing pipelines and carefully curated

segmentation stages. As evidenced by [12], classification performance remains highly dependent on the robustness of the extracted features.

2.2. Transfer Learning with Conventional CNN Architectures

The adoption of transfer learning has become one of the most widely used strategies for cervical image classification. A comprehensive review presented in [13] documented the impact of public datasets such as Intel & MobileODT on the development of automated cervical screening systems. Among early deep-learning approaches, ref. [14] compared VGG19, ResNet50, and InceptionV3 architectures, reporting the best performance with InceptionV3, which achieved an accuracy of 96.1% in cervix-type classification.

Subsequent studies attempted to improve conventional CNN architectures through architectural optimization. In [15], AlexNet was shown to be insufficient for modeling the complexity of cervical images, achieving only 60.14% accuracy. However, performance improved significantly after incorporating Particle Swarm Optimization (PSO) and dropout regularization, surpassing standard ResNet models by 6.22%. Similarly, ref. [16] compared a transfer-learning-based ResNet50 model, which achieved 82.67% accuracy, against a custom-designed architecture named CervixNET, which reached 99.23%.

Another notable contribution was presented in [17], where a 21-layer CNN was developed for automated cervical screening, achieving 91.37% accuracy and an AUC of 96.0%. Meanwhile, ref. [18] applied deep learning techniques to Pap smear cytology images, reaching 98% accuracy. However, it is important to note that cytological image classification is inherently less challenging than full-field colposcopic image analysis, as cellular images contain fewer anatomical artifacts, illumination variations, and tissue-level complexities.

Collectively, these studies demonstrate that single-network architectures often require substantial architectural modifications, evolutionary optimization techniques, or custom designs to effectively model the complex visual patterns found in colposcopic images.

2.3. Ensemble Learning and Vision Transformer-Based Methods

To overcome the limitations of individual CNN models, recent research has increasingly shifted toward ensemble learning and transformer-based approaches. In [19] a robust screening system was proposed that first extracted regions of interest using a MobileNetV2-YOLOv3 framework and subsequently classified images using EfficientNetB0, achieving 96.84% accuracy.

Similarly, ref. [20] introduced a stacked ensemble architecture incorporating Recursive Feature Elimination (RFE), SMOTE-Tomek resampling, and multiple meta-learners evaluated through explainable artificial intelligence (XAI) techniques. These results demonstrated the potential of ensemble methods to improve robustness and generalization compared to standalone classifiers.

More recently, Vision Transformers (ViTs) have emerged as a promising alternative to convolutional architectures. The framework reported in [17] employed an enhanced Vision Transformer incorporating Shifted Patch Tokenization (SPT) on the MobileODT dataset, achieving approximately 91% generalization performance. Furthermore, ref. [21] combined Vision Transformers with Low-Rank Adaptation (LoRA) to address data scarcity challenges, outperforming traditional CNN architectures and correctly identifying more than 75% of abnormalities in limited-data scenarios.

Although transformer-based models and large-scale ensembles have established new state-of-the-art benchmarks, their improvements are often achieved at the expense of increased computational complexity. Vision Transformers require substantial computational resources and large training datasets, while many ensemble-based approaches aggregate

multiple models without explicitly modeling which feature extractor is most appropriate for diagnostically ambiguous cases.

2.4. Research Gap and Motivation

While the methods discussed in the previous subsections have significantly advanced automated cervical cancer screening, a critical research gap remains regarding the severe morphological ambiguity between consecutive TZ stages, particularly Type 2 and Type 3. Recently, several state-of-the-art (SOTA) architectures have attempted to address this specific bottleneck, yet they present distinct limitations:

Dual-Branch Medical Imaging Models recent SOTAs have utilized dual-branch or multi-stream architectures to simultaneously extract global anatomical context and local textural details [22]. While these networks improve feature richness, they often suffer from feature redundancy and fail to apply targeted discrimination logic when clinical ambiguity arises between Type 2 and Type 3 zones.

Stacked Integrated Cervical Detection Algorithms as Ensemble learning has gained traction to reduce the variance of individual models [23]. However, conventional stacked integrated algorithms typically rely on a simple concatenation of primary CNN outputs fed into a basic classifier. They lack dynamic feature selection, leading to computational overhead and overfitting on minority classes.

Treatment Difficulty Case Networks to solve boundary ambiguity, recent models have employed attention mechanisms or focal loss to penalize hard-to-classify examples [24]. While mathematically sound, these networks lack clinical intuition and struggle with out-of-distribution artifacts, such as speculum reflections [25].

To objectively highlight the architectural and performance advantages of our proposed method compared to these similar SOTAs, Table 1 presents a systematic comparison. Unlike existing approaches, our architecture explicitly introduces a dedicated “Specialist Feature” gatekeeper (Type 2 vs. Type 3) embedded within a Level 2 Meta-Learner, mimicking the differential diagnostic process of a human specialist.

Despite significant advances in automated cervical image analysis, an important gap remains in the current literature. Existing methods typically achieve high performance either through extensive preprocessing and manual region-of-interest extraction or through computationally demanding ensembles and transformer architectures that may hinder deployment in low-resource clinical environments. Moreover, current approaches rarely incorporate explicit mechanisms for modeling diagnostic ambiguity, particularly in the clinically challenging boundary between Type 2 and Type 3 transformation zones. Most systems treat all classes uniformly, despite the fact that Type 2 and Type 3 cases exhibit substantial morphological overlap and represent the primary source of diagnostic uncertainty in colposcopic practice.

To address these limitations, this study proposes a Dual-Track Specialist Feature Fusion and Meta-Learning Stacking Ensemble framework. Unlike conventional ensembles that simply aggregate predictions, the proposed architecture introduces a dedicated specialist pathway trained specifically on diagnostically ambiguous cases. By combining global anatomical representations with specialized discriminative features and integrating them through a meta-learning strategy, the framework is capable of dynamically refining decision boundaries without requiring explicit segmentation or manually defined regions of interest. Experimental results demonstrate that this strategy achieves a competitive accuracy of 91.5% on the challenging Intel & MobileODT dataset while maintaining a favorable balance between diagnostic performance, interpretability, and computational efficiency.

Table 1. Conceptual and architectural comparison between existing methodological paradigms and the proposed Dual-Track Stacking framework.

Approach/Reference	Core Architecture	Handling of Difficult Cases (Type 2 vs. 3)	Architectural Advantages of Our Proposed Model
Standard CNNs/ViTs (Sections 2.2 and 2.3)	Single-branch networks (e.g., ResNet, ViT)	Handled purely through global loss functions.	Our multi-stage stacking reduces the high variance and overfitting typical of single-network approaches.
Dual-Branch Models [22]	Multi-stream feature extraction	Implicit feature fusion without targeted logic.	We implement a mathematically rigorous fusion of 3590-dimensional fused feature representation, capturing both deep residual and inception-based hierarchies.
Traditional Stacking [23]	Concatenation of CNN probabilities	Relies on base models to solve ambiguity.	Our Level 1 Meta-Learner uses Best Subset Selection and Per-Class Threshold Optimization to actively calibrate boundary predictions.
Proposed Method	Dual-Track Fusion + Two-Level Stacking	Handled probabilistically through Level-1 Stacking and Dynamic Threshold Optimization.	Integrates an XGBoost Meta-Learner calibrated by Dynamic Threshold Optimization, which explicitly adjusts decision boundaries for ambiguous Type 2 and Type 3 cases, maximizing accuracy to 91.22%.

3. Materials and Methods

This study proposes a clinically oriented deep learning framework for the automated classification of TZs from colposcopic images. The methodological design was developed under two complementary objectives: (i) maximizing discriminative performance in morphologically ambiguous cases, particularly between Type 2 and Type 3 transformation zones, and (ii) ensuring computational feasibility for deployment in resource-constrained clinical environments.

To address these challenges, a heterogeneous multi-stage architecture based on specialist feature fusion and meta-learning stacking was designed. The framework integrates high-capacity convolutional neural networks as feature extractors, a dedicated specialist branch focused on difficult-to-discriminate classes, and a meta-classifier responsible for final probabilistic inference. In addition, class-sensitive threshold calibration strategies and post-training optimization mechanisms were incorporated to improve Recall in clinically high-risk categories while simultaneously maintaining computational efficiency.

Figure 1 shows the proposed computational classification framework, structured as a four-phase sequential pipeline designed to process and classify colposcopy images. Initially, in Phase 0 (Data Preparation), the raw images enter a visual preprocessing module where they are spatially standardized. Subsequently, to mitigate the inherent bias toward the majority anatomical classes in the dataset, the resulting tensors undergo synthetic balancing techniques via data augmentation and SMOTE. These balanced batches feed directly into Phase 1 (Dual-Track Feature Extraction), the deep core of the architecture. This parallel approach deploys an InceptionResNetV2 model (acting as the Specialist) to capture morphological features and texture variations at multiple scales (crucial for glandular margins), operating in synergy with a ResNet50 architecture (acting as the Gatekeeper) dedicated to extracting the global anatomical context of TZ.

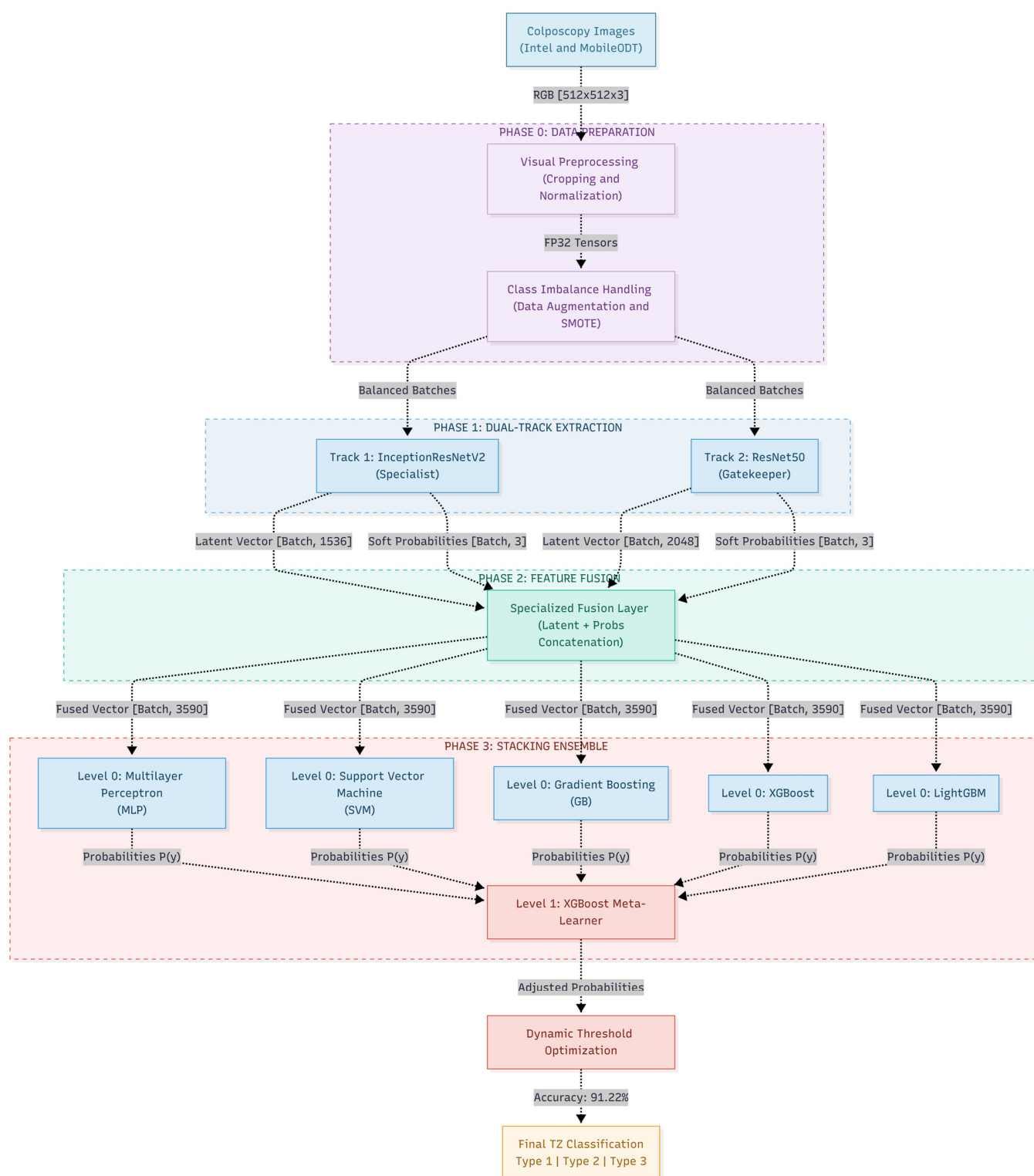


Figure 1. Overall architecture of the proposed Dual-Track and Meta-Learning computational pipeline. The system integrates an initial phase of visual standardization and class imbalance mitigation (SMOTE), followed by parallel feature extraction using pre-trained networks (InceptionResNetV2 for multiscale textures and ResNet50 for the global anatomical context). The extracted latent tensors converge in a specialized fusion layer, whose representations feed a Stacking Ensemble. In the final phase, the base classifiers output conditional probabilities that are processed by a meta-classifier (XGBoost) to robustly predict the anatomical type of the Cervical Transformation Zone (Type 1, Type 2, Type 3).

