

Doctoral dissertation review

Betelhem Zewdu Wubineh, M.Sc.

titled

*Segmentation and Classification of Pap Smear Images to Detect Cervical Cancer
using Deep Learning*

This review of the doctoral dissertation, authored by Ms Betelhem Zewdu Wubineh, M.Sc., was commissioned by the Chairman of the Disciplinary Council for Information and Communication Technology, Prof. Wojciech Bożejko (Letter No. W4N-KP-SW.4110.7.2026 dated 25 June 2026). The supervisor of this dissertation is Prof. Andrzej Rusiecki, and the auxiliary supervisor is Dr Krzysztof Halawa. The doctoral dissertation is written in English.

1. The research problem and its significance

The main problem addressed in the dissertation is the development and evaluation of deep learning methods for the automatic analysis of cervical cytology images. Particular attention is given to the segmentation of cellular structures and the classification of cervical cells. This problem is considered in the context of supporting cervical cancer diagnosis, where sufficiently early detection is important for effective treatment. Conventional assessment of cytological smears requires appropriately qualified specialists, is time-consuming, and is affected by subjective interpretation and inter-observer variability, which may result in different assessments of the same case by different diagnosticians. Therefore, the automation of cytological image analysis using artificial intelligence methods represents a natural direction in the development of computer-aided diagnosis.

From the perspective of the discipline of Information and Communication Technology, the main problem of the dissertation is not medical diagnosis itself, but rather computer image processing and analysis, machine learning, and the design, training, and evaluation of deep neural networks for cytological images. These images have specific properties, as they are characterized by complex cell morphology, frequent overlapping structures, staining artifacts, and considerable visual similarity between different classes. An additional problem is the limited availability of properly annotated data and frequent class imbalance, which makes it more difficult to train models with a large number of parameters and increases the risk of overfitting. The dissertation is therefore strongly related to modern computer vision and deep learning, while the medical domain provides a challenging environment for the evaluation of the developed methods.

The problem considered in the dissertation is divided into several related research areas. The first concerns the segmentation of cytological images, i.e., the automatic extraction of relevant cellular structures, in particular the nucleus and cytoplasm, and the determination of which deep neural network architectures and their modifications can provide high-quality segmentation. The second

WPLYNĘŁO

07-10-2026

RDN-III/118/2026

area concerns the classification of cervical cells, including the discrimination between normal and cancer cells as well as more detailed cytological classes. This requires the development of feature representations capable of distinguishing often subtle morphological differences in cellular images. The third problem concerns the limited availability and imbalance of training data and investigates whether synthetically generated images can increase data diversity, improve model training, and enhance generalization. These three areas are directly reflected in the research questions formulated in the dissertation, concerning segmentation, classification, and generative data augmentation, respectively.

An important extension of the problem formulated in this way is the integration of segmentation and classification. The Candidate investigates whether the information obtained during the segmentation of cellular structures can provide useful input for the subsequent classification stage and to what extent segmentation quality affects the final classifier decision. Both a two-stage approach, in which the segmentation result is passed to the classifier, and a multi-task learning approach, in which a single model based on a shared feature representation is trained simultaneously for image segmentation and classification, are considered. Thus, the dissertation addresses not only two separate image analysis tasks, but also the problem of organizing a complex processing pipeline and jointly learning related tasks. This scope is consistent with the research areas declared by the Candidate: data augmentation, segmentation, classification, and an integrated segmentation–classification approach.

The problem considered in the dissertation is undoubtedly scientific in nature. The work is not limited to the implementation of existing deep learning methods for a specific dataset, but includes the formulation and experimental verification of questions concerning the properties and effectiveness of particular classes of computer image analysis methods. The Candidate addresses the problem of improving the quality of cytological image segmentation and classification, the influence of modifications to deep neural network architectures on the obtained results, the possibility of using artificially generated images to improve model training, and the evaluation of the relationship between segmentation and subsequent classification.

Chapter 1 defines both the main objective and the specific objectives of the research. The main objective is to develop and improve deep learning methods for the segmentation and classification of cytological images. The specific objectives include the development of data augmentation methods, the development of methods for the segmentation of cellular structures, the development of classification models using, among others, attention mechanisms and ensemble learning, as well as the integration of segmentation and classification and the analysis of the influence of segmentation quality on classification performance.

