IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 March 2025, accepted 12 April 2025, date of publication 21 April 2025, date of current version 5 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3562762

RESEARCH ARTICLE

Segmentation Techniques Applied to CNNs for Cervical Cancer Classification

ANA ORTIZ-GONZÁLEZ[1], RAQUEL MARTÍNEZ-ESPAÑA[2], JUAN MORALES-GARCÍA[3], BALDOMERO IMBERNÓN[4], JOSÉ MARTÍNEZ-MÁS[5,6], MAURICIO A. ÁLVAREZ[7], OSCAR DAVID ROMERO[4], JUAN PEDRO MARTÍNEZ-CENDÁN[5,8], AND ANDRÉS BUENO-CRESPO[4]

[1]Pathology Department, University Hospital Complex of Cartagena, 30202 Cartagena, Spain
[2]Information and Communications Engineering Department, University of Murcia, 30100 Murcia, Spain
[3]Department of Software and Computing Systems, University of Alicante, 03690 Alicante, Spain
[4]Escuela Politécnica Superior (Computer Science), Universidad Católica de Murcia, 30107 Murcia, Spain
[5]Faculty of Medicine, Catholic University of Murcia, 30107 Murcia, Spain
[6]Obstetrics and Gynecology Department, CIAGO Gynecological Center, 30107 Murcia, Spain
[7]Computer Science Department, The University of Manchester, M13 9PL Manchester, U.K.
[8]Obstetrics and Gynecology Department, University General Hospital Santa Lucía, 30202 Murcia, Spain

Corresponding author: Andrés Bueno-Crespo (abueno@ucam.edu)

This work was supported in part by the PMAFI-21/21 Project through the Research Help Program of the Catholic University of Murcia; and in part by the Program for Mobility, Cooperation, and Internationalization "Jiménez de la Espada" under Grant 22466/EE/24 of Andrés Bueno-Crespo funded by the Seneca Foundation – Agency for Science and Technology in the Region of Murcia.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee at Catholic University of Murcia under Approval No. CE012005.

ABSTRACT Cervical cancer continues to be a significant global health issue, ranking as the fourth most prevalent cancer affecting women. Enhancing population screening programs by refining the examination of cervical samples conducted by skilled pathologists offers a compelling alternative for early detection of this disease. Deep Learning facilitates the development of automatic classification models to aid experts in this task. However, it is increasingly important to bring explainability to the model both to understand how the network learns to identify pathology and to bring confidence to the diagnosis. In this paper, we design an automatic segmentation masks for the classification of cervicovaginal cell images. This automatic segmentation is combined in a classification model that allows the models to improve their performance thanks to the morphological information provided by the combined segmentation in a Global Average Pooling layer with the convolutional network analysis of the original image. The models will be trained with real data so that learning can recognize the diversity of colors, shapes and sizes of human cell nuclei. The results show a robust and explainable model with satisfactory results, obtaining an F1 Score value of 0.935 in binary classification of revisable and non-revisable cell.

INDEX TERMS Artificial intelligence, deep learning, artificial vision, supervised classification, convolutional neural networks, segmentation, cervical cancer, papanicolaou smear.

I. INTRODUCTION

The cervix is a fibromuscular structure that connects the uterus to the vagina, which is lined by epithelial cells that provide protection to the tissue. When these cells are infected by the Human Papillomavirus (HPV), the viral genetic material can become integrated into the cervical cell, which along with other associated genetic and environmental factors, acquires the potential risk of developing cancer. Fortunately, the infected cell goes through stages prior to the development of cancer and are known as dysplastic cells.

Cervical cancer remains a significant global health issue and is the fourth most prevalent cancer among women, according to recent estimates [1]. In 2020 alone, approximately 604,000 new cases were reported, leading to over 340,000 deaths. This type of cancer accounts for 6.5%

The associate editor coordinating the review of this manuscript and approving it for publication was Kumaradevan Punithakumar.

of all tumors affecting women of all ages and showed a global cancer mortality rate of 7.7%, based on GLOBOCAN data from 2020. Cervical cancer is the leading cause of cancer-related deaths among women in several African nations, with 85-90% of cervical cancer deaths occurring in underdeveloped countries, where mortality rates are 18 times higher than in developed nations [1]. Routine medical screenings and the HPV vaccine have the potential to lower both the incidence and mortality of cervical cancer in high-income countries.

Despite advances in medical care, survival rates for cervical cancer drop significantly if it is not diagnosed in the early stages [2]. This highlights the critical role of prevention with regular screenings in fighting the disease. Preventive strategies include community education, social mobilization, and, importantly, vaccination [3]. Currently, four vaccines are available that protect against HPV, the virus responsible for more than 95% of cases of cervical cancer. These vaccines have demonstrated high effectiveness, reducing cancer incidence by up to 78% for individuals vaccinated before age 17, and by 60% for those vaccinated between ages 17 and 30. [4].

Screening strategies detect at-risk healthy individuals for early intervention, requiring a detectable precancerous stage and a reliable test.Various tests are used for early cervical cancer detection, such as HPV detection by PCR, vaginal cytology, and colposcopy [5]. Among these, HPV detection via PCR is more sensitive and performs better than cytology for identifying precancerous lesions, while also being simpler and more cost-effective than visual inspection techniques [3]. However, HPV detection is more costly and less specific than cytology, making Pap smears the most widely used screening method today. This low specificity is justified because the infectious status of HPV does not necessarily imply per se, the development of dysplasia or neoplasia since, in a significant percentage of patients, the infection resolves itself without producing organic lesions.

Cervicovaginal cytology involves collecting cervical cells with a brush, preserving them in a fixative, and staining them using the Papanicolaou technique.This staining enables the cells to be viewed under a microscope in various colors. By examining these cells and analyzing their cytological characteristics, clinicians can assess the tissue's condition and detect precancerous or dysplastic lesions, which, if left untreated, may progress to cancer. The Bethesda System evaluates all elements found in Papanicolaou cytology and defines morphological and diagnostic categories to classify cytology based on cell characteristics. It introduces the term ''squamous intraepithelial lesion (SIL)'' to describe nuclear and cytoplasmic changes in cervical squamous cells in response to HPV infection, categorizing them as low-grade (L-SIL) or high-grade (H-SIL). These nuclear characteristics reflect the infective state of these cells by the HPV virus, which carries a lower or higher risk of developing cancer respectively depending on the diagnostic category in which the cell is classified.In L-SIL, the observed changes suggest HPV infection without integration of the viral DNA into host DNA, whereas in H-SIL, the changes indicate viral DNA integration into host cells. L-SIL has a low risk of progression to cancer and frequently resolves spontaneously, while H-SIL carries a higher risk of developing into cervical cancer. For uncertain cases with findings that suggest dysplasia but are inconclusive, the Bethesda System uses the term ASC-US, i.e, Atypical Squamous Cells of Undetermined Significance.