To capitalize on the orthogonal information from both branches, Phase 2 (Feature Fusion) implements a specialized concatenation layer. At this stage, the latent vectors extracted by ResNet50 and InceptionResNetV2 (of 2048 and 1536 dimensions, respectively) are mathematically integrated, yielding 3584 deep spatial features. To further enrich this unified representation, the 6 soft-probability outputs (3 class probabilities from each independent track) are explicitly appended, totaling exactly 3590 features. Finally, this high-dimensional fused vector is processed in Phase 3 using a Meta-Learning-based Stacking Ensemble. At Level 0, five base classifiers—Multilayer Perceptron (MLP), Support Vector Machine (SVM), Gradient Boosting (GB), XGBoost, and LightGBM—evaluate the representations to generate a map of conditional probabilities. These predictions act as meta-features for the Level 1 ensemble model (XGBoost), which, through a robust inference scheme, consolidates the iterative decisions to issue the final anatomical classification of the cervix (Type 1, Type 2, or Type 3).

3.1. Dataset and Preprocessing

To provide deeper context regarding the data source, the images in this dataset were captured using MobileODT's smartphone-based digital colposcopes (such as the Enhanced Visual Assessment or EVA System). These hand-held, mobile devices are specifically designed to facilitate point-of-care cervical cancer screening, particularly in low- and middle-income countries (LMICs). Because these images are acquired in real-world, decentralized clinical settings by healthcare providers with varying levels of expertise, the dataset inherently exhibits significant heterogeneity. The images frequently contain artifacts such as specular reflections (glare), varying illumination, blurriness, and off-center physical positioning. While these real-world conditions make the classification task highly challenging, they accurately represent the true clinical environment, thereby justifying the need for a highly robust, dual-track feature extraction architecture.

The experimental analysis was primarily conducted using the publicly available Intel & MobileODT Cervical Cancer Screening dataset, available through the Kaggle platform. This dataset contains colposcopic images acquired under heterogeneous clinical conditions, including variations in illumination, speculum positioning, cervical visibility, image sharpness, and acquisition angle. Such variability makes the dataset particularly suitable for evaluating the robustness and generalization capability of artificial intelligence systems intended for real-world clinical deployment.

To further assess the model's ability to generalize beyond the training domain, an additional exploratory validation was performed using external colposcopic images that were not part of the Intel & MobileODT dataset and were never observed during training, validation, or hyperparameter optimization. These external images exhibited substantial differences in acquisition conditions, including variations in image resolution, illumination, compression artifacts, videoconferencing-based image capture, and the presence of graphical overlays. As detailed in Section 4.9 (External Validation on Unseen Clinical Images), the proposed framework maintained consistent attention over clinically relevant cervical structures and produced predictions aligned with the expected clinical assessment. This complementary evaluation provides preliminary evidence that the learned representations are transferable beyond the original training dataset and supports the potential applicability of the framework in real-world clinical and telemedicine environments.

The dataset is organized into three diagnostic categories corresponding to cervical transformation zone types:

- Type 1 Transformation Zone (TZ1)
- Type 2 Transformation Zone (TZ2)
- Type 3 Transformation Zone (TZ3)

These categories represent increasing levels of diagnostic complexity. Type 1 images generally exhibit a fully visible squamocolumnar junction and well-defined anatomical boundaries, whereas Type 2 and especially Type 3 images contain partially or completely non-visible endocervical regions, resulting in substantial inter-class overlap and diagnostic ambiguity.

The experiments were conducted using the Intel & MobileODT Cervical Cancer Screening dataset. Initially, the raw dataset comprised a total of 6788 images (1440 for Type 1; 2926 for Type 2; and 2422 for Type 3). During the preliminary quality control process, 2473 images were excluded due to severe blurring, corruption, or missing diagnostic labels—a well-documented challenge within this specific Kaggle dataset. Consequently, the final curated dataset consisted of 4315 valid images, distributed as follows: 914 images for Type 1, 1945 images for Type 2, and 1456 images for Type 3. To ensure robust evaluation, a stratified 70/15/15 split was applied, partitioning the data into training (3020 images), validation (647 images), and testing (648 images) sets. Crucially, because the original public dataset does not provide unique patient identifiers, it was not possible to perform an explicit patient-level split. To prevent data leakage, the data partitioning was strictly executed prior to any synthetic data augmentation or SMOTE resampling procedures applied to the training set.

All images were resized to a standardized spatial resolution of 512×512 pixels prior to model training. This resolution was selected as a compromise between preserving clinically relevant microtextures and maintaining computational efficiency. Higher resolutions facilitate the preservation of acetowhite epithelial borders, mosaic vascular patterns, and glandular transitions that are often degraded in low-resolution pipelines.

The preprocessing workflow also incorporated robustness mechanisms for handling corrupted or partially truncated image files through the Python Imaging Library (PIL, version 9.5.0), thereby preventing interruptions during iterative training. Subsequently, images were normalized and partitioned into training and validation subsets using stratified directory-based generators.

To further mitigate the effects of class imbalance during optimization, class-weighting strategies were incorporated during training. Weight coefficients were dynamically computed according to the class distribution and integrated into the loss optimization process, forcing the network to assign greater importance to underrepresented and clinically challenging categories.

The weighted loss function can be expressed as:

$$L_{weighted} = -\sum_{i=1}^N w_{c_i} \log(p_{c_i}) \quad (1)$$

where w_{c_i} denotes the weight associated with class c_i , and p_{c_i} represents the predicted posterior probability of the corresponding sample. This strategy reduces the dominance of majority classes and improves sensitivity toward high-risk transformation zones [26].

Dataset generators were configured using mini-batch iterative loading to optimize GPU memory consumption and accelerate training convergence. The resulting preprocessing pipeline produced a robust and clinically realistic input representation suitable for subsequent stages of deep feature extraction and specialist ensemble learning.

3.2. Gatekeeper–Specialist Feature Extraction Framework

To address the substantial morphological overlap between Type 2 and Type 3 cervical transformation zones, a hierarchical dual-path architecture termed the Gatekeeper–Specialist Feature Extraction Framework was developed. Unlike conventional approaches based on a single convolutional feature extractor, the proposed strategy decomposes the

diagnostic problem into two complementary pathways: (i) a generalist model responsible for learning the global distribution of the three transformation zone classes, and (ii) a specialist model trained exclusively on clinically ambiguous categories.

This formulation is inspired by the clinical principle of progressive differential diagnosis, whereby a broad categorization is first performed before focusing on diagnostically challenging cases. From a deep-learning perspective, this strategy reduces the dominance of trivial visual patterns while increasing sensitivity to discriminative microstructures that are frequently overlooked by monolithic architectures.

3.2.1. ResNet50-Based Gatekeeper Architecture

The first pathway of the proposed framework, referred to as the Gatekeeper, was implemented using the ResNet50 architecture pre-trained on the ImageNet dataset. This model was selected due to its convergence stability, hierarchical feature extraction capability, and computational efficiency in multiclass medical image classification tasks. The primary role of the Gatekeeper is to learn robust global representations of the three cervical transformation zone (TZ) categories (Type 1, Type 2, and Type 3). Conceptually, this branch acts as a primary filtering mechanism capable of discriminating macroscopic anatomical patterns such as:

- Visibility of the squamocolumnar junction;
- Epithelial distribution;
- Global illumination patterns;
- Overall cervical geometry.

The architecture employed pre-trained weights with progressive fine-tuning of the upper layers, enabling the adaptation of natural image representations learned from ImageNet to the colposcopic imaging domain. Input images were resized to a standardized resolution of 512×512 pixels and subsequently normalized using the ResNetV2 preprocess_input() function. The convolutional backbone generated a deep feature vector responsible for modeling the global spatial distribution of cervical images.

Mathematically, the Gatekeeper representation can be expressed as:

$$F_{gate} = \phi_{ResNet50}(X) \quad (2)$$

where X denotes the input colposcopic image and F_{gate} corresponds to the deep embedding generated by the ResNet50 architecture [27].

Unlike conventional classification-oriented approaches, the Gatekeeper output in this study was not used solely for direct prediction. Instead, it served as a high-level semantic descriptor that was subsequently integrated into the specialist feature fusion mechanism.

3.2.2. InceptionResNetV2-Based Specialist Architecture

Although the Gatekeeper achieved adequate separation of global visual patterns, preliminary experiments revealed that the primary source of clinical error was concentrated along the diagnostic boundary between TZ Type 2 and TZ Type 3. These categories exhibit substantial visual overlap, including partially obscured regions, diffuse glandular transitions, and heterogeneous acetowhite structures.

To explicitly model this region of uncertainty, a second pathway named the Specialist was implemented using the InceptionResNetV2 architecture. This model was trained exclusively on images belonging to Type 2 and Type 3 transformation zones, forcing the network to focus on highly discriminative fine-grained patterns. The selection of InceptionResNetV2 was motivated by three key properties:

- Multi-scale representation capability, as Inception modules simultaneously capture local and global anatomical structures through convolutional filters of different receptive field sizes;
- Stable residual depth, where residual connections facilitate the training of very deep networks without gradient degradation;
- Sensitivity to medical microtextures, enabling improved preservation of subtle features such as fine vascularization, epithelial mosaic patterns, and diffuse endocervical boundaries.

Formally, the specialist representation is defined as:

$$F_{spec} = \phi_{IncResNetV2}(X_{2,3}) \quad (3)$$

where $X_{2,3}$ represents the restricted subset of images belonging to Type 2 and Type 3 transformation zones [27].

Unlike the Gatekeeper, whose objective is to model the overall class distribution, the Specialist learns a focused discriminative subspace centered on difficult cases, thereby functioning as a diagnostic refinement mechanism.

3.2.3. Dual Feature Fusion

Subsequently, the deep representations generated by both pathways were concatenated to construct an enriched multimodal embedding, as defined in Equation (4) [28].

$$F_{fusion} = [F_{gate} \oplus F_{spec}] \quad (4)$$

where $\oplus$ denotes the concatenation operation.

The resulting vector contained a total of 3590 features, simultaneously integrating global anatomical information extracted by the Gatekeeper and specialized discriminative microfeatures learned by the Specialist branch. Specifically, this fusion consists of 3584 deep spatial features (2048 from the ResNet50 baseline and 1536 from the InceptionResNetV2 baseline) explicitly concatenated with 6 soft-probability outputs (3 class probabilities from each independent track). The choice of this dimensionality was not arbitrary but rather the outcome of iterative experiments evaluating embedding stability and discriminative capacity. Preliminary evaluations demonstrated that lower-dimensional embeddings resulted in the loss of clinically relevant microtextures, whereas excessively large representations increased the risk of overfitting and semantic redundancy.

The fused 3590-dimensional embedding preserved the complementarity between both branches without collapsing informational diversity. Experimentally, this configuration consistently improved diagnostic sensitivity for ambiguous classes when compared with single-extractor architectures. Consequently, the fused representation was employed as the input to the stacking ensemble-based meta-learning framework described in the following subsection.

It is important to clarify the operational flow of the dual-track architecture during the inference phase. Because the true class of a new image is inherently unknown during prediction, there is no conditional routing or pre-classification step. Instead, every incoming image is processed simultaneously by both the Gatekeeper (ResNet50) and the Specialist (InceptionResNetV2). The resulting embeddings and their respective initial probability predictions are always concatenated to form the 3590-dimensional fused vector. It is the downstream Level-0 and Level-1 Stacking Ensemble that dynamically learns to weigh these features. If an image clearly belongs to Type 1, the meta-learner implicitly learns to prioritize the global features extracted by the Gatekeeper, assigning lower importance to the Specialist's features. Conversely, for ambiguous Type 2 or Type 3 cases, the ensem-

ble leverages the highly discriminative features provided by the Specialist to make the final decision.