These objectives are further formulated as three research questions and corresponding hypotheses. The first question concerns the effectiveness of deep learning methods for nucleus and cytoplasm segmentation and the possibility of improving the results through appropriate modifications of neural network architectures. The second concerns the possibility of improving cervical cell classification performance through the development of feature representations and modifications of classification models. The third concerns the effect of artificially generated images on the performance and generalization ability of classification models and the comparison of different data

generation methods. The hypotheses assume, respectively, that segmentation performance can be improved through architectural modifications, classification performance can be improved through better feature representation, and classification results can be improved by increasing data diversity and reducing class imbalance using synthetic images.

The research questions formulated in this way can be experimentally verified, which the Candidate does in the subsequent parts of the dissertation by designing and modifying deep neural network architectures, training them on multiple datasets, and quantitatively evaluating the results using standard segmentation and classification metrics. An important part of the research also includes comparisons with baseline methods, experiments using different data splitting strategies, cross-validation, and, in selected cases, cross-dataset validation. Such an experimental methodology makes it possible to empirically verify the adopted assumptions and hypotheses.

A certain weakness is the rather general formulation of some of the research hypotheses. For example, the assumption that an architectural modification may improve model performance compared with a baseline solution does not precisely define the expected mechanism or the conditions under which the improvement should occur. Similarly, the third research question aims to determine which image generation method is the most effective, while the corresponding hypothesis does not indicate a specific class of methods that is expected to perform best. This does not undermine the scientific nature of the dissertation, but indicates that the research questions and hypotheses could have been formulated more precisely.

In addition to its scientific character, the doctoral dissertation also has significant practical relevance. The methods developed by the Candidate for the segmentation and classification of cytological images are intended to support the detection of cervical cancer-related lesions. Their potential application is in computer-aided diagnosis systems, in which artificial intelligence methods could support specialists by automatically extracting relevant cellular structures, classifying cells, and identifying cases requiring particular attention. However, the dissertation does not present a complete system ready for direct clinical use. The work focuses on the development and experimental evaluation of methods that may become components of such a system. This is an appropriate approach considering the scope of the dissertation, especially since the Candidate points out the need for further validation on larger, multi-center datasets and for direct evaluation by medical specialists before the developed solutions can be used in clinical practice.

From a practical perspective, the robustness of the models to changes in the data source is particularly important. Images originating from different datasets may differ in sample preparation, staining, resolution, imaging device characteristics, and population characteristics. The Candidate conducted experiments in which a model trained on one dataset was evaluated on another dataset. For example, the hybrid ViT model trained on the SiPaKMeD dataset was tested on the Pomeranian dataset, and the best variant using the BBox region achieved an accuracy of 93.79%. In another experiment, training was performed on SiPaKMeD and testing on Herlev. Such experiments are of particular practical importance because they evaluate whether the model remains useful under a change in data distribution, rather than only for images originating from the same dataset as the training data.

Another interesting element from the perspective of future applications is the use of segmentation results to restrict the image region analyzed by the classifier. The Candidate investigates how regions containing a segmented cell together with a small area of its surroundings can provide better classification performance. This suggests that it may be beneficial to reduce the influence of irrelevant background while preserving some contextual information surrounding the cell. This may be important when transferring models between image datasets with different characteristics.

The broader practical relevance of the research is also supported by experiments performed on other types of medical imaging. The segmentation methods were evaluated on the HAM10000 dermoscopic dataset of skin lesions, while the proposed classification approach based on the residual ensemble was tested on brain tumor MRI images. The results of these experiments indicate that some of the developed solutions are not strictly limited to cervical cytology and may provide a basis for applications in other medical image analysis tasks.

The practical relevance of the dissertation should therefore be assessed highly. At the present stage, however, it mainly concerns the development and validation of methods that may in the future become components of computer-aided diagnosis systems, rather than the presentation of a complete clinical solution.

2. Author's contribution

In Section 1.5, the Candidate divides the declared contribution of the dissertation into two main groups: image segmentation methods and classification methods. The scope of the presented contributions is very broad and includes modifications of existing deep neural network architectures, development of hybrid architectures, application of attention mechanisms, ensemble learning methods and data augmentation using generative models, as well as mechanisms for integrating segmentation and classification.

In the area of segmentation, the Candidate declares, among others, the modification of the U-Net architecture by using Group Normalization and dilated convolutions, the development of the CelluVGG19 variant using separable convolutions and residual blocks, the development of SPP-SegNet combining SegNet with multi-scale processing mechanisms, the modification of DeepLabV3+ by introducing Squeeze-and-Excitation blocks, and the development of a hybrid ViT-U-Net approach. An important element of this part of the dissertation is also the proposed uncertainty-aware multi-task learning (UAMT) model, which combines segmentation, classification, and prediction uncertainty estimation within a single architecture.