While the Pap smear is effective in detecting abnormalities, it requires careful examination to identify lesions, making it a time-consuming and labor-intensive process that relies on skilled professionals. In cytopathology, significant inter- and intra-observer variability exists, as the procedure involves a degree of subjectivity influenced by factors such as observer fatigue, sample quality, and processing artifacts in cytology, among others.

To improve diagnostic efficiency, machine learning (ML) has gained considerable attention for its applicability, particularly convolutional neural networks (CNNs), a specialized type of ML algorithm designed for computer vision and recognized as powerful tools in image-based medical diagnostics. CNNs have been successfully applied in diagnosing various diseases, including COVID-19 through X-ray images [6], CT scans [7], brain tumor classification [8], as well as Parkinson's disease, Alzheimer's disease, and schizophrenia using magnetic resonance imaging [9], [10]. They have also been used to analyze cervical cancer using both public image datasets [11] and newly created image datasets [12]. Public available datasets are composed by pre-processed images that omit folded, overlapped or blurred cells, usually found in real smears, so, these methods [11] could not be compared with the results obtained after classifying real images without pre-processing. Although CNNs demonstrate impressive diagnostic capabilities, often exceeding traditional methods in both accuracy and speed, these models generally lack transparency in their decision-making processes, functioning as 'non-interpretable' models. Applying Explainable AI (XAI) methods can help clarify the reasoning behind a model's decisions. In the medical field, where trust is paramount, this transparency is essential for clinicians to confidently rely on AI models. Numerous studies have worked to enhance the explainability of CNN and ML models in healthcare, as emphasized in recent surveys [13], [14], [15]. Within the XIA, image segmentation is a branch which involves dividing an entire image into distinct regions. This process is essential across a variety of applications. A prominent example in this field is medical image segmentation, which provides numerous benefits for clinical practice. Automated segmentation streamlines data processing and aids clinicians by offering task-specific visualizations and measurements. In nearly all clinical applications, the visualization algorithm not only highlights abnormal areas in human tissues but also assists healthcare professionals in monitoring cancer progression [16]. Image

segmentation, applied to the cells of the problem addressed in this work, makes it possible to discern and provide information on the different details of the cell. This information highlighted in the segmentation can help to add information and improve the classification results. Following this idea, this work proposes an approach to perform automatic cell segmentation, mainly highlighting and differentiating the nucleus from the cytoplasm. Furthermore, using cell segmentation, it is proposed to design a classification method that combines the information from the original images together with the segmented images to create an improved classification system. To understand these proposals, it is necessary to understand the challenges of the problem faced in this work. There is a major problem in obtaining real data in healthcare settings. This results in not having a large number of images of all the categories covered. Moreover, in a cytology the number of benign cells is much higher than dysplastic cells and this results in an unequal database between each diagnostic category. On the other hand, the images treated in this work, are images without preprocessing, obtained directly from the scanner, so they appear folded, blurred and/or overlapped cells, this makes the segmentation process more expensive and the classification method has more problems to distinguish the categories. The reason for not preprocessing the images is that we try to get as close as possible to real diagnostic conditions, where the pathologist may be confronted with cells with artifacts that make viewing difficult. Furthermore, in the L-SIL, ASC-US categories, even expert pathologists have difficulty discerning one category from another, due to their high similarities between cells of both categories in some cases.

Thus, having highlighted the challenges to be met, the main findings of the study are:

- Designing, implementing and constructing models trained with real images obtained from human clinical trials, rather than from preprocessed datasets, that:
  - Automatically segment images of cervicovaginal cells.
  - Improve the basic classification model, adding the combination of the segmentation approach with an image classification strategy.
  - Add explainability to the results through segmentation.
- Analyzing and compare the proposals made with a basic image classification approach.
- Applying the automatic segmentation approach to the educational world to train future pathologists.

The rest of the paper is organised as follows. Section II shows a review of some work on the use of XIA. Section III describes the discussion of the different models studied and compared, along with the description of the dataset and the tools used for its evaluation. Section IV the results obtained are shown, analysed and discussed. Finally, Section V presents the main conclusions and discusses future work.

II. RELATED WORKS

Deep learning techniques are highly accurate and achieve strong classification and regression results across various applications. However, they are often considered non-interpretable methods, making it challenging and costly to understand the reasons behind their outcomes. In recent years, efforts have been made to develop techniques that enhance the interpretability of these approaches [17]. Such interpretability methods fall under the umbrella of explainable artificial intelligence systems (XAI). The field of XAI includes various approaches aimed at shedding light on how a model operates, clarifying its results, and enhancing overall interpretability for end-users, including human decision-makers [18]. As previously mentioned, image segmentation is one of the areas that XIA is exploring to improve the interpretability. Image segmentation can be categorised into two main classes, semantic segmentation and instance segmentation. In instance segmentation, several instances of each class in the image are searched for, grouping a set of pixels to that particular instance, as opposed to semantic segmentation where pixels are only assigned to one class [19].

In literature we can find various applications of segmentation in medical imaging, we will mention some of them focusing on the segmentation technique used. In [20], a hybrid approach for segmenting brain lesions across various imaging modalities is introduced, combining median filtering, k-means clustering, Sobel edge detection, and morphological operations. Median filtering serves as a crucial preprocessing step, effectively removing impulsive noise from the acquired brain images. This is followed by k-means segmentation, Sobel edge detection, and morphological processing. The performance of the proposed automated system is evaluated on standard datasets using metrics like segmentation accuracy and execution time. The method achieves an impressive 94% accuracy when compared to manual delineation conducted by an expert radiologist. Other work where authors perform medical image segmentation es presented in [21]. There, the authors propose a segmentation method called N-net based on the U-net model. The authors start from a dual encoder model to deepen the network and improve the feature extraction capability. In our implementation, the Squeeze-and-Excitation (SE) module is added to the dual encoder model to obtain global channel-level features. In addition, the introduction of full-scale hopping connections favours the integration of low-level details and high-level semantic information. They compare the proposal with lung and liver image datasets and show that their proposal can be improved with respect to several U-Net based segmentation models. The authors of [22] propose a segmentation method that combines two U-Net architectures stacked on top of each other. The first U-Net uses a pre-trained VGG-19 as an encoder and to capture more semantic information efficiently, the authors propose to add another U-Net at the bottom. This method has been tested using four image datasets related to colonoscopy, dermatoscopy and microscopy images. Another

study is presented in [23]. In this study, a Multiscale Residual Fusion Network architecture designed for medical image segmentation is proposed. This network is able to exchange multiscale features of varying receptive fields by means of a dual-scale dense fusion block. This allows for resolution preserving, enhanced information flow and propagation of high and low level features for accurate segmentation maps. In addition, the approach allows capturing object variabilities and provides better results on different biomedical datasets. This network has been tested on 4 public image datasets with satisfactory results.