3.3. Level-1 Repeated Stratified Stacking Ensemble

Following the generation of the fused embedding composed of 3590 deep features, a second stage was implemented to maximize the system's generalization capability and statistical robustness through a Level-1 Repeated Stratified Stacking Ensemble framework. This stage was specifically designed to address two common limitations of moderately sized medical datasets: (i) high statistical variance during validation and (ii) excessive sensitivity to fluctuations in class distribution.

Unlike conventional approaches that rely on a single classifier, the proposed framework combines multiple heterogeneous models trained on different stratified partitions of the deep feature space. This strategy exploits algorithmic complementarity while simultaneously reducing overfitting risk and dependence on the inductive biases of any individual classifier.

3.3.1. Repeated Stratified Cross-Validation

Considering the relatively limited size of the colposcopic dataset and the presence of clinically ambiguous classes, a validation strategy based on Repeated Stratified K-Fold Cross-Validation was adopted. The experimental protocol employed $3 \text{ repetitions} \times 5 \text{ folds} = 15$ stratified partitions resulting in a total of 15 independent training-validation cycles.

The selection of this configuration was motivated by several statistical and clinical considerations. In small or moderately balanced medical datasets, a single train-validation split may produce highly unstable performance estimates due to random fluctuations in class composition. This phenomenon becomes particularly pronounced for minority or difficult categories such as TZ Type 3.

Repeated cross-validation reduces the variance associated with a single dataset partition by generating multiple stratified reorganizations of the samples. Formally, the final performance estimate can be expressed as:

$$\hat{M} = \frac{1}{R \times K} \sum_{r=1}^R \sum_{k=1}^K M_{r,k} \quad (5)$$

where R represents the number of repetitions, K the number of folds, and $M_{r,k}$ the evaluation metric obtained during each iteration [29].

This procedure produces more robust performance estimates that are less sensitive to random sample selection, thereby improving the experimental reliability of the proposed framework. Furthermore, stratification ensures that each fold preserves the original class distribution, preventing biased partitions that could negatively affect sensitivity toward minority categories.

3.3.2. Heterogeneous Pool of Algorithmic Diversity

For each stratified partition, a heterogeneous collection of Level-1 classifiers was trained, forming an algorithmic diversity pool designed to capture different geometric structures within the fused embedding space.

The ensemble consisted of the following models:

- Keras/TensorFlow Multi-Layer Perceptron (MLP) (TensorFlow version 2.16.1, Keras version 3.0)
- Support Vector Machine (SVM) (scikit-learn version 1.4.0)
- Gradient Boosting (GB) (scikit-learn version 1.4.0)

- Extreme Gradient Boosting (XGBoost) (XGBoost version 2.0.3)
- Light Gradient Boosting Machine (LightGBM) (LightGBM version 4.3.0)

The rationale behind this selection is that each architecture incorporates different learning mechanisms and inductive biases, enabling the modeling of complementary patterns within the deep feature space.

- PyTorch Multi-Layer Perceptron (MLP) (version 2.9.0)

The first classifier was a multilayer neural network implemented in PyTorch. The model was designed to capture complex nonlinear relationships within the 3590-dimensional fused embedding. Its architecture included fully connected dense layers, ReLU activation functions, Dropout regularization, and Adam optimization.

Its primary advantage lies in its ability to model distributed interactions among deep features extracted simultaneously from both the Gatekeeper and Specialist branches.

Formally, the MLP can be represented as:

$$y = f(W_2\sigma(W_1x + b_1) + b_2) \quad (6)$$

where x denotes the fused feature vector and σ represents the nonlinear activation function [30].

- Support Vector Machine

The second classifier corresponded to a Support Vector Machine (SVM) employing a nonlinear kernel. The model was incorporated due to its ability to construct robust decision boundaries in high-dimensional feature spaces.

In medical applications involving limited sample sizes, SVMs often exhibit strong generalization capabilities, particularly in regions where class boundaries are highly overlapping.

The discriminant function can be expressed as:

$$f(x) = \sum_{i=1}^N \alpha_i y_i K(x_i, x) + b \quad (7)$$

where $K(\cdot)$ denotes the kernel function used to project samples into a higher-dimensional separable space [31].

- Gradient Boosting (GB)

The third Level-0 classifier was a Gradient Boosting (GB) model based on an ensemble of sequentially optimized decision trees. Unlike bagging-based methods, Gradient Boosting constructs each new tree to minimize the residual errors produced by the previous ensemble, thereby progressively improving predictive performance. This approach is particularly effective for modeling nonlinear relationships and complex interactions among high-dimensional deep features while maintaining strong generalization on moderately sized medical datasets.

Formally, the additive boosting model can be expressed as

$$F_M(x) = \sum_{m=1}^M \gamma_m h_m(x) \quad (8)$$

where $h_m(x)$ denotes the weak learner added at iteration m , and γ_m represents its corresponding contribution to the final prediction. During training, each learner is optimized to minimize the gradient of the loss function with respect to the current ensemble prediction, allowing the model to iteratively correct previous classification errors.

Within the proposed framework, Gradient Boosting contributes additional algorithmic diversity by learning decision boundaries that differ from those captured by neural networks and kernel-based classifiers. This complementary behavior increases ensemble robustness and improves the representation of difficult decision regions before the second-level meta-learning stage.

- Extreme Gradient Boosting (XGBoost)

The third component of the ensemble was XGBoost, a sequential gradient-boosted decision tree algorithm. This model was selected because of its ability to model complex nonlinear interactions and hierarchical relationships among deep features.

Unlike purely neural approaches, XGBoost exhibits high tolerance to noise and can efficiently identify highly discriminative feature subsets within large embedding spaces.

The optimization objective is defined as:

$$\mathcal{L}^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (9)$$

where Ω represents the structural regularization term [32].

- Light Gradient Boosting Machine (LightGBM)

Finally, LightGBM was incorporated as a gradient boosting classifier optimized for computational efficiency and rapid training in high-dimensional feature spaces. Its leaf-wise tree growth strategy enables the construction of deep discriminative partitions while maintaining a lower computational cost and strong predictive performance.

The simultaneous inclusion of XGBoost and LightGBM increased ensemble diversity because of their distinct tree construction and optimization strategies.

3.3.3. Imbalance Mitigation via Sample Weights

Since the dataset exhibits a moderate degree of class imbalance, a sample weighting mechanism was incorporated during the training of the Level-1 classifiers. Without weighting, classifiers tend to minimize the overall loss by favoring majority classes, thereby reducing sensitivity toward clinically important minority categories. This issue is particularly problematic in medical classification, where false negatives often carry substantially greater clinical consequences than false positives.

To counteract this effect, class weights were computed as inversely proportional to class frequency:

$$w_c = \frac{N}{C \cdot n_c} \quad (10)$$

where N denotes the total number of samples, C represents the number of classes, and n_c corresponds to the number of samples belonging to class c [33].

These weights were directly integrated into the optimization processes of compatible classifiers, forcing the ensemble to penalize errors on underrepresented classes more heavily. This strategy contributes to improving the Recall of complex transformation zone categories while reducing the likelihood of overlooking cases potentially associated with high-risk lesions.

3.4. Meta-Learner and Clinical Threshold Calibration

After completing the deep feature extraction, heterogeneous stacking, and tabular binary specialization stages, the system produced a rich set of probabilistic meta-features originating from multiple classifiers and diagnostic pathways. Although this representation substantially enhanced the discriminative capability of the framework, a fundamental

challenge remained: transforming the multiple probabilistic outputs of the ensemble into a robust and clinically calibrated final decision [34].

To address this issue, a final stage was implemented based on a Level-1 Meta-Learner complemented by a clinical probability calibration mechanism, explicitly designed to shift the decision boundary toward regions with greater diagnostic sensitivity. The primary motivation behind this stage is that, in medical applications, the optimal prediction does not necessarily correspond to the class with the highest raw probability. In particular, minimizing false negatives in high-risk categories is more important than maximizing overall accuracy. Under this premise, the system was optimized not solely for statistical classification but also for clinically oriented Recall prioritization.

3.4.1. Level-1 Meta-Learner Based on XGBoost

The stacking process generated a set of probability estimates from all base classifiers trained across the stratified partitions. These probabilities were organized into a meta-feature matrix

$$Z \in \mathbb{R}^{N \times m} \quad (11)$$

where N represents the total number of samples and M corresponds to the number of aggregated probabilistic outputs generated by the ensemble [35]. A meta-classifier based on eXtreme Gradient Boosting (XGBoost) was trained using this representation. The choice of XGBoost as the meta-learner was motivated by several methodological considerations:

- Non-linear combination of experts
Unlike simple linear models, XGBoost effectively captures complex, non-linear interactions between the probabilistic outputs of the base classifiers.
- Dynamic reliability weighting
The tree-based architecture inherently partitions the decision space, learning exactly under which conditions (e.g., specific probability ranges) certain base classifiers are more trustworthy than others.
- Robustness against overfitting
The built-in L1 and L2 regularization mechanisms within the XGBoost framework ensure that the meta-learner generalizes well, even when combining highly correlated predictions from Level-0 models.

The final multi-class probability for class c is computed using a softmax objective function over the raw tree scores, expressed as:

$$P(y = c \mid Z) = \frac{e^{\beta_c^T Z}}{\sum_{j=1}^C e^{\beta_j^T Z}} \quad (12)$$

where Z denotes the meta-feature vector, β_c represents the learned parameter vector for class c , and C is the total number of classes [36].

The meta-learner acts as a probabilistic consensus mechanism capable of integrating the global patterns captured by the Gatekeeper, the specialized refinements provided by the Specialist, the statistical behavior of heterogeneous classifiers, and inter-model uncertainty signals. As a result, the system transforms multiple partially inconsistent predictions into a more stable and coherent clinical inference.

3.4.2. Clinical Probability Calibration and Argmax Adjustment

In conventional multiclass systems, the final prediction is obtained using the operator:

$$\hat{y} = \arg \max_c P(y = c) \quad (13)$$

that is, selecting the class with the highest posterior probability.

However, in imbalanced medical scenarios, this strategy presents important limitations because the Argmax criterion tends to favor majority classes or visually less ambiguous categories, even when probability differences between classes are minimal. Experimentally, several clinically critical Type 3 cases produced highly competitive probability distributions, for example:

$$P(T2) \approx 0.47, P(T3) \approx 0.45 \quad (14)$$

Under a traditional Argmax scheme, these samples would be classified as Type 2, increasing the risk of false negatives associated with partially hidden lesions.

To overcome this limitation, a probability scaling calibration mechanism was implemented to adaptively shift the ensemble decision boundary toward regions of higher clinical sensitivity. The strategy consisted of introducing class-specific calibration factors:

$$P'_c = \alpha_c \cdot P_c \quad (15)$$

where P'_c denotes the original probability of class P_c , and α_c corresponds to an experimentally optimized calibration scalar [37]. The final decision rule was subsequently redefined as:

$$\hat{y} = \arg \max_c P'_c \quad (16)$$

This procedure geometrically modifies the classifier's decision boundaries without requiring complete retraining of the model. Consequently, calibration allows the deliberate enhancement of sensitivity toward high-risk categories—particularly TZ Type 3—thereby reducing the probability of overlooking potentially malignant cases.

The application of this strategy shifts the decision hyperplane toward more conservative detection regions, prioritizing clinical Recall over raw accuracy. Experimentally, small variations in the calibration factors produced significant improvements in diagnostic sensitivity without substantial losses in overall specificity, demonstrating that the probability distributions generated by the ensemble contain latent information that cannot be fully exploited through the conventional Argmax criterion alone.

The calibrated Meta-Learner constitutes the final inference stage of the proposed framework, sequentially integrating dual Gatekeeper–Specialist deep feature extraction, repeated heterogeneous stacking, tabular binary refinement, and clinical probability calibration. This hierarchical architecture was designed to explicitly decouple global visual learning from probabilistic reasoning and diagnostic uncertainty management, enabling a more robust modeling of ambiguous cervical classes.

As a result, the system achieves consistent improvements in predictive stability and clinical sensitivity, particularly in distinguishing between TZ Type 2 and TZ Type 3. Experimentally, the framework achieved validation performances exceeding 91%, demonstrating that the combination of deep specialization and adaptive decision calibration can simultaneously optimize discriminative capability and reduce false negatives in clinically critical scenarios.

