Short Report
UDC 618.19-006.6:530.182
DOI 10.7251/IJEEC2402062G

Optimization and Construction of a Deep Learning Model for Breast Cancer Segmentation

Luka Glišić[1], Ivana Berković[1], and Biljana Radulović[1]

[1] Department for information technologies,Tehnical Faculty “Mihajlo Pupin”, University of Novi Sad, Zrenjanin, Serbia

E-mail address: lukaglisic.srb@gmail.com, ivana.berkovic62@gmail.com, biljana.radulovic@tfzr.rs

Abstract—**The research focuses on optimizing and developing a breast cancer segmentation model utilizing multi-GPU arrays to maximize hardware resource utilization. By exploring various architectural strategies, the study aims to enhance resource allocation efficiency despite existing limitations. Detailed evaluations using mammography datasets have demonstrated significant improvements in tumor detection capabilities. This technology holds the potential to revolutionize breast cancer detection, a critical advancement given the global impact of the disease. Training data analysis confirms the scalability of these results across diverse hardware configurations, ensuring high efficiency and reliability. The study employs modern architectures, contributing valuable insights to the field of breast cancer segmentation and advancing medical imaging technologies. The development of efficient and clinically viable solutions is imperative in contemporary medical image analysis. While state-of-the-art deep learning architectures offer impressive capabilities, their substantial computational demands pose barriers to widespread clinical adoption. This research addresses the need for solutions that efficiently process large datasets while maintaining diagnostic accuracy, facilitating integration into clinical workflows and reducing operational costs.**

Keywords- machine learing; AI; image segmentation;

I. Introduction

The exponential advancement in computational capabilities has catalyzed unprecedented progress in artificial intelligence (AI) applications across medical imaging domains. Deep learning architectures have demonstrated remarkable efficacy in medical image analysis, particularly in oncological diagnostics [1]. Recent developments in computer-aided diagnosis (CAD) systems have shown significant potential in augmenting clinical decision-making processes, with particular emphasis on breast cancer detection and segmentation methodologies [2].

The optimization of deep learning models for medical image segmentation presents multifaceted challenges, encompassing computational complexity, model generalization, and clinical reliability. Contemporary research indicates that while deep neural networks achieve state-of-the-art performance in segmentation tasks, their deployment in clinical settings necessitates sophisticated optimization strategies [3]. The integration of transformer-based architectures and attention mechanisms has further expanded the computational paradigm, introducing additional optimization considerations in the context of medical image analysis [4]. This investigation addresses the fundamental challenges in optimizing deep learning models for breast cancer segmentation, with particular emphasis on the intricate balance between computational efficiency and diagnostic accuracy. The research methodology encompasses systematic analysis and implementation of advanced optimization techniques, including parallel processing architectures, memory management strategies, and precision optimization through weight quantization .

A. Motivation

The development of efficient and clinically viable solutions represents a critical imperative in contemporary medical image analysis. While state-of-the-art deep learning architectures demonstrate impressive capabilities, their implementation often requires substantial computational resources, creating significant barriers to widespread clinical adoption. This limitation becomes particularly acute in healthcare facilities with standard computational infrastructure, where the deployment of resource-intensive models proves impractical.

The increasing volume of medical imaging data in clinical practice necessitates solutions that can process large datasets efficiently while maintaining diagnostic accuracy. Current complex architectures, despite their sophisticated nature, frequently present challenges in real-time processing scenarios, limiting their integration into time-sensitive clinical workflows. Furthermore, the operational costs and energy consumption associated with running computationally intensive models pose additional constraints on healthcare providers. These practical limitations underscore the necessity for optimized approaches that achieve a more favorable balance between model performance and computational efficiency. The development of streamlined architectures that maintain high diagnostic accuracy while reducing computational overhead would significantly enhance the accessibility and utility of AI-based diagnostic tools across diverse clinical settings.

II. METHODOLOGY

In this research, a distributed strategy for training deep neural networks was applied. The initial step involved selecting the platform for the construction and training process. Subsequently, a credible source of mammographic images was identified, which were then preprocessed and prepared for the upcoming analysis. Upon completing the image import into the operational memory, a model architecture with pre-trained parameters (Imagenet) was created. However, the number of images was insufficient to achieve high precision generalization. Hence, data augmentation was applied to create batches for partial model training, which still has limitations on overall anomaly generalization in mammographic images. Nonetheless, there is potential for expanding the model to larger hardware architectures.