Advanced U-Net models, such as the modified Double U-Net proposed by the authors of [24], combine two stacked U-Net models, with the first utilizing an ensemble of pre-trained models (Xception, DenseNet, VGG-19) and the second capturing additional information for enhanced segmentation. Evaluated on five medical datasets, the model achieves state-of-the-art results. The authors of [25] introduce Swin-Unet, a Transformer-based U-shaped architecture for medical image segmentation, which outperforms other methods using full convolution or hybrid Transformer-CNN models in multi-organ and cardiac segmentation tasks. Additionally, hybrid Transformer-CNN architectures, like HAU-Net [26] and PFormer [27], further enhance segmentation performance in breast ultrasound lesion and 3D medical image segmentation, respectively, demonstrating superior results with reduced computational costs. In the past year, contributions have emerged related to medical image analysis, evaluating the most widely used and promising architectures. For example, [28] highlights U-Net, Xception, and ResNet as the leading approaches in the analysis of cervical cytology images. Similarly, studies [29] and [30] focus on U-Net and transformer architectures, exploring their integration into medical image segmentation. Finally, [31] performs a comprehensive analysis of the models for image-based gynecological cancer diagnosis, showing that the models based on ResNet and U-Net achieve better performance in terms of prediction and efficiency.

This concise literary summary demonstrates that explainable AI techniques are effective not only in identifying the reasons behind classifications and diagnoses but also in enhancing those diagnoses. This underscores the value of these techniques for classifying cervical images, facilitating quicker and more accurate identification of the most relevant regions indicative of disease.

III. MATERIALS AND METHODS

This section presents the characteristics of the base dataset used, the segmentation approach proposed and the combined architecture approach to improve the classification by means of segmentation.

A. BASE DATASET

The base dataset used for this project was generously provided by the Pathology Service of the Complejo Hospitalario Universitario Santa Lucía-Santa María del Rosell, located in Cartagena, Murcia. Its consists of 1743 cell images, each meticulously labeled by cervical cancer experts into one of four distinct classes according to the Bethesda System. The hospital selected attend a multiethnical population, and the sample includes Subsaharian, Moroccan, European and from Eastern Europe women. However, there is no racial difference between cervical cells under microscope examination. The class distribution by severity is as follows: 389 H-SIL cells, 293 ASC-US cells, 182 L-SIL cells and 879 Benign cells.

One of the main challenges in developing deep learning models for cervical diagnostics is the class imbalance, which reflects clinical reality. In cytology, the vast majority of samples correspond to benign cells, while cases of malignant or suspected malignancy are significantly less frequent. It is common to find only a few abnormal cells among thousands of benign cells. This pattern has been respected in our proprietary database, ensuring that our DL model learns to correctly identify these minority cases without compromising its ability to recognise normal samples. To address this imbalance, data augmentation techniques, by data generation, have been applied to optimise the accuracy of the model without introducing biases that may affect its applicability in real-world clinical settings. Data augmentation is done so that all classes match the majority class, increasing the minority classes to the same volume as the majority class (benign class). This involved applying random transformations to each image, including: rotations of up to 10 degrees, zooming by as much as 15%, shifts (both horizontal and vertical) of up to 20%, shearing up to 15%, and both horizontal and vertical flips. To address any gaps created by these transformations, the nearest pixel will fill the area by stretching. In addition, it must be taken into account that the images used are real images obtained directly from the scanner, which results in folded, overlapping, blurred and/or dye-stained cells. Figure 1 shows for each category 2 examples of cells, one example where the cell is seen alone and another where the cells are folded, overlapped and/or stained. For the test samples, 30 samples of each class have been selected by expert pathologists to ensure that there is a full range of types. With the remaining samples, 80% were selected for training, while the remaining 20% were used for validation.

B. MASK DATASET GENERATION

Starting from the base dataset shown in Sec. III-A, the creation of an image mask dataset for each image is proposed. For the creation of this mask dataset, a semi-automatic method based on two steps is proposed: manually labeling a representative sample of the images and automatically generating the masks for the remaining images. To initiate the manual labeling process, a representative sample from the available image set was selected. Specifically, 20 images from each class were chosen, totaling 80 images, for manual mask generation. This mask creation was carried out by medical specialists with experience in identifying the different components of a cell: the nucleus and cytoplasm, as well as the background (areas outside the cell).

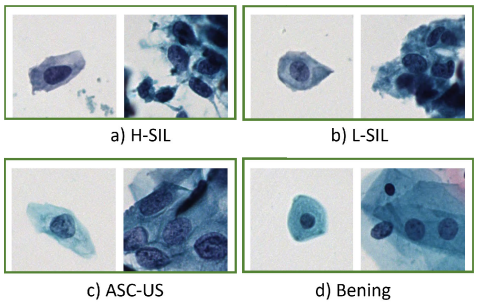


FIGURE 1. Types of images for each diagnostic category. Those on the left show a single cell, centered in the image, of high quality in which the nuclear and cytoplasmic cellular features are clearly visible. The images on the right are complex as they contain multiple cells of poorer quality due to factors such as: folding, overlapping, blurred and intensely contrasted areas. Our data set contains both types.