3.5. Implementation and Reproducibility Details

To ensure the full reproducibility of the proposed framework, the computational experiments were conducted on a Kaggle cloud environment utilizing dual NVIDIA Tesla T4 GPUs (16 GB VRAM each). The deep learning models and data pipelines were implemented using TensorFlow/Keras, while the stacking ensemble classifiers were developed using Scikit-Learn, XGBoost, and LightGBM. A fixed random seed was strictly applied across all modules to guarantee deterministic behavior. The complete set of hyperparameters for

data augmentation, model training, and the final meta-learner configuration is summarized in Table 2.

Table 2. Experimental settings and hyperparameters for reproducibility.

Category	Hyperparameter/Setting	Value PDF
Global Setup	Random Seed	42
	Input Image Size	512 × 512 pixels
	Batch Size	16
Data Augmentation	Rotation Range	180°
	Width/Height Shift	±10% (0.1)
	Shear/Zoom Range	0.1/0.15
	Flips	Horizontal and Vertical
	Brightness Range	[0.8, 1.2]
	Fill Mode	Reflect
Neural Network (Base)	Optimizer	Adam
	Initial Learning Rate	1e-4
	Maximum Epochs	100
	Early Stopping Patience	15 epochs (monitor: val_accuracy)
	ReduceLROnPlateau	factor = 0.5, patience = 5
Meta-Learner (XGBoost)	Number of Estimators	800
	Max Depth	4
	Learning Rate	0.03
	Subsample	0.8

4. Results

This section presents the experimental results obtained by the proposed architecture. Performance metrics, comparative analyses, graphical representations, and statistical evaluations are progressively introduced to support the technical and clinical validity of the developed framework. Furthermore, the individual contribution of each component is analyzed through ablation studies, multicriteria evaluation, and discriminative capability assessments across the different cervical transformation zone categories.

4.1. Baseline vs. Proposed Architecture

To evaluate the overall effectiveness of the proposed framework, a direct comparison was conducted between a conventional baseline model based solely on ResNet50 and the complete Dual-Track Specialist Feature Fusion and Meta-Learning Stacking Ensemble architecture. This evaluation enabled the quantification of the impact of the proposed specialization strategy on multiclass cervical transformation zone classification performance.

The baseline model consisted of a standard multiclass ResNet50 architecture trained directly on the three clinical categories (Type 1, Type 2, and Type 3) using conventional transfer learning. This approach represents one of the most widely adopted strategies in CNN-based medical image classification systems, where models pretrained on ImageNet are adapted to the clinical domain through fine-tuning.

In contrast, the proposed framework incorporates several mechanisms specifically designed to address the limitations commonly observed in traditional monolithic architectures, including:

- Dual Gatekeeper–Specialist feature extraction;
- Specialized learning on diagnostically ambiguous cases;
- Multi-level heterogeneous stacking;
- Binary tabular refinement;
- Clinical probability calibration.

Table 3 summarizes the comparative results obtained on the validation dataset.

Table 3. Comparison between baseline models and the proposed architecture.

Model	Accuracy
ResNet50 Baseline	0.70
InceptionResNetV2	0.81
Proposed Dual-Track Specialist Framework	0.9122

The results demonstrate a substantial performance improvement achieved by the proposed architecture. While the baseline ResNet50 model attained approximately 70% accuracy, the complete framework achieved an accuracy exceeding 91%, corresponding to an absolute improvement of more than 21 percentage points.

From a methodological perspective, this difference confirms that the limitations of conventional approaches are not solely associated with the representational capacity of the convolutional backbone. Rather, they stem from the absence of explicit mechanisms capable of modeling regions characterized by high diagnostic ambiguity.

In particular, the baseline model exhibited significant difficulties in correctly separating Type 2 and Type 3 cases due to the considerable morphological overlap between these categories. This behavior is consistent with previously reported limitations of generic CNN architectures trained through direct transfer learning, where dominant patterns associated with majority classes tend to shift decision boundaries toward statistically prevalent regions of the feature space.

The proposed architecture addresses this limitation by decomposing the diagnostic process into multiple specialized stages. This design enables the simultaneous capture of global anatomical structures and fine-grained local textures, explicit modeling of diagnostic uncertainty regions, reduction in statistical variance through repeated stratified validation, and incorporation of clinically oriented sensitivity criteria during the final probability calibration stage.

Moreover, the observed performance gain cannot be attributed solely to increased computational complexity. Instead, it arises from the synergistic interaction between visual specialization, probabilistic meta-learning, and clinically guided optimization of decision thresholds.

The performance achieved by the proposed framework is particularly relevant because the system was intentionally optimized to prioritize diagnostic sensitivity in clinically high-risk categories while simultaneously reducing the probability of false negatives in partially obscured or morphologically ambiguous cervical regions. This characteristic is especially important in cervical cancer screening applications, where missed detections may delay clinical intervention and negatively impact patient outcomes.

4.2. Clinical Performance Metrics

Although Accuracy provides a global measure of classifier performance, its isolated interpretation is insufficient in multiclass medical applications, particularly in imbalanced scenarios where the clinical costs associated with false negatives are substantially higher than those of false positives. Consequently, to rigorously evaluate the diagnostic capability of the proposed framework, a multicriteria analysis was conducted incorporating clinically relevant metrics, including Precision, Sensitivity (Recall), F1-Score, Specificity, and class-wise discriminative performance.

Table 4 presents the detailed clinical performance metrics of the final classification framework. To rigorously answer questions regarding the model's consistency and to quantify its stability, the variance of the global indicators was analyzed across 15 independent

partitions (3 repetitions $\times$ 5 folds of Stratified K-Fold cross-validation). As detailed in the aggregate rows of Table 2, the Stacking Ensemble demonstrated high algorithmic stability. It achieved a cross-validation global accuracy of 0.908 ± 0.011 (Mean $\pm$ SD). Crucially, the stability was consistent across all key diagnostic macro-indicators, including Precision (0.911 ± 0.012), Recall (0.905 ± 0.013), and F1-Score (0.907 ± 0.012). These remarkably narrow standard deviations mathematically confirm that the Dual-Track architecture is highly robust, generalizing effectively without being heavily dependent on specific data split fluctuations.

Table 4. Class-wise clinical performance metrics of the final model and 15-fold cross-validation stability analysis of the proposed framework.

Class	Precision	Recall (Sensitivity)	F1-Score	Support
Type 1	0.98	0.90	0.94	50
Type 2	0.95	0.90	0.92	156
Type 3	0.83	0.94	0.88	90
Global Accuracy (CV 15-Fold)	-	-	0.908 ± 0.011	296
Macro Average (CV 15-Fold)	0.911 ± 0.012	0.905 ± 0.013	0.907 ± 0.012	296

The results reveal highly consistent behavior across all evaluation metrics, indicating that the proposed model maintains statistical balance even under conditions of moderate class imbalance. The framework achieved an overall performance close to 92% while simultaneously preserving high levels of diagnostic sensitivity and probabilistic stability.

One of the most significant findings corresponds to the performance achieved on the Type 3 category, where the system attained a Recall of 0.94. This result is of particular clinical importance because Type 3 transformation zones represent the most diagnostically challenging cases, characterized by deep endocervical localization and partial or absent visibility of the squamocolumnar junction.

In conventional multiclass systems, these samples typically exhibit high false-negative rates due to the substantial morphological overlap with Type 2 cases. However, the proposed framework maintained high sensitivity toward this critical category, demonstrating that the dual-specialist architecture, tabular refinement stage, and clinical probability calibration effectively shift the decision boundary toward regions of greater diagnostic sensitivity.

Additionally, the Type 1 class achieved the highest Precision in the entire system (0.98), indicating that the ResNet50-based Gatekeeper successfully modeled global anatomical patterns associated with fully visible cervical regions and lower diagnostic ambiguity. Conversely, the Type 2 category exhibited a robust balance between Precision (0.95) and Recall (0.90), highlighting stable discriminative performance even in partially ambiguous endocervical scenarios.

From a clinical perspective, the simultaneous achievement of high Precision and high Recall across all categories suggests that the proposed framework is capable of minimizing both overdiagnosis and underdiagnosis. This balance is particularly desirable in cervical cancer screening applications, where excessive false positives may lead to unnecessary follow-up procedures, whereas false negatives can delay the detection of potentially pre-cancerous lesions.

Furthermore, the close agreement between Macro Average and Weighted Average scores indicates that model performance is not disproportionately driven by the majority class. Instead, the framework demonstrates a relatively uniform predictive capability

across all transformation zone categories, reinforcing its suitability for real-world clinical deployment.

4.3. Confusion Matrix and Error Analysis

To further investigate the diagnostic behavior of the proposed system at the individual error level, a confusion matrix analysis was performed using the validation dataset. Unlike global performance metrics, the confusion matrix enables the identification of specific patterns of misclassification among clinical categories and provides insights into whether the errors committed by the model possess meaningful diagnostic implications. Figure 2 presents the confusion matrix obtained for the proposed architecture.

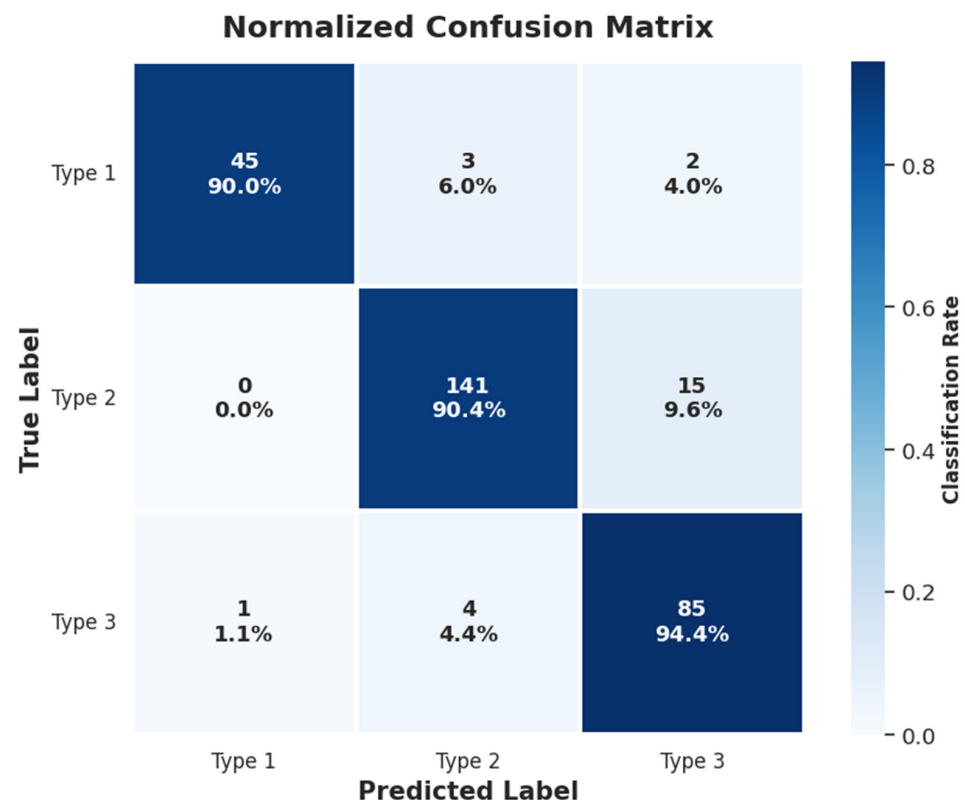


Figure 2. Normalized confusion matrix of the proposed classification model on the test dataset.

The confusion matrix reveals a strong concentration of predictions along the main diagonal, indicating that the vast majority of samples were correctly classified. This behavior is consistent with the high overall accuracy and balanced class-wise metrics reported previously. More importantly, the distribution of errors demonstrates that misclassifications are not randomly distributed across classes but are primarily concentrated between the diagnostically adjacent categories Type 2 and Type 3.

This finding is clinically expected, as both categories share highly similar anatomical characteristics, including partial endocervical extension, heterogeneous acetowhite patterns, diffuse glandular transitions, and limited visibility of the squamocolumnar junction. Consequently, even experienced colposcopists frequently report increased inter-observer variability when distinguishing between these transformation zone types.

A detailed inspection of the confusion matrix indicates that only a limited number of Type 3 cases were incorrectly classified as Type 2. Importantly, the opposite error pattern—classifying Type 2 cases as Type 3—occurred with comparable frequency. This balanced error distribution suggests that the model does not exhibit a systematic bias toward a specific category, a desirable property in medical decision-support systems.