The above solutions do not have the same level of methodological originality. A significant part of them is based on combining previously known components and mechanisms in new configurations adapted to the specific characteristics of cytological images. Examples include the use of SPP in SegNet, SE blocks in DeepLabV3+, and the combination of transformer mechanisms with U-Net. However, I do not consider this to be a significant weakness of the Candidate's contribution. In deep learning research, scientific value may also result from properly designing configurations of existing mechanisms, adapting them to a specific class of problems, and carrying out broad and systematic experimental validation. I consider the UAMT model to be particularly interesting in this group. In

this model, classification and segmentation are trained jointly, while the model additionally estimates prediction uncertainty. This solution goes beyond the classical sequential segmentation–classification scheme and is one of the more interesting elements of the dissertation.

The contribution in the area of classification has a similar character. The Candidate developed and investigated a number of variants using, among others, a combination of VGG19 and elements of the AlexNet architecture, spatial and channel attention mechanisms, integration of ResNet50V2 with SE blocks and a Vision Transformer, multi-scale feature fusion, and different variants of ensemble learning. The studies on ensemble learning are particularly extensive and include both models with learnable weights and the residual ensemble approach, in which features extracted independently by multiple pretrained networks are further processed and fused.

The Candidate’s contribution also includes studies on generative data augmentation, in particular the development of the RES_DCGAN variant by introducing residual blocks into the generator, as well as broad comparisons of the influence of data generated using GAN, DCGAN, RES_DCGAN, and a diffusion model on subsequent classification performance. Also in this case, in my opinion, the main value does not come only from a single modification of the generator architecture, but from conducting a broad analysis that makes it possible to determine in which configurations and for which datasets synthetic enlargement of the training set actually provides benefits.

I consider the most important contribution of the Candidate not to be a single modification of a particular architecture, but the implementation of an extensive research program covering the whole process of cytological image analysis. In this context, the studies on using segmentation results as input data for classification are particularly valuable. The Candidate compares classification using whole images, segmented ROI regions, and BBox regions preserving a limited amount of context around the cell. This makes it possible not only to propose further models, but also to answer a more general question of whether segmentation is actually needed for effective classification and how its result should be used. I consider the results showing the benefits of preserving a small amount of surrounding context around the segmented cell to be interesting from both the scientific and potential application perspectives.

An important part of the contribution is also the evaluation of the ability of the models to operate after changing the dataset. The dissertation includes experiments in which a model trained on the SiPaKMeD dataset was subsequently tested on data from the Herlev or Pomeranian datasets. This is particularly valuable in the context of medical imaging, because differences between datasets may result from different procedures of sample preparation, staining, and image acquisition. In the SiPaKMeD–Pomeranian experiment, the best BBox-based variant achieved 93.79% accuracy. I assess this type of experiment more highly than another comparison of architectures performed only on a random split of a single dataset, because it directly addresses one of the fundamental problems of modern machine learning in medical imaging, namely model robustness to domain shift.

The scale of the experimental work deserves particular emphasis. The Candidate did not limit the research to a single architecture or a single dataset. The Candidate evaluated many variants of segmentation and classification models, different attention mechanisms, ensemble learning and augmentation strategies, different input data representations, and different validation approaches.

The studies included several cytological datasets, while the possibility of broader application of the developed methods was also evaluated on the HAM10000 dermoscopic dataset and on MRI images of brain tumors. In the case of HAM10000, the possibility of using the segmentation methods in another imaging modality was evaluated, while for MRI data the proposed classification methods were tested. Regardless of the comments concerning the design of some experiments, which I present later in this review, the scope of the performed work demonstrates a very large research effort and good knowledge of modern deep learning methods.

In the context of such an extensive methodological contribution, the lack of a reference in the dissertation to a public repository containing implementations of the proposed models and code allowing the experiments to be reproduced is noticeable. Some of the developed solutions have a complex structure and are created by integrating many components, as in the case of hybrid models, extensive network ensembles, pipelines combining segmentation and classification, and the UAMT model. The methodology chapter itself includes many independent segmentation, classification, generative, and integrated pipeline architectures. Reproducing them precisely only on the basis of the descriptions, diagrams, and parameters presented in the dissertation may be time-consuming for other researchers and, in some cases, leaves room for different implementation decisions. Providing the source code, experiment configurations, and definitions of the proposed architectures would significantly improve the reproducibility of the results and the practical scientific value of the work. It would also allow other research teams to use the proposed models more easily as reference solutions and compare their own methods with them. I do not consider the lack of such a repository to undermine the presented contribution, but considering the scale and complexity of the research, making it available would be a very valuable addition to the dissertation.