For the automatic mask generation process, based on the understanding of convolutional neural networks (CNNs) applied to image classification, an efficient and simple segmentation model was sought to avoid long training periods and ensure full comprehension of the model. The simplest model for segmentation is a Fully Convolutional Network (FCN). Both FCNs and CNNs for image classification share an input layer followed by convolutional and MaxPooling layers for feature extraction. However, instead of using fully connected layers to predict a class, FCNs use a decoder to return the image to its original size, as shown in Figure 2 in the segmentation part.This type of segmentation model faces a problem: as features are extracted, the output size reduces, leading to less precise feature maps at greater depths. Consequently, when the decoder is used to resize the output back to the image size, the output loses significant precision. Among the models developed after traditional FCNs, U-Net, created by [32], was selected. This network, originally designed to identify tumors in the lungs and brain, is a modification of FCNs.

In the proposed model in this paper, U-Net uses convolutional blocks and MaxPooling for feature extraction, along with upsampling convolutional blocks. However, to avoid the issue faced by FCNs, U-Net concatenates the output of the convolutional block with the input of the corresponding upsampling convolutional block to retain spatial information. Another advantage of U-Net is its efficiency. Being relatively small and simple, it delivers high-quality results in a short period of time. For these reasons, U-Net was chosen for the current problem.

For these reasons, the U-Net model is selected for the automatic generation of masks for the remaining images. Specifically, this model is trained using the manually labeled image set, with the original images provided as input and the image masks as the expected output. Once the model has been trained in this way, the remaining original images are processed to automatically generate their associated masks.

In Figure 2, this process is shown, where the U-NET is trained by a manual process and subsequently, the network creates the automatic segmentation from an input cell image. As observed, the image on the left represents the original image provided to the model for mask generation; the center image shows the mask created manually by an expert, and the image on the right displays the mask automatically generated by the model. A high similarity is evident between the expert-generated mask (center image) and the automatically generated mask (right image), with an IoU (Intersection over Union, also known as the Jaccard Index) value of 0.94.

C. MODEL ARCHITECTURE

For the comparative analysis of image classification models, three advanced architectures were meticulously developed, each offering distinct strengths and refined design considerations tailored to tasks requiring both segmentation and classification of complex cell images. Figure 2 shows the scheme of the 3 proposed models. We will now describe each of them.

The **Classical Model** is a classical architecture based on an Xception network, which is specifically designed for image classification. It processes only image inputs, without the need for masks, and uses a densely connected classifier for prediction. The training was initialised with "ImageNet" weights, with the first 40 layers frozen to retain essential foundational features, such as edges and textures, while allowing later layers to specialise on the nuances of the cellular image. Althoughsuitable for general image classification tasks, this architecture does not take full advantage of the advanced integration of segmentation and feature extraction seen in the other proposed models, which process both images and segmentation masks in parallel.

The **Proposed Model 1** integrates the U-Net segmentation network with Xception, which allows simultaneous processing of segmentation and feature extraction. Unlike the classical model, the proposed model 1 models both images and segmentation masks in parallel. The U-Net component segments the image while extracting features from both the segmented mask and the original image, which increases the model's ability to capture intricate cellular patterns and improves performance in both segmentation and classification tasks. In addition, the U-Net output is connected to a ResNet network that takes advantage of gradient preservation properties to reduce information loss, which is crucial for capturing details of cellular anomalies that are crucial for applications requiring high-precision segmentation. Finally, a classifier receives the output from Xception and ResNet to generate the diagnosis. This classifier uses a flattened layer and finally a dense layer. Although it does not employ the Global Average Pooling (GAP) classifier, it benefits from a more sophisticated integration of segmentation and feature extraction. However, its computational efficiency is less optimised compared to models employing GAP classifiers, and it may require longer training times due to the complexity of the architecture.

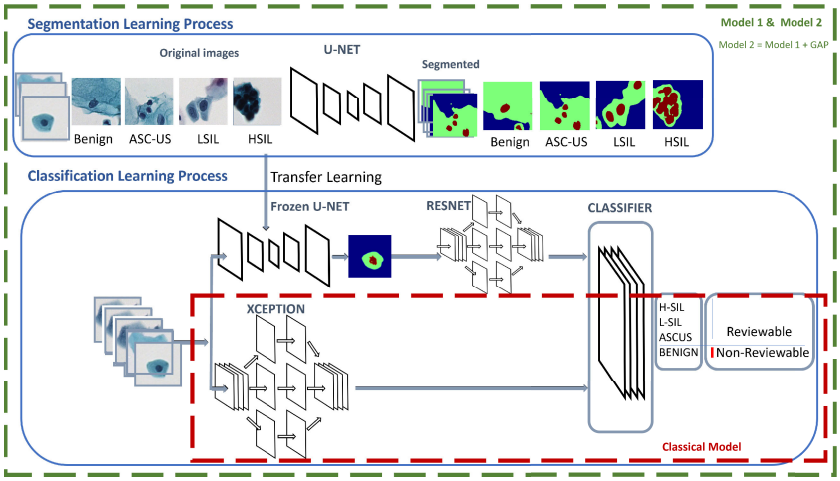


FIGURE 2. Proposed models, where a U-NET is trained with 5% of the samples segmented by hand. Once it learns to segment it is merged with an Xception using a classifier.

The **Proposed Model 2** refines the previous architecture (Proposed Model 1) by incorporating a GAP layer as a classification mechanism. GAP transforms each feature map into a single averaged value, which reduces the dimensionality of the model and improves computational efficiency while preserving spatial information. This strategic design allows efficient processing of complex spatial relationships. By preserving spatial data without flattening the feature maps, GAP minimizes the risk of overfitting and speeds up the training process. The ability to process both segmentation and feature extraction in parallel, combined with the computational efficiency provided by GAP, makes the proposed model 2 a very effective choice for large-scale cellular image classification tasks. In addition, its ability to achieve high accuracy with computational efficiency is a significant advantage in high-throughput environments.

All models were trained with a batch size of 32 for a maximum of 500 epochs, using EarlyStopping with a patience of 20 epochs and monitoring the validation loss to restore the best weights. The Adam optimizer was used with an initial learning rate of $1e$–4, and the loss function was categorical crossentropy, except for the U–Net, which used a custom combination of Dice Loss (class–weighted) and Categorical Focal Loss ($\alpha = 0.25$, $\gamma = 2.0$). Models that incorporated pretrained weights (Xception and ResNet50) used imagenet, freezing the first 40 layers. Throughout, a validation split of 20% was used. In summary, the classical model offers a robust and efficient model using an Imagenet-trained Xception in a classical classification approach. The proposed model 1, offers parallel segmentation and classification by incorporating a U-Net and ResNet, improving the classical model. Finally, the proposed model 2, optimizes the previous model by incorporating a GAP layer to reduce dimensionality and improve computational efficiency, making it ideal for high-accuracy and large-scale applications.