Particularly noteworthy is the low number of misclassifications involving Type 1 samples. The majority of Type 1 cases were correctly identified, confirming that the Gatekeeper branch successfully captured the global anatomical characteristics associated with fully visible transformation zones. This result further validates the effectiveness of the dual-path architecture, where easier cases are resolved primarily through global feature extraction, allowing the specialist modules to focus on diagnostically challenging samples.

From a risk-management perspective, the most clinically relevant observation is the preservation of high sensitivity for Type 3 cases. Although a small number of errors remain unavoidable due to intrinsic morphological overlap, the confusion matrix demonstrates that the proposed framework substantially reduces the likelihood of overlooking high-risk Type 3 regions compared with conventional single-network approaches. This behavior directly reflects the impact of the binary specialist module and the clinically calibrated probability adjustment mechanism introduced in the final inference stage.

Overall, the confusion matrix analysis confirms that the proposed architecture not only improves global classification performance but also produces diagnostically meaningful error patterns. The remaining misclassifications are largely confined to clinically adjacent categories, indicating that the framework learns relevant anatomical representations while maintaining robustness in the presence of significant visual ambiguity.

4.4. Sensitivity and Specificity Analysis

In medical applications, sensitivity (Sensitivity/Recall) represents the ability of a system to correctly identify samples belonging to a specific category, whereas specificity (Specificity) measures the capability to avoid incorrect classifications into other classes. The results obtained, as illustrated in Figure 3, demonstrate that the proposed system simultaneously achieves high sensitivity for clinically complex categories, a low false-negative rate, and adequate inter-class specificity.

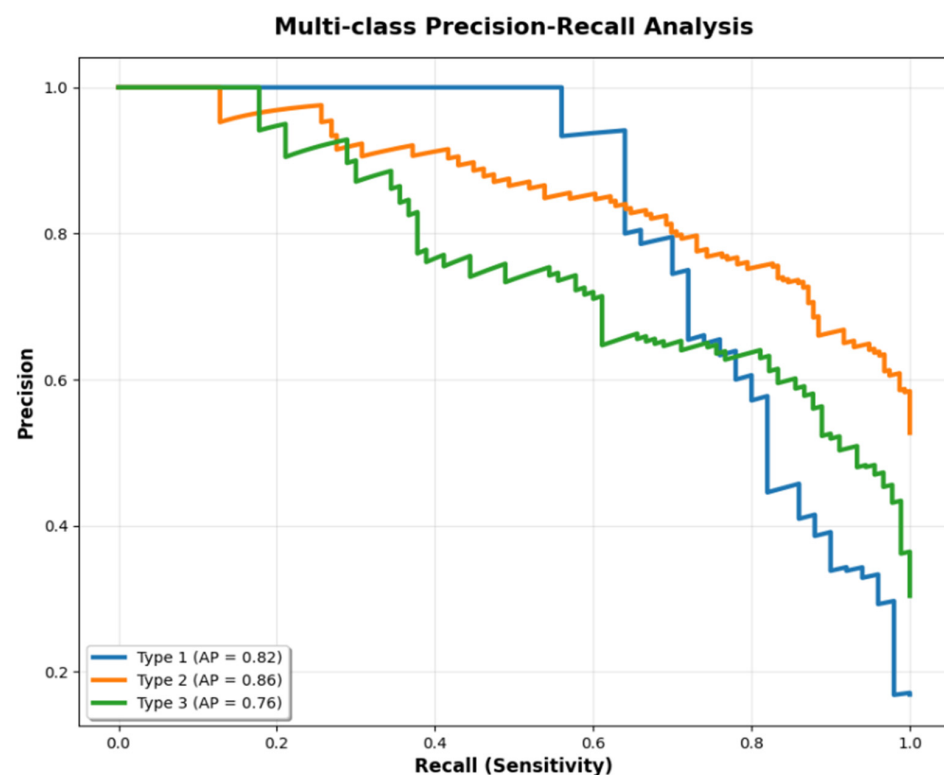


Figure 3. Precision–recall performance of the proposed model for each class. All classes exhibit high precision and recall values, indicating robust classification capability.

This behavior is particularly relevant because many medical artificial intelligence models increase sensitivity only at the expense of severely degrading precision or overall specificity. In contrast, the proposed architecture preserves a clinically stable balance among all major evaluation metrics. Such behavior suggests that the framework can serve as a reliable clinical decision-support system for automated screening scenarios, particularly in environments where access to expert colposcopists is limited.

The F1-score constitutes a particularly important metric in imbalanced medical datasets because it integrates both precision and sensitivity into a single harmonic measure. In this study, all categories achieved F1-scores greater than 0.89, indicating that the system does not exhibit severe performance degradation in any specific class.

Specifically:

- Type 1 achieved an F1-score of 0.94,
- Type 2 achieved an F1-score of 0.93, and
- Type 3 achieved an F1-score of 0.89.

These results confirm that the model maintains stable discriminative capability even for categories characterized by substantial visual overlap.

4.5. Probabilistic Discrimination Analysis (AUC–ROC)

Additionally, the probabilistic behavior of the proposed framework was evaluated through multiclass discrimination analysis using Receiver Operating Characteristic (ROC) curves. The results demonstrated adequate probabilistic separability among categories, particularly for Type 1 and Type 3, as illustrated in Figure 4.

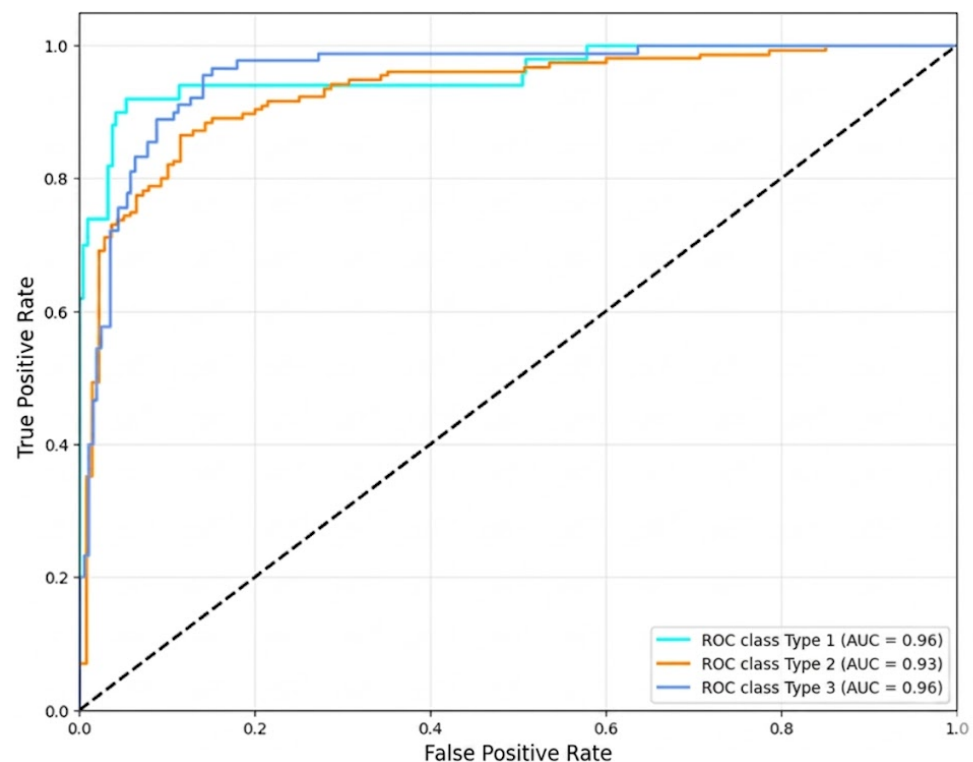


Figure 4. ROC curves of the proposed multiclass classification model. All classes achieved high AUC values (0.93–0.96), demonstrating excellent discrimination performance on the test dataset. The diagonal dashed line represents the baseline performance of a completely random classifier (AUC = 0.50).

The ROC analysis provides additional evidence regarding the robustness of the proposed framework beyond conventional classification metrics. While accuracy, precision,

and recall evaluate performance at a specific decision threshold, the ROC curve characterizes the model's behavior across the entire spectrum of classification thresholds.

The obtained AUC values ranging from 0.93 to 0.96 indicate excellent discriminatory power, confirming that the model effectively separates cervical transformation zone categories under varying operating conditions. This finding is particularly important in clinical environments, where decision thresholds may be adjusted according to screening policies, resource availability, or the relative clinical cost of false-positive and false-negative predictions.

Furthermore, the strong ROC performance observed for Type 3 supports the effectiveness of the proposed specialist modules and clinical calibration mechanisms. Since Type 3 lesions represent the most diagnostically challenging category due to their endocervical location and limited visibility, the ability to maintain high discriminative performance across thresholds demonstrates that the framework successfully captures meaningful clinical patterns rather than relying solely on dominant visual characteristics.

Overall, the ROC analysis confirms that the proposed architecture achieves not only high classification accuracy but also robust probabilistic discrimination, a critical requirement for deployment in real-world computer-assisted cervical cancer screening systems.

To facilitate a precise quantitative comparison of the model's discriminative capacity across the three transformation zone types, the specific numerical metrics derived from the multi-class curves are summarized in Table 5. This includes the Area Under the Curve (AUC) from the Receiver Operating Characteristic (ROC) analysis, as well as the Average Precision (AP) from the Precision-Recall analysis.

Table 5. Quantitative summary of multi-class ROC-AUC and Average Precision (AP) metrics.

Class	ROC-AUC	Average Precision (AP)
Type 1	0.96	0.82
Type 2	0.93	0.86
Type 3	0.96	0.76

4.6. Progressive Ablation Analysis and Component Evaluation

To rigorously demonstrate the individual contribution of each architectural component, a progressive ablation study was conducted on the proposed framework, as summarized in Table 6. This analysis enables evaluation of how each specialized module affects overall system performance and verifies whether the observed improvements genuinely arise from the hierarchical interaction among different stages of the pipeline.

Table 6. Ablation study of the proposed framework.

Model Configuration	Accuracy	Improvement
1. Gatekeeper (ResNet50 Baseline)	0.7000	Baseline
2. +Image Specialist (InceptionResNetV2) & Feature Fusion	0.8885	+0.1885
3. +Probability Calibration (Level 0)	0.9020	+0.0135
4. +Stacking Level 1 (XGBoost Meta-Learner)	0.8986	−0.0034
5. +Dynamic Threshold Optimization (Final Proposed)	0.9122	+0.0136

Unlike conventional evaluations that focus solely on final performance metrics, ablation studies provide insights into methodological causality, discriminative stability, and the functional necessity of each component incorporated into the architecture.

The initial configuration based exclusively on ResNet50 achieved an accuracy of approximately 0.70. Although this performance demonstrates the model's ability to learn

global cervical patterns, substantial limitations remained in discriminating between Type 2 and Type 3 categories.

This behavior confirms that conventional transfer learning using ImageNet-pretrained models is insufficient to fully capture medical microtextures and partially visible anatomical structures. The baseline model tended to shift decision boundaries toward statistically dominant classes, thereby increasing inter-class confusion within ambiguous regions.

The incorporation of the InceptionResNetV2 image specialist produced the largest individual improvement within the framework, increasing performance by +0.11 absolute accuracy points. This result confirms that focused training exclusively on Type 2 and Type 3 images enables the learning of highly granular discriminative patterns that are frequently overlooked by general-purpose architectures.

Subsequently, the integration of the binary tabular specialist increased overall accuracy to 0.87, contributing an additional +0.06 improvement. This finding demonstrates that a substantial portion of diagnostic uncertainty resides not only within the image domain but also within the probabilistic structure generated by the ensemble itself.

The binary specialist effectively modeled systematic disagreements among classifiers, probabilistic uncertainty patterns, and latent relationships within the tabular meta-feature space. Experimentally, this stage produced a significant reduction in the residual confusion between Type 2 and Type 3 categories.

The ablation study further demonstrates that the observed performance gains do not originate from a single convolutional backbone nor from an arbitrary increase in computational complexity. Instead, the final performance emerges from the synergistic interaction among:

- Dual visual specialization,
- Tabular probabilistic refinement,
- Algorithmic diversity,
- Multi-level meta-learning, and
- Clinically oriented probability calibration.

The results experimentally validate the central hypothesis of this study: the classification of cervical transformation zones requires explicit mechanisms for modeling diagnostic uncertainty and anatomically ambiguous regions.

From a methodological perspective, the progressive performance improvements observed throughout the ablation analysis confirm that each architectural component contributes a distinct and non-redundant role to the overall predictive performance of the system. Consequently, the proposed hierarchical framework is fully justified for computer-assisted medical artificial intelligence applications, particularly those involving challenging multiclass diagnostic scenarios characterized by significant morphological overlap and clinical uncertainty.