The Candidate's contribution is also confirmed by a very extensive publication record directly related to the topic of the dissertation. The list of publications includes 14 papers related to the dissertation, including six journal articles and eight conference publications. The publication venues include, among others, Scientific Reports, Cancers, IEEE Access, and Neural Computing and Applications. The publications include both experimental papers presenting solutions later developed in the dissertation and review papers concerning segmentation, classification, and data augmentation in cervical cancer diagnosis. The conference publications cover, among others, cytological image segmentation, ViT-SE_Res models, DenseNet201 with a spatial attention mechanism, the VGG19-AlexNet hybrid model, and DCGAN-based augmentation. Such a large number of publications directly related to the dissertation provides additional confirmation of the Candidate's research activity and indicates that a significant part of the presented results was systematically evaluated by the scientific community.

In summary, I assess the Candidate's contribution to the development of the discipline of Information and Communication Technology positively. I do not consider its main value to be the development of one fundamentally new neural network architecture. In my opinion, the most valuable aspects are the creative adaptation and integration of modern deep learning methods for the challenging problem of cytological image analysis, very broad experimental validation, studies on the relationship between segmentation and classification, cross-dataset experiments, and addressing the problem of joint learning of segmentation, classification, and uncertainty estimation. Taken

together, these elements constitute a significant and well-documented contribution to the development of methods for computer analysis of medical images.

3. Correctness

I assess the research presented in the dissertation as generally methodologically correct, extensive, and conducted in accordance with practices commonly used in modern research on deep learning methods. The Candidate uses several publicly available datasets and one dataset developed by the Candidate, applies standard metrics for the evaluation of segmentation and classification, compares the proposed solutions with baseline models and results reported in the literature, and in some experiments uses five-fold cross-validation and statistical analysis methods. Cross-dataset experiments, in which models trained on one dataset are evaluated on another dataset, are also a valuable element of the work, as are attempts to transfer the developed solutions to other medical imaging modalities.

The very large scale of the experimental program deserves particular emphasis. The Candidate analyzes many segmentation and classification architectures, attention mechanisms, ensemble learning methods, and generative data augmentation approaches, and then combines these elements into more complex segmentation–classification pipelines. In some experiments, the results are presented for different validation schemes, together with mean values and standard deviations, while the Friedman test and the Nemenyi post-hoc test are used for selected comparisons. Particularly valuable are the experiments presented in Chapter 6, in which classification based on original images, ROI regions, and BBox regions obtained after segmentation is compared, and external validation between different datasets is also performed.

I also assess positively the organization of the main methodological part of the dissertation. Chapter 3 describes the datasets used, preprocessing, segmentation and classification models, augmentation methods, pipelines combining segmentation with classification, evaluation metrics, and implementation parameters. The dissertation also provides information about the computational environment, software libraries, and, for many models, detailed training parameters. This does not mean, however, that all experiments are described with the same level of detail. Due to the very large number of analyzed variants, information concerning training procedures, data splits, and experimental configurations is sometimes distributed across different parts of the dissertation, which makes direct comparison more difficult.

In my opinion, the presented results in most cases provide a basis for a positive assessment of the effectiveness of the investigated solutions. More caution is required, however, when interpreting the reasons for the observed improvements and when formulating conclusions concerning generalization, statistical significance, and clinical usefulness. In many cases, the proposed model differs from the baseline solution in several components at the same time. The observed improvement then allows one to conclude that the complete modified configuration performs better, but it does not always allow the improvement to be attributed to a specific component. In such cases, more systematic ablation studies would be desirable.

The most important detailed comments concerning the correctness and interpretation of the results are presented below:

1) Method used to generate ground-truth masks for the Pomeranian dataset.

An important limitation of the segmentation experiments is the way in which the masks for the Pomeranian dataset were prepared. Cytoplasm and nucleus masks were generated automatically in OpenCV using Otsu thresholding and morphological operations, and were then used as pseudo-ground truth without manual verification by experts. This means that high Dice or IoU values in experiments on this dataset primarily indicate agreement between the network predictions and the output of the algorithm used to generate the masks, rather than direct agreement with expert annotations of biological structures. The Candidate also identifies this issue as a limitation in Chapter 7. Therefore, conclusions concerning segmentation quality for the Pomeranian dataset should be interpreted more cautiously than results obtained for datasets with expert-annotated reference masks.