D. METRICS

In order to thoroughly evaluate and compare the performance of the models, a set of key metrics was chosen, each offering unique insights into the model's strengths and potential limitations. These metrics focus on different aspects of classification performance, including general accuracy, class-specific performance, and the balance between precision and recall. The selected metrics are as follows:

- **Accuracy**: It is the proportion of the total number of correct predictions against the total number of predictions, and is calculated as Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$, where TP, TN, FP and FN are True Positive, True Negative, False Positive and False Negative, respectively. where TP, TN, FP and FN, are True Positive, True Negative, False Positive and False Negative, respectively.
- **Precision**: It is the proportion of the number of elements that truly belong to a class among the total number of elements of that class, including the model's predictions. It is calculated as Precision $= \frac{TP}{TP+FP}$.
- **Recall** (Sensibility): Also known as True Positive Rate (TPR), measures the degree to which the model correctly identifies the elements of a class. It is calculated as Recall $= \frac{TP}{TP+FN}$.
- **F1 Score**: Combines precision and recall to obtain an overall performance metric of the model. It is calculated as F1 Score $= 2 \times \frac{Precision \times Recall}{Precision+Recall}$.
- **Area Under the ROC Curve (AUC)**: Represents the probability that the model, given a randomly chosen positive and negative example, will rank the positive higher than the negative. It is calculated as the area under the ROC curve, which plots the TPR against the False Positive Rate ($1 - recall$).

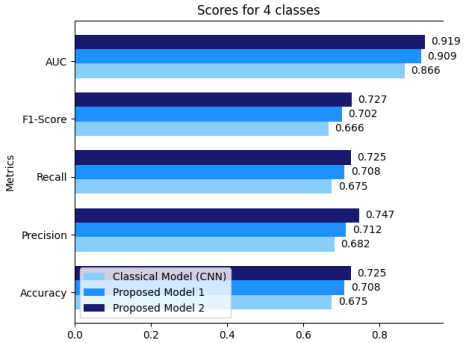
IV. EVALUATION AND DISCUSSION

This section shows a comparison between the three proposed models. The metrics defined in section III-D are used in

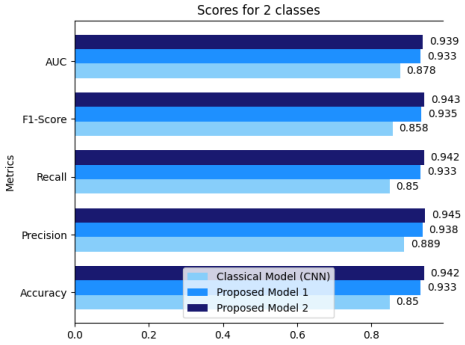
this comparison. The comparison was made from two points of view. On the one hand, a binary classification into Revisable and Not Revisable cells by the pathologist was considered. On the other hand, a quaternary classification has been considered, classifying into 4 classes corresponding to benign, ASC-US, L-SIL and H-SIL. The classification into two classes is useful on a practical level since it serves as a filter for the pathologist, so that he only reviews the cytologies that have or may have dysplasia (ASC-US, L-SIL or H-SIL), since it will determine that these patients have a closer clinical follow-up.

Figure 3.a shows the comparison of different metrics in 4-class models. The proposed models (1 and 2) consistently outperform the classical model (CNN Xception) in all metrics: accuracy, precision, recall, and F1-score. Specifically, Proposed Model 2 achieved the highest scores, including an accuracy of 0.725 and an F1-score of 0.727. Proposed Model 1, while slightly lower in performance than Proposed Model 2, still achieves a commendable accuracy of 0.708 and F1-score of 0.702. These results suggest that the enhancements in the proposed models lead to improved multi-class classification, particularly in distinguishing among complex categories where there is a risk of overlapping features. The Classical Model, with an accuracy of 0.675 and F1-score of 0.667, shows relatively lower precision and recall, which could lead to higher false positives or negatives in more ambiguous cases. Comparing the results of the quaternary classification using the Wilcoxon test, a p-value of 0.1 and 0.05 is obtained in comparison of the classical modelwith the proposed models 1 and 2, respectively. This indicates that the Classical modelis significantly worse than model 2 at a 95% confidence level and model 1 at a 90% confidence level. However, the Wilcoxon statistical test indicates that there is no significant difference between the proposed models 1 and 2 for classification into 4 classes. Although Models 1 and 2 have no significant differences between them, the best selected model must be Model 2, considering that it has slightly higher metrics than Model 1, but it is also computationally more efficient due to the GAP layer. Thus, in computational time, we see that the classical modeltakes 0.083 seconds to infer an image, model 1 0.1447 seconds and model 2 0.1241 seconds. These times have been taken in a GPU T4, 8GB RAM.

For the simpler binary classification task (2-class, Figure 3.b), the differences between the models remain evident, and both proposed models achieve high precision and balanced performance. The proposed model 2 reaches an impressive accuracy of 0.942 and an F1-score of 0.943, suggesting strong reliability in binary classification tasks where precise differentiation is crucial. Proposed Model 1 also performs well, with an accuracy of 0.933 and F1-score of 0.935, making it competitive for applications where binary classification suffices. In comparison, the Classical Model, while decent with an accuracy of 0.850 and F1-score of 0.858, lags behind the proposed models, suggesting that the improvements in Proposed Models 1 and 2 contribute to a better balance between sensitivity and specificity, minimizing misclassifications. This analysis underlines the robustness of the proposed models using segmentation in the binary classification training process compared to a classical model. Comparing the results of the binary classification using the Wilcoxon test, a p-value of 0.03 and 0.01 was obtained in comparison of the classical model with the proposed models 1 and 2, respectively. This indicates that the classical model is significantly worse than the other two models at a 95% confidence level. However, the statistical test indicates that there is no significant difference between the proposed models 1 and 2.



(a) Classification into 4 classes



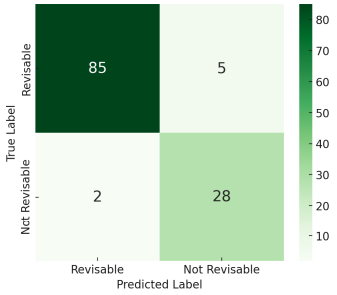
(b) Classification into 2 classes

FIGURE 3. Comparison on different metrics for both four and two class classification of diagnoses on the classical and the proposed model.