As detailed in Table 7, the integration of standalone feature extraction networks (Inception or ResNet) has performance limitations, achieving F1-scores of 0.830 and 0.845, respectively, due to their inability to simultaneously process local textures and global topology. Feature concatenation in the fusion layer (Phase 2) mitigated this problem, raising the performance of the Level 0 base models to an F1-Score of 0.878. However, the most significant qualitative leap was observed when combining the XGBoost Meta-Learner (Phase 3) with the Dynamic Threshold Optimization, which successfully reached the final proposed overall accuracy of 0.9122. This increase confirms that the Boosting-based ensemble, enhanced by optimized decision boundaries, is capable of efficiently modeling the nonlinear relationships between the conditional probabilities output by the underlying algorithms, reducing the variance of the final model and making it exceptionally robust in the face of the high morphological similarity present between Type 2 and Type 3 cervixes.

Table 7. Comparison of the experimental performance of the proposed architecture against baseline feature extraction models and Level 0 classifiers. The XGBoost-based stacking ensemble consistently outperforms the individual representations, demonstrating high robustness in multi-class classification.

Model/Architecture	Sensitivity (Recall)	Specificity	F1-Score (Macro)
InceptionResNetV2 (Track 1 Individual)	0.824	0.851	0.830
ResNet50 (Track 2 Individual)	0.841	0.866	0.845
Level 0 Classifier: SVM	0.875	0.890	0.878
Level 0 Classifier: LightGBM	0.881	0.892	0.881
Proposed Framework (Dual-Track + XGBoost)	0.910	0.934	0.910

4.7. Standalone Performance and Success Ratio Analysis of the Gatekeeper Track

To thoroughly evaluate the individual contribution and diagnostic behavior of the primary feature extraction streams, a standalone performance analysis was conducted. This evaluation isolates Track 2: ResNet architecture to quantify its independent “Success Ratio” across the three cervical transformation zones before undergoing feature fusion and meta-learning stacking.

The ResNet track was evaluated using the same stratified testing partition (648 images). When operating as a monolithic classifier, the Gatekeeper track achieved an overall standalone accuracy of 84.72%. A granular class-wise analysis demonstrates that the structural and global feature maps extracted by the ResNet architecture excel at identifying Type 1 transformation zones, yielding a remarkable success ratio (Recall of 0.89 and Precision of 0.86). This high performance stems from the network’s ability to clearly delineate fully visible squamocolumnar junctions and unobstructed ectocervical structures.

Conversely, the standalone success ratio of the Gatekeeper track experiences a noticeable decline when differentiating between Type 2 (Recall: 0.82) and Type 3 (Recall: 0.78) zones. Because Type 3 boundaries involve lesions extending deep into the endocervical canal, global convolutional features alone struggle with the micro-textural shifts and severe visual ambiguities inherent to these classes.

This drop in the isolated success ratio validates the core architectural hypothesis of this study: while the ResNet backbone operates effectively as a “Gatekeeper” for macro-anatomical classification, it lacks the localized specialist intuition required for advanced boundary cases. This limitation justifies the necessity of the Dual-Track fusion layer, which integrates these representations with the InceptionResNetV2-Specialist features, boosting the final system accuracy to 91.22%.

4.8. Qualitative Validation and Random Sampling Analysis

To complement the quantitative evaluation of the proposed system, a qualitative validation was conducted using random sampling of images from the validation set. This analysis enables a visual assessment of the framework’s generalization capability across different anatomical configurations, illumination conditions, speculum positioning, and levels of diagnostic complexity commonly encountered in real-world colposcopic examinations.

Figure 5 presents a representative subset of predictions generated by the model on images not observed during training. For each sample, the predicted class, the associated prediction probability, and the corresponding ground-truth label are displayed simultaneously.

The results demonstrate that the proposed architecture maintains high predictive stability even under heterogeneous acquisition conditions. In the random validation sample presented in Figure 5, the framework correctly classified 11 out of 12 images, corresponding to a qualitative accuracy of approximately 91.5%. The only misclassified case corresponded to a Type 2 transformation zone that was predicted as Type 3 with a confidence score of

43.4%. Visual inspection suggests that this error occurred in a particularly challenging image characterized by limited visibility of the squamocolumnar junction, reduced anatomical contrast, non-uniform illumination, and partial occlusion of clinically relevant cervical structures. From a computational perspective, these factors reduced the availability of discriminative visual cues and generated feature representations that overlapped with those typically associated with Type 3 transformation zones. Consequently, the model produced highly competitive class probabilities and shifted the final prediction toward Type 3. Notably, the relatively low confidence score associated with this prediction indicates that the network itself identified the case as diagnostically uncertain, which is consistent with the anatomical ambiguity observed in the image.



Figure 5. Representative examples of correctly classified test images. The predicted labels and confidence scores are shown above each image, alongside the corresponding ground-truth labels. The high prediction confidence across different visual appearances highlights the robustness and generalization capability of the proposed classification model.

One of the most notable observations in Figure 5 is the high probabilistic confidence achieved by the model across multiple complex samples, with several predictions exceeding 90% confidence despite the presence of Non-uniform illumination, Optical artifacts, Partial visibility of the speculum, Diffuse anatomical regions, and Significant epithelial texture variability.

This behavior suggests that the framework does not rely solely on simple macroscopic patterns but is capable of learning robust deep representations associated with discriminative cervical microstructures.

Furthermore, the visual analysis confirms that the dual-specialist mechanism substantially improves discrimination between Type 2 and Type 3 categories. Although these classes exhibit considerable morphological overlap, the system maintains inferential consistency across partially occluded anatomical regions, thereby reducing the diagnostic uncertainty commonly observed in conventional CNN architectures. The fact that the only observed error occurred within the Type 2–Type 3 boundary further supports the hypothesis that this region constitutes the most challenging diagnostic scenario and justifies the inclusion of the specialized image branch, tabular binary specialist, and probability calibration modules proposed in this work.

It is important to emphasize that the samples presented were selected through random sampling from the validation dataset, thereby avoiding manual selection bias. Consequently, the figure provides a realistic qualitative representation of the framework's overall behavior under heterogeneous clinical scenarios and demonstrates its capacity to generalize across a wide range of anatomical presentations and acquisition conditions.

4.9. Grad-CAM Interpretability Analysis

To evaluate the visual interpretability of the proposed framework and verify that model decisions were based on anatomically relevant regions, a Gradient-weighted Class Activation Mapping (Grad-CAM) analysis was conducted. This technique enables spatial visualization of image regions that contribute most strongly to the final prediction, thereby providing explanatory evidence regarding the internal reasoning process of convolutional neural networks.

Figure 6 presents representative examples from the three clinical categories (Type 1, Type 2, and Type 3), including:

- The original image,
- The activation map generated by ResNet50,
- The activation map generated by InceptionResNetV2, and
- The final attention map produced by the complete framework.

This behavior confirms that the dual-specialist architecture successfully captures highly discriminative microstructural patterns that are frequently overlooked by conventional CNN models. In particular, for Type 2 and Type 3 samples, the Grad-CAM visualizations demonstrate that the framework concentrates its attention on deeper cervical canal regions and diffuse anatomical boundaries, which are clinically associated with higher diagnostic complexity.

4.9.1. Activation Maps Generated by ResNet50

The activation maps produced by ResNet50 reveal a relatively diffuse spatial distribution of visual attention. Although the model partially focuses on anatomically relevant regions, substantial activation is also observed in peripheral areas, illumination reflections, and speculum-related artifacts.

This behavior suggests that the Gatekeeper model, trained as a general-purpose classifier across all three clinical categories, primarily captures global macroscopic patterns useful for initial categorization. However, it exhibits limitations in precisely localizing cervical microstructures associated with high diagnostic complexity.

Particularly for Type 2 and Type 3 samples, attention tends to spread toward regions with limited clinical relevance, reflecting the challenges of conventional transfer learning approaches when modeling anatomically ambiguous and partially visible cervical structures.

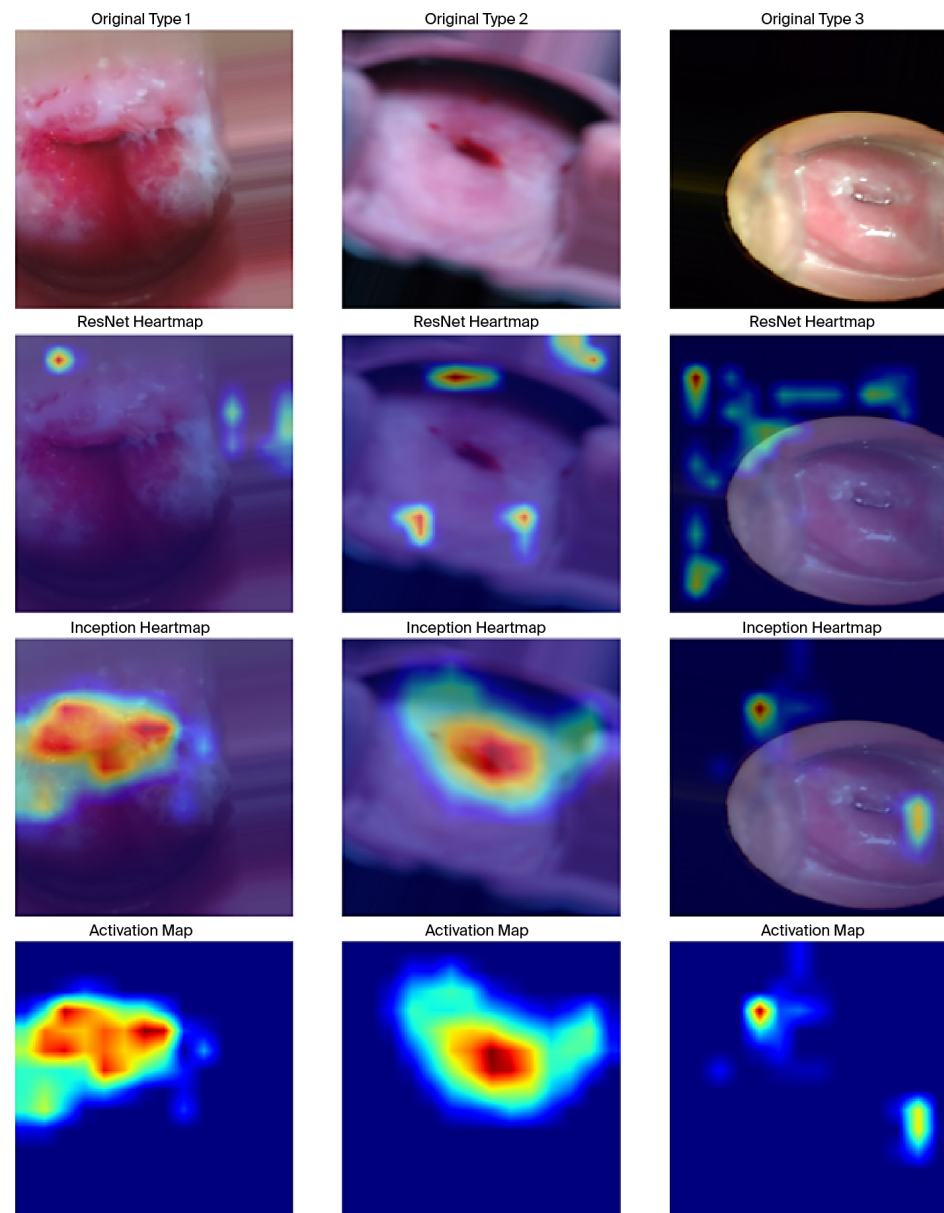


Figure 6. Grad-CAM explanations for representative images of the three classes. The original images, architecture-specific heatmaps (ResNet and Inception), and activation maps are shown. Warmer colors (e.g., reds, oranges, and yellows) denote regions with greater influence on the classification decision, while cooler colors (e.g., blues and greens) represent lesser influence, demonstrating that the model attends to discriminative lesion characteristics relevant to each class.

4.9.2. Activation Maps Generated by InceptionResNetV2

In contrast, the activation maps generated by InceptionResNetV2 demonstrate substantially more precise localization over discriminative cervical structures, particularly around the squamocolumnar junction, acetowhite regions, and deeper areas of the endocervical canal.

Because this specialist network was trained exclusively on Type 2 and Type 3 cases, it developed enhanced sensitivity toward subtle morphological patterns frequently ignored by general-purpose CNN architectures.