2) Lack of systematic ablation studies for some of the proposed models.

In many experiments, the proposed models differ from the baseline architecture in several components at the same time. This applies, among others, to CelluVGG19, SPP-SegNet, SE-DeepLabV3+, ViT-SE_Res, some ensemble methods, and the UAMT model. In such cases, comparison of the complete configuration with the baseline model demonstrates the benefit of using the whole set of modifications, but it does not make it possible to determine unambiguously which component is responsible for the improvement. For example, in ViT-SE_Res, ResNet50V2, SE blocks, and the transformer part are introduced together, and the obtained results are interpreted as the effect of their combined operation. More systematic ablation experiments would be useful to provide a clearer explanation of the sources of the observed improvement.

3) Non-uniform experimental protocol.

The dissertation uses both a 64%/16%/20% split of image datasets and five-fold cross-validation, as well as, in some cases, other splits resulting from the specific characteristics of individual datasets. The use of different procedures is itself justified, but not all results that are later compared were obtained using the same protocol. A particularly clear example is Section 5.3.1: the results obtained using only real data in Table 5.17 come from 5-fold cross-validation, while the results after augmentation in Table 5.18 were obtained using a single 80/20 split. Therefore, the observed improvement does not represent a fully controlled measurement of the effect of synthetically generated data alone. A better approach would be to evaluate both variants using identical data splits.

4) Inconsistencies in the description of data splitting.

Section 3.8 defines the main protocol as 64% training data, 16% validation data, and 20% test data. However, Section 5.2.1 for the DenseNet201+SAM experiment and Section 5.2.4 for ViT-SE_Res report a 64%/16%/16% split, which sums to 96%. This is most likely an error in the description and 20% of the data were actually used for testing, but this issue requires a clear explanation.

5) Data splitting for datasets containing dependent images.

For some datasets, it is particularly important that the training/test split is performed at the appropriate level. The Mendeley image dataset contains 963 images obtained from 460 patients, while Cervix93 consists of 93 image stacks, each containing approximately 10–20 images of the same field of view acquired at different focal depths. I did not find clear information in the dissertation on whether a patient-level split was used for Mendeley and a stack-level split for Cervix93. This is important because random assignment of strongly related images from the same patient or the same stack to both training and test sets could lead to overestimated performance. This issue is particularly important for Cervix93, for which a classification accuracy of 100% is reported.

6) Unspecified aggregation method for metrics in multi-class classification.

Section 3.7 presents definitions of Accuracy, Precision, Recall, and F1-score for the binary classification problem, while the dissertation also contains many multi-class experiments. I did not find clear information in the dissertation on how these metrics are calculated in the multi-class case. This is particularly important for imbalanced datasets and should be precisely specified.

7) The comparison of generative augmentation methods is partly affected by different numbers of synthetic images.

In Table 3.4, the number of images generated by GAN, DCGAN, and RES_DCGAN is the same, while the number generated by DDPM is considerably smaller due to its higher computational cost. For Pomeranian, the corresponding values are 2664 images for GAN/DCGAN/RES_DCGAN and 1239 for DDPM, and similar differences occur for SIPaKMeD and Herlev. As a result, the experiments in Section 5.3.3 compare not only the type of image generator, but also the size of the augmented training dataset. This makes it difficult to determine whether the difference in performance results only from the quality of a particular generative model. For a strict comparison, the same number of synthetic images should be used for all methods.

8) Limited statistical analysis in some experiments.

I assess positively the use of 5-fold cross-validation in selected experiments, the reporting of mean values and standard deviations, and the use of the Friedman and Nemenyi tests. However, this approach is not applied consistently throughout the dissertation. In many other places, small differences between individual results are interpreted as improvements or as significant results, although an appropriate statistical significance test was not performed.

9) Influence of segmentation on classification performance.

An important conclusion resulting from the conducted experiments is that the influence of segmentation on classification performance is not clearly positive and depends on the way the segmentation result is used, the applied model, and the analyzed dataset. In Table 6.8, the use of BBox improves the result for Pomeranian, SIPaKMeD multiclass, and Herlev, but for SIPaKMeD binary the variant without segmentation achieves a slightly better result. Similarly, comparison of Tables 6.11 and 6.12 shows that in some configurations classification of the full image is equally good or better than classification after segmentation. Therefore, general statements in the dissertation that integration of segmentation and classification “consistently” or “significantly” improves performance are too categorical. A more appropriate conclusion is that the effect of segmentation depends on the

dataset, classifier, and representation of the segmentation result, while particularly interesting results are obtained using BBox, which preserves a limited amount of contextual information.