The four-class confusion matrix, covering H-SIL, ASC-US, L-SIL, and Benign, reveals that the model is generally effective but faces specific challenges in differentiating between certain categories (Figure 4.a). Notably, the model performs well in identifying Benign cases, with a strong true positive rate, suggesting it is reliable in classifying non-threatening cases and reducing the risk of unnecessary intervention. H-SIL cases also show a reasonable true positive rate.

However, there is a degree of confusion with ASC-US and L-SIL, which may reflect an overlap of nuclear and cytoplasmic characteristics between both categories, which sometimes makes their distinction difficult, also among expert pathologists. The ASC-US category, often characterized by diagnostic ambiguity, is frequently misclassified

(a) Classification into 4 classes



(b) Classification into 2 classes

FIGURE 4. Confusion matrix of the proposed model, where the rows represent the true labels and the columns represent the predicted labels. Each cell shows the number of instances for each combination of true and predicted labels, allowing for the calculation of the model accuracy. The diagonal shows the number of correctly classified instances, while the off-diagonal elements represent classification errors.

as L-SIL or even Benign, which could be concerning from a clinical perspective, as it risks overlooking cases that may need further follow-up. L-SIL, while showing decent recognition, is occasionally confused with ASC-US and Benign, hinting at a need for refinement in feature extraction to improve the model's ability to distinguish these categories.

The proposed model makes more errors in the classification of ASC-US and L-SIL cells, because they represent diagnostic categories with more ambiguous cytologic features; especially ASC-US, which pose a diagnostic challenge for expert pathologists. Their nuclear features sometimes overlap with each other, and factors such as poor cellularity or poor quality of the problem cell (with areas of folding, nuclear overlap or lack of contrast) can make viewing and diagnosis very difficult. In addition, the volume of data for each diagnostic category in 4 classes is very uneven, because L-SIL cells and especially ASC-US cells are very infrequent within the sample, representing only about 5 cells out of about 5000 cells that a cytology may contain.

Overall, this matrix suggests that while the model can broadly differentiate benign from potentially concerning categories, the boundaries among H-SIL, ASC-US, and L-SIL remain challenging, likely requiring additional clinical features or data to improve accuracy. For example, patients with a diagnosis of AS-CUS will require cytology follow-ups, and with the diagnosis of H-SIL will require a close examination of the cervix with a colposcopy.

In contrast, the two-class confusion matrix (Figure 4.b), simplifying the task to Revisable and Not Revisable cases, indicates a high overall performance with minimal misclassification. The model effectively identifies Revisable cases, showing strong sensitivity, and correctly flags a high number of cases that may require revision. The high specificity observed in the Not Revisable category further reinforces the model's ability to reduce false positives, ensuring that non-revisable cases are accurately classified and preventing unnecessary follow-ups. This binary classification task, being less complex than the four-class version, yields a more straightforward and reliable outcome, which may be advantageous for applications where a binary decision suffices.

Multiclass classification is less accurate because the training database is disparate; finding classes such as ASC-US and L-SIL with fewer training cells. A solution could be to increase the collection of cells from the more minority diagnostic categories. In addition, image quality alterations could hinder the model prediction results so some sources employ preprocessed databases to try to improve the results. However, we have used images of cells in real conditions without preprocessing, because they fit the real diagnostic scenario faced by the pathologist.

When we compare the results obtained in the classification in 4 categories with those obtained in 2 classes, the results improve significantly, especially in relation to ASC-US and L-SIL, since both are included together with H-SIL in the same category: ''reviewable''. This type of simplified classification, in addition to improving the results, is considered a more practical classification because in the practice of cervical cancer screening, the aim is to detect patients with a potential risk of developing cancer (the ''reviewable'' category with ASC-US, L-SIL and H-SIL) in order to follow them closely. However, the simplicity of this classification model may overlook the nuances between categories such as H-SIL, ASC-US, and L-SIL, which could hold differing clinical implications. Therefore, while the two-class model is robust and practical for general classification, the four-class model provides a more nuanced perspective but requires further refinement to achieve a similar level of reliability. By accurately segmenting the relevant regions of the image, U-Net directly influences the performance of the classification model. Effective segmentation allows the system to focus on the most informative structures, reducing the impact of noise and improving feature extraction. In this way, the U-Net acts as a filter that makes it easier for the classifier to analyse key visual patterns for decision making, resulting in improved accuracy. When the segmentation provided by the U-Net is not used, it is observed that the classifier is forced to interpret images without an aid of the presented morphology, which hinders its ability to correctly identify classes. Therefore, the inclusion of U-Net brings added value to the classification process.

V. CONCLUSION AND FUTURE WORK

This article proposes the use of CNNs to develop more advanced models of interpretability and to create a model

that works on both the detection and classification of cells through images, as well as the incorporation of data interpretability using segmentation techniques. The use of pre-existing Convolutional Neural Network (CNN) models allows for the classification of cervical cancer cells and the use of these metrics as a benchmark. Incorporating interpretability techniques into segmentation models not only helps explain the results obtained but also facilitates clinical validation and fosters trust among medical professionals in the use of these tools. In this way, models segment more transparently, allowing clinicians to visualize the key areas on which the model based its decision. This aspect is especially relevant in education, where students in fields related to computer science and biomedicine can benefit from better understanding the relationship between image processing and its clinical interpretation. Working with real-world data is crucial for enabling models to handle patient-derived data with potential pathologies, allowing the model to adapt more accurately to human cellular typologies, with varying nucleus and cytoplasm configurations. Working with real data directly involves complicated situations where cells may be overlapping and not clearly differentiable. However, the proposed approach of combining automatic segmentation models with image classification models has obtained satisfactory results, obtaining interpretable data for clinicians, as well as classifications with an F1 score value of 0.935. Finally, future research should focus on refining the accuracy of segmentation models by developing composite and fused models, different from those explored in this study, which integrate multiple classification and segmentation approaches, such as the modified Double U-Net, Swin-Unet, and Hybrid Transformer-CNN architectures. This would further enhance performance and interpretability, solidifying the role of artificial intelligence as a fundamental support in medical practice, developing diagnostic helping applications, and in training future professionals and students, using these tools for training and examining them. In addition, the possibility of increasing the dataset for the more complicated L-SIL and ASC-US classes to improve multiclass classification should also be explored. Finally, the models proposed will be tested with others datasets and the used dataset tested with new models.