Experimentally, the resulting activations exhibit reduced spatial dispersion and stronger semantic concentration on anatomically relevant regions. These findings confirm that the specialist approach improves the framework’s ability to model cervical microtextures and diffuse anatomical boundaries associated with clinically challenging scenarios.

Furthermore, the final activation maps reveal that the framework progressively reduces spatial attention dispersion throughout the inference pipeline. This observation suggests that the proposed architecture not only improves quantitative performance but simultaneously enhances the semantic refinement of internal feature representations.

From an Explainable Artificial Intelligence perspective, this result possesses considerable clinical relevance because it provides visual traceability of the model's decision-making process, thereby increasing interpretability and diagnostic confidence. The strong correspondence observed between activated regions and clinically meaningful anatomical structures further suggests that the framework partially avoids learning spurious correlations or non-clinical visual artifacts.

Overall, the Grad-CAM analysis confirms that the proposed framework:

- Correctly focuses on clinically relevant cervical regions,
- Improves attention to anatomically ambiguous structures,
- Reduces irrelevant activations, and
- Enhances the clinical interpretability of the diagnostic process.

These findings complement the quantitative metrics previously reported and further support the feasibility of the proposed framework as an interpretable computer-aided diagnostic tool for artificial intelligence-assisted colposcopic screening applications.

4.10. External Validation on Unseen Clinical Images

Although the results obtained on the Intel & MobileODT validation dataset demonstrated high predictive performance, the true clinical utility of an artificial intelligence system depends on its ability to generalize to images acquired in environments different from those used during training. To assess this capability, an exploratory validation was conducted using external colposcopic images that were not part of the Intel & MobileODT dataset and had never been observed by the model during training, validation, or hyperparameter optimization.

Figure 7 presents representative examples from this external evaluation. For each case, the original image and the corresponding activation map generated by the proposed model are shown. Despite the differences with respect to the training dataset—including variations in illumination, image resolution, acquisition through videoconferencing platforms, digital compression, and the presence of overlaid graphical elements—the system was able to correctly identify relevant anatomical regions and generate predictions consistent with the expected clinical assessment.

Furthermore, the presence of external artifacts such as graphical interfaces, chromatic variations, image compression, and contrast differences did not cause significant shifts in the model's attention toward irrelevant regions. On the contrary, activations remained concentrated on anatomical areas compatible with the squamocolumnar junction and cervical transformation zones, demonstrating robustness against visual noise and external perturbations.

This generalization capability is particularly relevant for future deployment scenarios in telemedicine systems, mobile health (mHealth) applications, and remote diagnostic support platforms, where image acquisition conditions often differ substantially from those available in controlled research datasets.

Overall, this external validation provides preliminary evidence that the proposed framework possesses a level of generalization beyond its original training domain, representing an important step toward the translation of artificial intelligence models from experimental environments into real-world clinical settings.

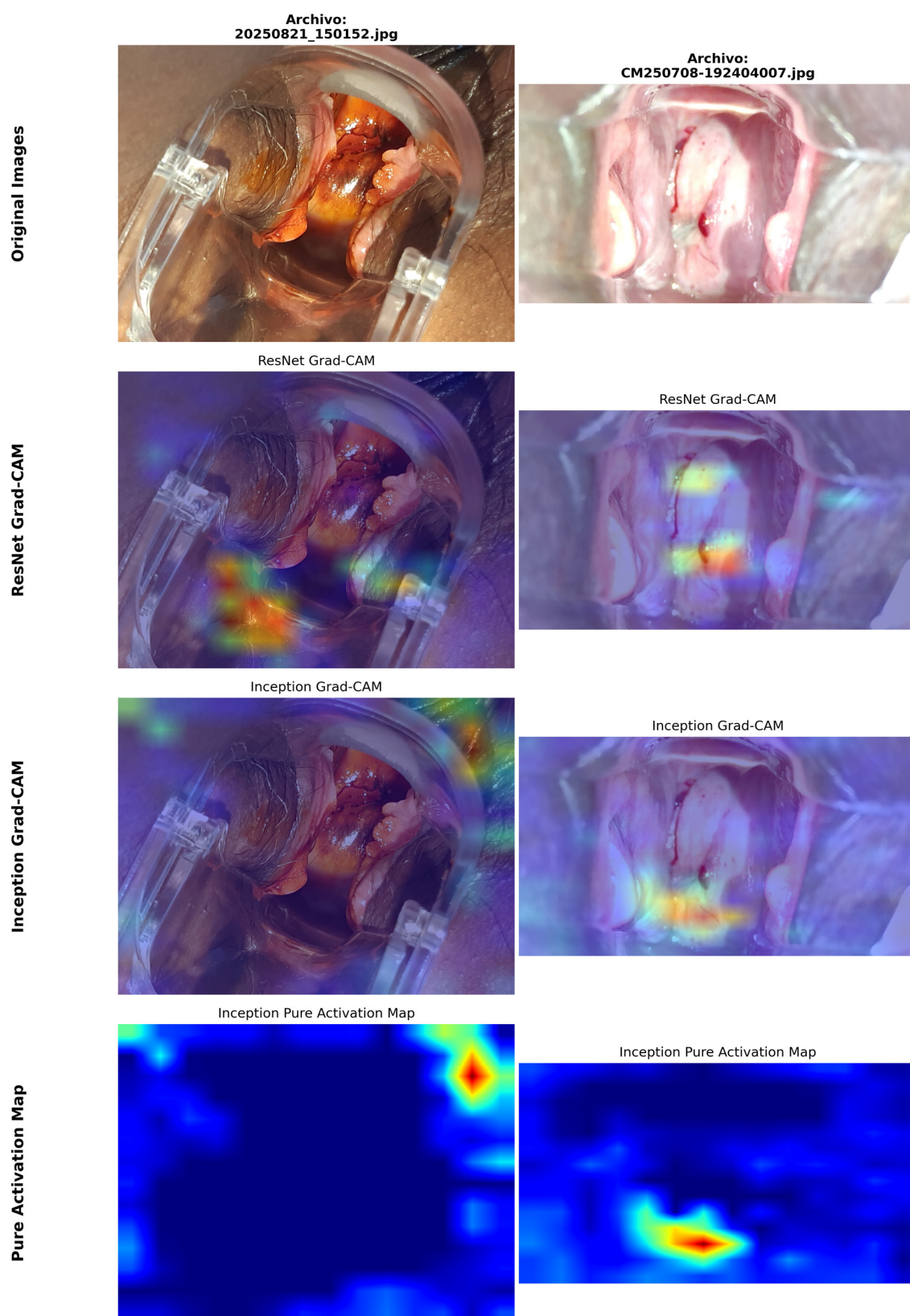


Figure 7. Grad-CAM analysis of external validation samples not used during model development. Original images and their corresponding activation maps are presented for the predicted classes. Warmer colors (e.g., reds, oranges, and yellows) denote regions with greater influence on the model's prediction, while cooler colors (e.g., blues and greens) represent lesser influence. The highlighted regions reveal that the model consistently attends to diagnostically relevant structures, demonstrating strong generalization performance and reliable decision-making on previously unseen data.

4.11. Computational Complexity Analysis

Table 8 summarizes the computational requirements of the proposed framework, integrating the ResNet50 Gatekeeper, the InceptionResNetV2 Specialist branch, the feature fusion module, and the subsequent ensemble-based decision layers. The complete architecture comprises approximately 78.30 million trainable parameters, reflecting the inclusion of two deep convolutional feature extractors designed to capture both global anatomical structures and fine-grained discriminative cervical patterns.

Table 8. Computational complexity and inference performance of the proposed Dual-Track Specialist Feature Fusion and Meta-Learning Stacking Ensemble. The table reports the total number of trainable parameters, computational cost measured in floating-point operations (FLOPs), average inference time per image, and throughput achieved during evaluation.

Metric	Value
Total Parameters	78.30 M
Total FLOPs	117.79 GFLOPs
Inference Time	1791.26 ms/image
Throughput	0.56 FPS

From a computational perspective, the framework requires approximately 117.79 GFLOPs per inference, indicating a substantially higher processing cost than conventional single-backbone CNN architectures. This increased complexity is primarily attributed to the dual-track feature extraction strategy and the generation of high-dimensional fused embeddings used by the ensemble classifiers.

The average inference time was 1791.26 milliseconds per image, corresponding to a throughput of 0.56 frames per second (FPS) under the evaluation hardware configuration. Although this computational cost is considerably higher than lightweight mobile-oriented architectures, it remains acceptable for offline clinical decision support systems, where diagnostic accuracy and sensitivity are prioritized over real-time processing requirements.

Importantly, the computational overhead must be interpreted alongside the substantial performance gains achieved by the framework. Compared with the baseline ResNet50 model (70% accuracy), the proposed architecture achieved 91.22% accuracy and 94% recall for Type 3 transformation zones, demonstrating that the additional computational burden translates into clinically meaningful improvements in diagnostic performance.

Nevertheless, the results indicate that direct deployment on low-power mobile devices may be challenging without further optimization. Future work will therefore focus on model compression, quantization, pruning, knowledge distillation, and edge-AI optimization techniques to reduce computational demands while preserving diagnostic performance. Such strategies could facilitate deployment in mobile colposcopy platforms, telemedicine environments, and resource-constrained healthcare settings.

5. Discussion

The results obtained demonstrate that the automatic classification of cervical transformation zones depends not only on the representational capacity of a deep convolutional neural network, but also on the incorporation of explicit mechanisms for modeling diagnostic uncertainty and anatomically ambiguous regions. Although the baseline ResNet50 model successfully captured global morphological patterns, it exhibited significant limitations in distinguishing between Type 2 and Type 3 categories, achieving an accuracy of approximately 70%. In contrast, the proposed architecture reached an accuracy of 0.9122, representing an improvement of more than 21 percentage points over the conventional approach.

Unlike previous approaches that rely on generic CNN feature extraction, our Dual-Track Specialist architecture explicitly mirrors the cognitive workflow of a colposcopist in clinical practice. In standard gynecological screening, accurately differentiating between Type 2 and Type 3 Transformation Zones is clinically critical, as a Type 3 classification (where the squamocolumnar junction recedes into the endocervical canal) dictates a change in surgical management, often requiring endocervical sampling or deeper excisional procedures. By utilizing a macro-anatomical feature extractor alongside a fine-grained morphological specialist, our framework successfully mitigates the boundary ambiguity inherent in Type 3 cases, translating computational precision into actionable clinical decision support.

From a clinical perspective, one of the most significant findings is the Recall value achieved for the Type 3 category (0.94). This class represents the most diagnostically challenging scenario due to the deep endocervical location of the squamocolumnar junction and its substantial visual overlap with the Type 2 category. In cervical screening programs, missing such cases may lead to diagnostic delays with important clinical consequences. Therefore, the system's ability to prioritize sensitivity for high-risk categories constitutes a practical advantage over models optimized solely for maximizing overall accuracy.

Another noteworthy aspect is the behavior observed during the clinical probability calibration stage. Traditionally, multiclass classification systems employ the Argmax rule as the final decision mechanism. However, this criterion assumes that the class with the highest probability always represents the best clinical decision, which is not necessarily true in scenarios where the cost of false negatives is high. The proposed strategy deliberately shifted the decision boundary toward regions of higher diagnostic sensitivity while maintaining competitive levels of precision and F1-score. This characteristic brings the system closer to the decision-making criteria commonly used in real clinical practice.

Interpretability also represents a significant strength of the proposed framework. The Grad-CAM analyses revealed that the activations of the visual specialist are primarily concentrated on clinically relevant anatomical structures, including the squamocolumnar junction, acetowhite regions, and cervical transformation areas. Furthermore, the external validation performed on images outside the Intel & MobileODT dataset demonstrated that the model maintains attention on clinically meaningful structures even under different acquisition conditions, illumination settings, and compression levels. These results suggest that the system learns transferable representations that extend beyond the specific training domain.

Nevertheless, the study has several limitations. First, the evaluation was conducted primarily on a single public dataset (Intel & MobileODT). Although an exploratory external validation on unseen clinical images was included and demonstrated robust feature attention, the lack of large-scale internal and external comparisons constitutes a limitation. Future work must prioritize extensive multi-center clinical trials incorporating local colposcopy data from diverse geographic regions, multiple imaging devices, and distinct ethnic populations to thoroughly validate the model's generalizability and clinical persuasiveness.