10) Cross-dataset validation is a valuable element of the work, but the applied processing pipeline requires a more complete description.

I assess very positively the experiments in which a model trained on SIPaKMeD is tested on Herlev or Pomeranian. For the learnable-weight ensemble, an accuracy of 84.71% was obtained for BBox input compared with 82.44% for the original image. For the hybrid ViT in the SIPaKMeD->Pomeranian experiment, the BBox-10 variant achieved 93.79%. However, in the case of the pipeline using segmentation, it is not completely clear whether the segmentation model was also trained only on the source dataset, or whether a model previously trained on data from the target dataset was used to obtain ROI/BBox regions for the target dataset. This distinction is essential for interpreting the experiment as full external validation of the complete system. In the case of SIPaKMeD->Pomeranian, the method used to harmonize classes between datasets with different original labeling schemes should also be clearly described.

11) Validation of the UAMT model is too limited considering the complexity of the proposed solution. UAMT combines a shared ResNet50 encoder, a U-Net segmentation decoder, an attention-guided evidential classifier, prediction uncertainty estimation, and learned weighting of segmentation and classification loss functions. This is one of the most interesting elements of the dissertation, but ablation studies are missing that would compare at least: single-task vs multi-task learning, fixed vs learned loss weights, and standard softmax classification vs evidential classification. The final values of the learned parameters s_1 and s_2 are also not reported, and therefore it is not known what effective weights were assigned to the two tasks. Expected Calibration Error (ECE) was used in the calibration analysis, but insufficient details are provided on how this metric was calculated. In particular, the number of bins, the method of defining them, and the procedure used to aggregate the results are not specified.

12) The interpretation of Explainable AI results is in some places too far-reaching.

Grad-CAM and saliency maps are valuable additions to the analysis and make it possible to identify the image regions on which the model focuses. However, they do not by themselves provide evidence that the regions used by the model are diagnostically correct. Section 3.9 contains statements suggesting that the visualizations “confirm” the diagnostic significance of particular regions and even “guarantee” the clinical significance of the predictions. Such a strong conclusion would require validation by a cytologist or pathologist. On the positive side, it should be noted that in a later part of the dissertation the Candidate recognizes this limitation and points out the lack of expert validation of the attention maps.

13) Comparisons with results reported in the literature.

Tables 4.11, 5.29 and 6.15 are very useful because they make it possible to place the obtained results in the context of the current state of the art. However, not all reported results were obtained using identical data splits, preprocessing procedures, class definitions, and metric aggregation methods. Therefore, small percentage differences between studies cannot be treated as strict evidence that one method is superior to another. These comparisons clearly demonstrate that the Candidate’s results are competitive, but they do not always allow an unambiguous ranking of the methods.

14) Inconsistencies between Precision, Recall, Dice and IoU in Table 4.7 and 4.8.

In Table 4.7, for SPP-SegNet in cytoplasm segmentation, Precision = 92.59% and Recall = 95.75% are reported, while the reported Dice is only 64.25% and IoU = 47.41%. A similar situation occurs for the standard SegNet: Precision = 92.03%, Recall = 91.54%, while Dice = 66.08% and IoU = 49.40%. A similar problem is present in Table 4.8 for the standard DeepLabV3+ on Pomeranian for cytoplasm segmentation: Precision = 87.68%, Recall = 98.13%, while Dice = 78.63% and IoU = 64.98%.

15) F1 values inconsistent with Precision and Recall in Table 5.17 and 6.13.

In the classification results for the real Pomeranian dataset, there are cases in which the reported F1-score is higher than both Precision and Recall. For MobileNet, Precision = 92.52%, Recall = 91.57%, and F1 = 93.66% are reported; for InceptionV3, the corresponding values are 85.71%, 85.54%, and 87.43%; while for ResNet50V2 they are 89.55%, 89.16%, and 90.92%. When F1 is calculated as the harmonic mean of Precision and Recall, such values are not possible. A similar inconsistency occurs in Table 6.13. For the ROI variant, Accuracy = 75.89%, Precision = 79.20%, Recall = 75.89%, while F1-score = 88.34%.

16) Lack of visual presentation of synthetic images generated by the generative models.