DECLARATIONS

ETHICAL APPROVAL

The protocol granted ethical approval by the Ethics Committee at Catholic University of Murcia, code CE012005.

AUTHORS' CONTRIBUTIONS

Conceptualization: Andrés Bueno-Crespo, Raquel Martínez-España, Ana Ortiz-González, and Juan Morales-García; Methodology: Baldomero Imbernón and Juan Pedro Martínez-Cendán; Software: Andrés Bueno-Crespo and Oscar David Romero; Validation: Baldomero Imbernón, Raquel Martínez-España, and Juan Morales-García; Formal Analysis: Ana Ortiz-González, Raquel Martínez-España, and Mauricio A. Álvarez; Investigation: Andrés Bueno-Crespo, Baldomero Imbernón, Ana Ortiz-González, and Oscar David Romero; Writing—Original Draft Preparation: Andrés Bueno-Crespo, Ana Ortiz-González, Raquel Martínez-España, Juan Morales-García, and José Martínez-Más; Writing—Review And Editing: Andrés Bueno-Crespo, Ana Ortiz-González, Raquel Martínez-España, Juan Morales-García, José Martínez-Más, Mauricio A. Álvarez, and Juan Pedro Martínez-Cendán; Visualization: Juan Morales-García, Andrés Bueno-Crespo, Raquel Martínez-España, and José Martínez-Más; Supervision: Mauricio A. Álvarez and Juan Pedro Martínez-Cendán; Project Administration: Andrés Bueno-Crespo; Funding Acquisition: Andrés Bueno-Crespo. All authors have read and agreed to the published version of the manuscript.

DATASET

The dataset is available upon request to the authors and subject to reasonable conditions.

REFERENCES

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 71, no. 3, pp. 209–249, Feb. 2021, doi: 10.3322/caac.21660.

[2] National Cancer Institute. (Apr. 2023). *Cervical Cancer Prognosis and Survival Rates*. Accessed: Jun. 9, 2023. [Online]. Available: https://www.cancer.gov/types/cervical/survival

[3] World Health Organization. (Feb. 2022). *Cervical Cancer—Who.int*. Accessed: Jun. 9, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cervical-cancer

[4] J. Lei, A. Ploner, K. M. Elfström, J. Wang, A. Roth, F. Fang, K. Sundström, J. Dillner, and P. Sparén, "HPV vaccination and the risk of invasive cervical cancer," *New England J. Med.*, vol. 383, no. 14, pp. 1340–1348, Oct. 2020, doi: 10.1056/nejmoa1917338.

[5] National Cancer Institute. (Apr. 2023). *Cervical Cancer Screening—Cancer.gov*. Accessed: Jun. 9, 2023. [Online]. Available: https://www.cancer.gov/types/cervical/screening

[6] A. A. Reshi, F. Rustam, A. Mehmood, A. Alhossan, Z. Alrabiah, A. Ahmad, H. Alsuwailem, and G. S. Choi, "An efficient CNN model for COVID-19 disease detection based on X-ray image classification," *Complexity*, vol. 2021, no. 1, May 2021, Art. no. 6621607, doi: 10.1155/2021/6621607.

[7] S. Kugunavar and C. J. Prabhakar, "Convolutional neural networks for the diagnosis and prognosis of the coronavirus disease pandemic," *Vis. Comput. for Ind., Biomed.*, vol. 4, no. 1, p. 12, May 2021.

[8] C. Ozdemir, "Classification of brain tumors from MR images using a new CNN architecture," *Traitement du Signal*, vol. 40, no. 2, pp. 611–618, Apr. 2023.

[9] M. B. T. Noor, N. Z. Zenia, M. S. Kaiser, S. A. Mamun, and M. Mahmud, "Application of deep learning in detecting neurological disorders from magnetic resonance images: A survey on the detection of Alzheimer's disease, Parkinson's disease and schizophrenia," *Brain Informat.*, vol. 7, no. 1, pp. 1–21, Dec. 2020.

[10] C. Ozdemir and Y. Dogan, "Advancing early diagnosis of Alzheimer's disease with next-generation deep learning methods," *Biomed. Signal Process. Control*, vol. 96, Oct. 2024, Art. no. 106614.

[11] B. Taha, J. Dias, and N. Werghi, "Classification of cervical-cancer using pap-smear images: A convolutional neural network approach," in *Proc. 21st Annu. Conf. Med. Image Understand. Anal.* Cham, Switzerland: Springer, Jan. 2017, pp. 261–272.

[12] J. Martínez-Más, A. Bueno-Crespo, R. Martínez-España, M. Remezal-Solano, A. Ortiz-González, S. Ortiz-Reina, and J.-P. Martínez-Cendán, "Classifying papanicolaou cervical smears through a cell merger approach by deep learning technique," *Expert Syst. Appl.*, vol. 160, Dec. 2020, Art. no. 113707.