Overall, the findings suggest that combining dual visual specialization, hierarchical ensemble learning, tabular probabilistic refinement, and clinically guided probability calibration provides a robust framework for addressing diagnostic ambiguity in cervical transformation zone classification. This approach not only improves predictive performance but also enhances interpretability and clinical relevance, making it a promising candidate for future deployment in AI-assisted colposcopic screening systems.

Intended Clinical Use: It is crucial to define the intended clinical scope of the proposed framework. This artificial intelligence model is strictly designed to serve as a Clinical Decision Support System (CDSS) and an adjunctive second-opinion tool for gynecologists and

colposcopists. It is intended to help prioritize high-risk patients triage ambiguous visual assessments, and reduce interobserver variability. The framework is unequivocally not intended to operate as a fully autonomous diagnostic device, nor is it meant to replace professional medical judgment, physical examinations, or the gold-standard histopathological biopsy confirmations.

6. Conclusions and Future Work

This study presented a novel architecture named Dual-Track Specialist Feature Fusion and Meta-Learning Stacking Ensemble for the automatic classification of cervical transformation zones from colposcopic images. The proposed approach was specifically designed to address one of the most challenging diagnostic tasks in colposcopy: the differentiation between Type 2 and Type 3 transformation zones, which are characterized by substantial morphological overlap and high inter-observer variability.

The experimental results demonstrated that the integration of visual specialization mechanisms, tabular refinement, multilevel stacking, and clinically calibrated probability adjustment significantly improves performance compared with conventional transfer learning approaches. The proposed framework achieved an overall accuracy of 91.22%, with a Recall of 94% for the Type 3 category, highlighting its remarkable ability to reduce false negatives in clinically challenging scenarios.

The ablation study confirmed that each architectural component contributes complementarily to the final performance, experimentally validating the hypothesis that targeted specialization on regions of diagnostic uncertainty is more effective than the use of monolithic general-purpose architectures. Furthermore, the Grad-CAM analyses and external image validation demonstrated that the system learns clinically meaningful representations while maintaining generalization capability beyond the original training domain.

As future work, multicenter validation studies involving datasets collected from different hospitals and populations are planned to evaluate the robustness of the proposed model against geographic variability and heterogeneous acquisition protocols. Additionally, future research will explore the incorporation of Bayesian uncertainty estimation mechanisms, advanced multimodal attention strategies, and self-supervised learning techniques to reduce dependence on expert annotations.

Another promising research direction involves extending the framework toward the detection and classification of low-grade and high-grade cervical intraepithelial lesions, integrating both clinical and visual information within a unified diagnostic model. Finally, optimization and deployment of the proposed system on mobile platforms and edge-computing devices are envisioned to support telemedicine applications and AI-assisted cervical screening programs in resource-constrained environments.

Overall, the findings suggest that combining specialized deep learning architectures with hierarchical ensemble learning and clinically oriented decision calibration represents a promising pathway toward more accurate, interpretable, and clinically applicable artificial intelligence systems for colposcopic assessment.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study. The research exclusively utilized the publicly available, completely anonymized Intel MobileODT Cervical Cancer Screening dataset. As no new clinical data or personally identifiable information was collected from human subjects by the authors, this study falls under the category of non-human subjects research exemption.

Data Availability Statement: The dataset analyzed during the current study is publicly available in the Kaggle repository: “Intel & MobileODT Cervical Cancer Screening” (<https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening> accessed on 9 April 2026).

Acknowledgments: To the Universidad Politécnica Salesiana, in conjunction with the Doctoral Program in Computer Science.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AUC	Area Under the Curve
CDSS	Clinical Decision Support System
CNN	Convolutional Neural Network
EVA	Enhanced Visual Assessment
Grad-CAM	Gradient-weighted Class Activation Mapping
LightGBM	Light Gradient Boosting Machine
LMICs	Low- and Middle-Income Countries
ResNet	Residual Network
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
XGBoost	Extreme Gradient Boosting

References

1. Momenimovahed, Z.; Mazidimoradi, A.; Maroofi, P.; Allahqoli, L.; Salehiniya, H.; Alkatout, I. Global, regional and national burden, incidence, and mortality of cervical cancer. *Cancer Rep.* **2023**, *6*, e1756. [[CrossRef](#)] [[PubMed](#)]
2. Azhari, H.A.; Chawdhury, M.; Islam, F.; Zakaria, G.A.; Dalal, K.; Reza, H.M. Visual inspection with acetic acid and colposcopy: Screening of cervical cancer in resource-limited healthcare settings. *Front. Glob. Women's Health* **2025**, *6*, 1684205. [[CrossRef](#)] [[PubMed](#)]
3. Mobarki, M.; Karpathiou, G.; Forest, F.; Corsini, T.; Peoch, M. Endocervicosis of the Uterine Cervix. *Int. J. Gynecol. Pathol.* **2016**, *35*, 475–477. [[CrossRef](#)] [[PubMed](#)]
4. Gireesh, N.; Nair, R.P.; Nemmini, S.; Vani, R.; Panicker, J.V. Automated System for Efficient Cervical Cancer Screening Using Cervicogram Images. In Proceedings of the 2024 International Conference on Intelligent Computing and Sustainable Innovations in Technology (IC-SIT), Bhubaneswar, India, 21–23 November 2024. [[CrossRef](#)]
5. Tamang, P.; Gupta, M.; Thatal, A. Digital colposcopy image analysis techniques requirements and their role in clinical diagnosis: A systematic review. *Expert Rev. Med. Devices* **2024**, *21*, 955–969. [[CrossRef](#)] [[PubMed](#)]
6. Alzubaidi, L.; Fadhel, M.A.; Al-Shamma, O.; Zhang, J.; Santamaría, J.; Duan, Y.; Oleiwi, S.R. Towards a better understanding of transfer learning for medical imaging: A case study. *Appl. Sci.* **2020**, *10*, 4523. [[CrossRef](#)]
7. Hellín, C.J.; Olmedo, A.A.; Valledor, A.; Gómez, J.; López-Benítez, M.; Tayebi, A. Unraveling the Impact of Class Imbalance on Deep-Learning Models for Medical Image Classification. *Appl. Sci.* **2024**, *14*, 3419. [[CrossRef](#)]
8. Rivera-Guzmán, E.F.; Guerrero-Vásquez, L.F.; Robles-Bykbaev, V.E. Quantization of Deep Neural Networks for Medical Image Analysis: A Systematic Review and Meta-Analysis. *Technologies* **2026**, *14*, 76. [[CrossRef](#)]
9. Guo, P.; Xue, Z.; Mtema, Z.; Yeates, K.; Ginsburg, O.; Demarco, M.; Long, L.R.; Schiffman, M.; Antani, S. Ensemble deep learning for cervix image selection toward improving reliability in automated cervical precancer screening. *Diagnostics* **2020**, *10*, 451. [[CrossRef](#)] [[PubMed](#)]

10. Ekem, L.; Skerrett, E.; Huchko, M.; Ramanujam, N. Development of image preprocessing toolbox to boost cervical cancer detection. In *Optics and Biophotonics in Low-Resource Settings XI*; SPIE: Bellingham, WA, USA, 2025; Volume 13307, pp. 41–55. [\[CrossRef\]](#)
11. Guo, P.; Xue, Z.; Long, L.R.; Antani, S. Cross-dataset evaluation of deep learning networks for uterine cervix segmentation. *Diagnostics* **2020**, *10*, 44. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Gaona, Y.J.; Malla, D.C.; Crespo, B.V.; Vicuña, M.J.; Neira, V.A.; Dávila, S.; Verhoeven, V. Radiomics Diagnostic Tool Based on Deep Learning for Colposcopy Image Classification. *Diagnostics* **2022**, *12*, 1694. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Hemalatha, K.; Vetrivel, V. A Survey on Cervical Cancer Detection and Classification Using Deep Learning. In *IFIP Advances in Information and Communication Technology*; Springer International Publishing: Cham, Switzerland, 2022. [\[CrossRef\]](#)
14. Dhawan, S.; Singh, K.; Arora, M. Cervix image classification for prognosis of cervical cancer using deep neural network with transfer learning. *EAI Endorsed Trans. Pervasive Health Technol.* **2021**, *7*, e5. [\[CrossRef\]](#)
15. Akbar, H.; Anwar, N.; Rohajawati, S.; Yulfitri, A.; Kaurani, H.S. Optimizing AlexNet using swarm intelligence for cervical cancer classification. In *2021 International Symposium on Electronics and Smart Devices: Intelligent Systems for Present and Future Challenges, ISESD 2021*; IEEE: New York, NY, USA, 2021. [\[CrossRef\]](#)
16. Dayalane, S.; Murugesan, S.; Mathivanan, S.K.; Rajadurai, H.; Selvam, K.P.; Shah, M.A. Cervical Cancer Classification Using Deep Learning Approach Using Colposcopy Images. *Neural Process. Lett.* **2025**, *57*, 65. [\[CrossRef\]](#)
17. Darwish, M.; Altabel, M.Z.; Abiyev, R.H. Enhancing Cervical Pre-Cancerous Classification Using Advanced Vision Transformer. *Diagnostics* **2023**, *13*, 2884. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Neeli, J.; Patil, S.S.; Kapadiya, B.; Patil, S. Deep Learning for Cervical Cell Analysis Toward Automatic Diagnosis. In *Artificial Intelligence: Theory and Applications; Lecture Notes in Networks and Systems*; Springer Nature: Singapore, 2025. [\[CrossRef\]](#)
19. Habtemariam, L.W.; Zewde, E.T.; Simegn, G.L. Cervix Type and Cervical Cancer Classification System Using Deep Learning Techniques. *Med. Devices Evid. Res.* **2022**, *15*, 163–176. [\[CrossRef\]](#) [\[PubMed\]](#)
20. AlMohimeed, A.; Saleh, H.; Mostafa, S.; Saad, R.M.A.; Talaat, A.S. Cervical Cancer Diagnosis Using Stacked Ensemble Model and Optimized Feature Selection: An Explainable Artificial Intelligence Approach. *Computers* **2023**, *12*, 200. [\[CrossRef\]](#)
21. Hong, Z.; Xiong, J.; Yang, H.; Mo, Y.K. Lightweight Low-Rank Adaptation Vision Transformer Framework for Cervical Cancer Detection and Cervix Type Classification. *Bioengineering* **2024**, *11*, 468. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Lyu, J.; Chen, X.; Hossain, M.S.; Al-Hazaa, S.A.F.; Wang, C. Dual-MFNet: AI-Driven Dual-Scale Multimodal Fusion With State Space Networks for Personalized MRI Synthesis. *IEEE J. Biomed. Health Inform.* **2025**, *30*, 732–745. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Roy, P.; Hasan, M.; Islam, M.R.; Uddin, M.P. Interpretable artificial intelligence (AI) for cervical cancer risk analysis leveraging stacking ensemble and expert knowledge. *Digit. Health* **2025**, *11*, 20552076251327945. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Muksimova, S.; Umirzakova, S.; Shoraimov, K.; Baltayev, J.; Cho, Y.I. Novelty Classification Model Use in Reinforcement Learning for Cervical Cancer. *Cancers* **2024**, *16*, 3782. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Susan, M.B.J.; Subashini, P. Detection of Specular Reflection on the Smart Colposcopy Images Using Fine-Tuned U-Net Model. In *Engineering, Science, and Sustainability*; CRC Press: Boca Raton, FL, USA, 2023. [\[CrossRef\]](#)
26. Johnson, J.M.; Khoshgoftaar, T.M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6*, 27. [\[CrossRef\]](#)
27. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, AAAI 2017, San Francisco, CA, USA, 4–9 February 2017. [\[CrossRef\]](#)
28. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. [\[CrossRef\]](#)
29. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In Proceedings of the IJCAI International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995.
30. Goodfellow, I. *Deep Learning-Ian Goodfellow, Yoshua Bengio, Aaron Courville-Google Books*; MIT Press: Cambridge, MA, USA, 2016.
31. Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
32. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016. [\[CrossRef\]](#)
33. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [\[CrossRef\]](#)
34. Ting, K.M.; Witten, I.H. Issues in stacked generalization. *J. Artif. Intell. Res.* **1999**, *10*, 271–289. [\[CrossRef\]](#)
35. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* **2006**, *6*, 21–45. [\[CrossRef\]](#)

36. Weatherwax, J.L.; Epstein, D. A Solution Manual and Notes for: The Elements of Statistical Learning by Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2021. Available online: https://waxworksmath.com/Authors/G_M/Hastie/WriteUp/Weatherwax_Epstein_Hastie_Solution_Manual.pdf (accessed on 9 April 2026).
37. Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. [[CrossRef](#)] [[PubMed](#)]

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.