Data augmentation using generative models, including GAN, DCGAN, DDPM, and the proposed RES_DCGAN, plays an important role in the dissertation. The Candidate describes the architectures of these models and analyzes their influence on subsequent classification results, and for DCGAN and RES_DCGAN additionally evaluates the quality of the generated data using the FID metric. However, I did not find examples of synthetic images generated by the individual models in the dissertation. I consider this to be an important omission because, in the case of medical image generation, quantitative evaluation alone does not provide a complete picture of the quality of the generated data. Presenting representative examples of real and synthetic images, preferably for different classes and different generators, would make it possible to evaluate the preservation of cell morphological features, the presence of artifacts, the diversity of generated samples, and the visual realism of the images. This would be particularly important in view of the statements made in the dissertation that RES_DCGAN generates more realistic and higher-quality images and that more advanced synthesis methods lead to better training data.

17) Lack of analysis of the impact of the proposed modifications on model computational complexity.

The dissertation proposes many modified deep neural network architectures, in which additional components such as attention blocks, SE modules, transformer elements, multi-scale processing mechanisms, or additional feature fusion paths are added to baseline models. An interesting extension of the conducted research would be an analysis of how the introduced modifications affect the computational complexity of the proposed solutions. For selected pairs of baseline and modified models, it would be possible to compare the number of parameters, estimate the number of computational operations, e.g. FLOPs, and evaluate memory requirements during inference or training. This would make it possible to determine whether the improvement in segmentation or classification performance is associated with a significant increase in computational cost and whether the proposed solutions provide a good trade-off between performance and computational efficiency. Such an analysis would be particularly valuable from the perspective of the potential application of the models in practical computer-aided diagnosis systems.

The above comments do not undermine the overall credibility of the conducted research program. They mainly indicate areas in which the very broad scope of the work was not accompanied by an equally detailed control and statistical analysis. The dissertation contains a large number of correctly conducted experiments, and in many cases the obtained results clearly demonstrate the benefits of the proposed configurations. More caution is required when attempting to attribute these benefits to individual components and when formulating broad conclusions concerning generalization and potential clinical usefulness. Several tables also contain inconsistencies between the reported values of individual performance metrics, which should be verified and explained by the Candidate; however, they do not affect the overall assessment of the very extensive experimental program. Overall, I assess the correctness of the research work positively.

4. Candidate knowledge

The main chapter of the dissertation documenting the Candidate's knowledge of the current state of the art is Chapter 2, "Literature Review". It covers, in sequence, issues related to the application of deep learning in medical image analysis, segmentation and classification of cervical cancer cell images, data augmentation methods, and a review of previous studies concerning the automatic analysis of cytological images. This scope corresponds well to the topic of the dissertation and covers important areas of modern computer science, in particular artificial intelligence and machine learning, deep neural networks, computer image analysis and processing, and the classification and segmentation of medical images.

I assess Chapter 2 positively. The literature review is extensive and is not limited only to presenting individual methods. The Candidate shows the evolution of solutions from classical machine learning methods and hand-crafted features to CNNs, transfer learning, transformer architectures, and hybrid CNN-Transformer models. In the area of segmentation, U-Net architectures and their extensions are discussed, while in the area of classification, modern convolutional networks, pretrained models, and attention mechanisms are analyzed. An important part of the review is also devoted to data augmentation methods, including GANs, their variants, and diffusion models.

Another valuable aspect of this chapter is the identification of open research problems, such as limited model generalization across datasets and different imaging conditions, propagation of segmentation errors to the classification stage, insufficient availability of labeled data, class imbalance, computational complexity of more advanced architectures, and the problem of the reliability of synthetic data. This shows that the Candidate not only knows the methods used in the dissertation, but also understands their limitations and the problems related to their application to real medical data.

The literature review is complemented by Chapter 3, "Methodology", which is mainly intended to present the methods used and developed in the dissertation, but also contains a significant amount of information about existing solutions. It discusses, among others, image data preparation and preprocessing, U-Net, SegNet and DeepLabV3+ architectures, networks using transfer learning, self-attention, spatial and channel attention mechanisms, Squeeze-and-Excitation and CBAM, Vision Transformers, ensemble learning methods, GAN, DCGAN and DDPM, model evaluation methods, and

Explainable AI. This chapter demonstrates the Candidate's good knowledge of both basic and more recent deep learning techniques used in computer image processing.

I assess the bibliography of the dissertation as extensive and, within the scope of the dissertation topic, generally complete. It includes both publications that form the basis of the discussed architectures and methods and many recent studies concerning segmentation, classification, and data augmentation in medical imaging, with particular emphasis on cervical cytology. An important advantage is the large number of publications from the last few years, including recent solutions based on transformers, attention mechanisms, generative models, and modern segmentation and classification methods. This indicates that the Candidate has followed the recent development of the research area in which the work was conducted. The review also refers to papers published in recognized journals and proceedings of reputable conferences.