[13] H. Hakkoum, I. Abnane, and A. Idri, "Interpretability in the medical field: A systematic mapping and review study," *Appl. Soft Comput.*, vol. 117, Mar. 2022, Art. no. 108391. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494621011522
[14] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, "A survey on the interpretability of deep learning in medical diagnosis," *Multimedia Syst.*, vol. 28, no. 6, pp. 2335–2355, Jun. 2022.
[15] A. Farrag, G. Gad, Z. M. Fadlullah, M. M. Fouda, and M. Alsabaan, "An explainable AI system for medical image segmentation with preserved local resolution: Mammogram tumor segmentation," *IEEE Access*, vol. 11, pp. 125543–125561, 2023.
[16] J. Shao, S. Chen, J. Zhou, H. Zhu, Z. Wang, and M. Brown, "Application of U-Net and optimized clustering in medical image segmentation: A review," *Comput. Model. Eng. Sci.*, vol. 136, no. 3, pp. 2173–2219, 2023.
[17] A. Rai, "Explainable AI: From black box to glass box," *J. Acad. Marketing Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020.
[18] R. Gipiškis, C.-W. Tsai, and O. Kurasova, "Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey," *ICT Exp.*, vol. 10, no. 6, pp. 1331–1354, Dec. 2024.
[19] I. Qureshi, J. Yan, Q. Abbas, K. Shaheed, A. B. Riaz, A. Wahid, M. W. J. Khan, and P. Szczuko, "Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends," *Inf. Fusion*, vol. 90, pp. 316–352, Feb. 2023.
[20] R. Agrawal, M. Sharma, and B. K. Singh, "Segmentation of brain lesions in MRI and CT scan images: A hybrid approach using k-means clustering and image morphology," *J. Inst. Eng., India, B*, vol. 99, no. 2, pp. 173–180, Apr. 2018.
[21] B. Liang, C. Tang, W. Zhang, M. Xu, and T. Wu, "N-net: An Unet architecture with dual encoder for medical image segmentation," *Signal, Image Video Process.*, vol. 17, no. 6, pp. 3073–3081, Sep. 2023.
[22] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-net: A deep convolutional neural network for medical image segmentation," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 558–564.
[23] A. Srivastava, D. Jha, S. Chanda, U. Pal, H. D. Johansen, D. Johansen, M. A. Riegler, S. Ali, and P. Halvorsen, "MSRF-net: A multi-scale residual fusion network for biomedical image segmentation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 5, pp. 2252–2263, May 2022.
[24] S. D. Deb and R. K. Jha, "Modified double U-Net architecture for medical image segmentation," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 7, no. 2, pp. 151–162, Feb. 2023.
[25] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-unet: Unet-like pure transformer for medical image segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2021, pp. 205–218.
[26] H. Zhang, J. Lian, Z. Yi, R. Wu, X. Lu, P. Ma, and Y. Ma, "HAU-net: Hybrid CNN-transformer for breast ultrasound image segmentation," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105427.
[27] Y. Gao, J. Zhang, S. Wei, and Z. Li, "PFormer: An efficient CNN-transformer hybrid network with content-driven P-attention for 3D medical image segmentation," *Biomed. Signal Process. Control*, vol. 101, Mar. 2025, Art. no. 107154.
[28] M. Fang, B. Liao, X. Lei, and F.-X. Wu, "A systematic review on deep learning based methods for cervical cell image analysis," *Neurocomputing*, vol. 610, Dec. 2024, Art. no. 128630.
[29] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of U-Net," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10076–10095, Dec. 2024.
[30] G. Sun, Y. Pan, W. Kong, Z. Xu, J. Ma, T. Racharak, L.-M. Nguyen, and J. Xin, "DA-TransUNet: Integrating spatial and channel dual attention with transformer U-net for medical image segmentation," *Frontiers Bioengineering Biotechnol.*, vol. 12, May 2024, Art. no. 1398237.
[31] A. A. Taddese, B. C. Tilahun, T. Awoke, A. Atnafu, A. Mamuye, and S. A. Mengiste, "Deep-learning models for image-based gynecological cancer diagnosis: A systematic review and meta- analysis," *Frontiers Oncol.*, vol. 13, Jan. 2024, Art. no. 1216326.
[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *18th Int. Conf. Med. image Comput. Comput.-Assist. Intervent.–MICCAI*, Munich, Germany. Cham, Switzerland: Springer, Jan. 2015, pp. 234–241.

ANA ORTIZ-GONZÁLEZ received the degree in medicine from the University of Murcia. She is a Specialist in pathology. She works in the pathology service of the University Hospital Complex of Cartagena, with special interests in dermatopathology, gynecological pathology, and cytopathology. She is a member of Spanish Society of Pathology and European Society of Pathology. She is a member of the editorial team of the journal *Archivos de Patología*.

RAQUEL MARTíNEZ-ESPAÑA received the B.S., M.S., and Ph.D. degrees in computer science from the University of Murcia, Spain, in 2009, 2010, and 2014, respectively. She is an Associate Professor and a Researcher with the Information and Communications Engineering Department, University of Murcia. Her current research interests include ML and DL, pattern recognition in medical problems, and IoT applications.

JUAN MORALES-GARCíA received the B.S., M.S., and Ph.D. degrees in computer science from the Catholic University of Murcia, in 2020, 2021, and 2023, respectively. He is an Assistant Professor and a Researcher with the Department of Software and Computing Systems, University of Alicante, Spain. His current research interests include ML and DL, edge computing, the IoT, and pervasive computing.

BALDOMERO IMBERNÓN received the B.Sc. and Ph.D. degrees in computer engineering from the Catholic University of Murcia, Spain, in 2012 and 2018, respectively. He is an Associate Professor in computer science engineering with the Catholic University of Murcia. His research focuses on high performance computing and optimization algorithms in GPU.

JOSÉ MARTíNEZ-MÁS received the degree in medicine from the University of Murcia, in 2011, and the Ph.D. degree in health science from the Catholic University of Murcia, in 2020. He is a Specialist in obstetrics and gynecology. He is the Medical Director with the CIAGO Gynecological Center, and a Lecturer with the Faculty of Medicine, Catholic University of Murcia. His current research interest includes ML and DL applied to health sciences.

MAURICIO A. ÁLVAREZ received the B.Eng. degree (Hons.) in electronics engineering from the Universidad Nacional de Colombia, in 2004, the M.Eng. degree in electrical engineering from the Universidad Tecnológica de Pereira, Colombia, in 2006, and the Ph.D. degree in computer science from The University of Manchester (UoM), U.K., in 2011. He was a Lecturer with The University of Sheffield, U.K. He is currently a Senior Lecturer in machine learning with the Department of Computer Science, UoM.

OSCAR DAVID ROMERO received the B.S. degree in computer science from the Catholic University of Murcia, Spain, in 2023. He combines his research with his work as a Computer Engineer, developing and implementing ML and DL solutions for enterprise environments, focusing on computer vision and natural language processing services.

JUAN PEDRO MARTíNEZ-CENDÁN received the degree in medicine and surgery and the Ph.D. degree in medicine from the University of Murcia. He is a Specialist in gynecology and obstetrics. He combines his work in healthcare with teaching on the degree in medicine with the Catholic University of Murcia, where he directs the master's degree in bioethics. He is also a Tutor Resident with the Hospital Universitario General Santa Lucía.

ANDRÉS BUENO-CRESPO received the B.Sc. degree in computer engineering from the University of Malaga, and the Ph.D. degree from the Polytechnic University of Cartagena, Spain, in 2013. He is an Associate Professor in computer science and the Coordinator of the distance learning degree in computer science engineering with the Catholic University of Murcia, Spain. His research focuses on pattern recognition in medical images and DL techniques for IoT infrastructures.