An additional argument confirming the Candidate's knowledge is the publication record directly related to the topic of the dissertation. The preparation of two systematic literature review papers provides further evidence of a good knowledge of the current state of research.

In summary, I assess the Candidate's knowledge in the discipline of Information and Communication Technology positively. In particular, the dissertation confirms very good knowledge of modern methods of artificial intelligence, deep learning, and computer image analysis, including both fundamental neural network architectures and more advanced solutions based on transfer learning, attention mechanisms, Vision Transformers, ensemble learning, and generative models. The broad literature review, knowledge of current research directions, ability to identify limitations of existing methods, and extensive publication record provide sufficient basis to conclude that the Candidate has appropriate general knowledge in the discipline, with particularly deep knowledge in the areas directly related to the topic of the dissertation.

5. Other notes

From an editorial and language perspective, I assess the dissertation positively. The work has a clear and logical structure and is divided into chapters corresponding to the successive stages of the research. A large number of tables, figures, and diagrams makes it easier to follow the presented methods and results. Despite the considerable length of the dissertation and the very large number of experiments, the presentation of the material remains generally consistent. The language is clear and allows the presented methods and results to be understood without ambiguity.

However, the dissertation contains some individual editorial and language errors and imperfections, which do not significantly affect its overall assessment. The following issues were identified in particular:

- 1) Inconsistent highlighting of the best results in tables. In some tables, including Tables 4.1 and 4.8, the values shown in bold do not always correspond to the best values of a given metric. With a large number of comparisons, this makes it more difficult to quickly compare the results.
- 2) Individual errors in numerical values reported in the text. For example, on page 88, the description of the results reports an accuracy of 964.68%.

3) Errors in numerical values reported in tables. For example, in Table 6.11 for Pomeranian, BBox-20, the F1 value reported for 5-fold cross-validation is given as "9637". The same table also contains minor language errors such as "Segmetation" and "Multicalss".

4) Typographical errors and inconsistencies in model names. For example, Table 6.6 repeatedly uses the name "SE-DesneNet201".

5) Inconsistency between the title and the content of Table 6.14. The table is titled "Segmentation results of external validation using SIPaKMeD as the training set and Pomeranian as the testing set", while it contains segmentation and classification results for Pomeranian, Herlev, and SIPaKMeD.

6) Incorrect expansion of an abbreviation. The list of abbreviations includes "Global Average Pooling (GAV)", while the standard abbreviation is GAP.

7) Incorrect caption of Figure 3.33, which reads "The proposed hybrid ViT-based approach for cervical cell detection".

8) Non-standard terminology for fully connected layers. In several places in the dissertation, the author uses terms that are not commonly accepted for fully connected layers. On pages 34 and 45, the term "fully linked layers" is used, while on page 56 the term "completely connected layers" appears. The standard terms are "fully connected layers" or "dense layers".

6. Summary

I assess positively the doctoral dissertation by **Betelhem Zewdu Wubineh** entitled "*Segmentation and Classification of Pap Smear Images to Detect Cervical Cancer using Deep Learning*". The dissertation provides an extensive and valuable study of the application of modern deep learning methods to the segmentation and classification of cytological images, presents original solutions to the research problems, and confirms the Candidate's appropriate knowledge and ability to conduct scientific research independently. The comments and reservations indicated in this review do not undermine the main scientific value of the dissertation or the credibility of the conducted research program.

In view of the above, I recommend acceptance of the doctoral dissertation and admission of the Candidate to its public defense.

Taking into account the opinions presented in the previous points and the requirements defined by Article 187 of the Act of 20 July 2018 - Law on Higher Education and Science (as amended)¹ my assessment of the dissertation in terms of three basic criteria is as follows:

A. Does the dissertation contain an original solution to a scientific problem? (select one option by marking X)

☒	☐	☐	☐	☐
Definitely YES	Probably YES	It's hard to say	I don't think so	Definitely NOT

¹<http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276>

B. After reading the dissertation, do you agree that the candidate has general theoretical knowledge in the discipline of Information and Communication Technology?

☒

Definitely YES

☐

Rather YES

☐

It's hard to say

☐

I don't think so

☐

Definitely NOT

C. Does the candidate have the ability to conduct research independently?

☒

Definitely YES

☐

Rather YES

☐

It's hard to say

☐

I don't think so

☐

Definitely NOT



Signature