A. Hadrware specifications

The Kaggle platform facilitates computationally intensive tasks through a synergistic combination of scalable cloud resources and advanced hardware accelerators. These resources encompass Central Processing Units (CPUs), Graphics Processing Units (GPUs), and Tensor Processing Units (TPUs), each playing a pivotal role in various stages of machine learning and data processing workflows.

The Intel Xeon 2.20 GHz CPU, featuring four virtual cores and 32 GB of Random Access Memory (RAM), is optimized for general-purpose computations and serves as the foundation for preprocessing, data Smanagement, and sequential computations. CPUs are particularly well-suited for tasks requiring flexibility, such as handling heterogeneous data pipelines, executing conditional logic, or performing operations that are not easily parallelizable. The CPU's multithreaded architecture enables concurrent task handling, ensuring efficient resource allocation and reduced latency during operations such as data loading, transformation, and feature extraction.

TABLE I. ACCESSIBLE HARDWARE SPECIFICATIONS

Hardware Components	**Number of Cores**	**Memory**
NVIDIA Tesla P100 GPU	3584 CUDA cores	16 GB VRAM
NVIDIA T4 x2 GPU	2560 CUDA cores	16 GB VRAM
Google TPU v3-8	8 TPU v3 cores	128GB VRAM
Intel Xeon 2.20 GHz CPU	4 vCPU cores	32GB RAM

B. Source and Processing of Images

For model training, mammographic images from the CBIS-DDSM (Curated Breast Imaging Subset DDSM Dataset) [5] dataset were utilized, following a preprocessing procedure involving extraction, normalization, and pairing for training. In total, 10,642 images were preprocessed and loaded into memory. Employing parallelism, in accordance with Amdahl's Law [6]:

$$S(N) = \frac{1}{(1-p) + \frac{p}{N}} \tag{1}$$

where S(N) represents the theoretical speedup factor, p is the proportion of the program that can be parallelized, and N is the number of processing cores. Using this equation with our quad-core processor (N=4) and a parallelizable portion of 85% (p=0.85), an acceleration of 392% is attained, in contrast to the image processing duration without parallelism utilization. Subsequent to partitioning the dataset, images undergo resizing to dimensions of 224x224 to ensure efficient and uniform model training.

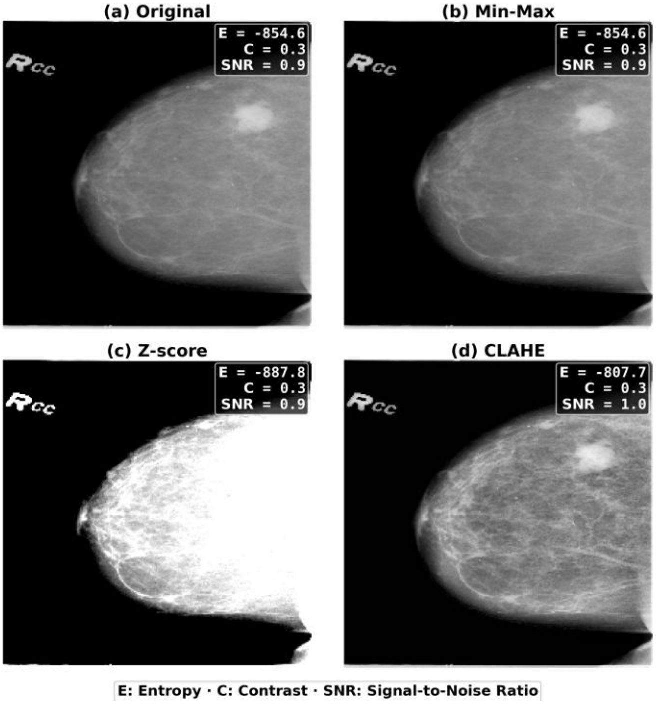


Figure 1. Mammogram Normalization - Original vs. Normalized Images

Fig. 1 demonstrates various normalization techniques applied to a breast image. The original image (a) serves as the baseline. Min-Max normalization (b) adjusts pixel values to a specific range, preserving appearance but not significantly enhancing contrast. Z-score normalization (c) standardizes pixel values, increasing the signal-to-noise ratio but potentially causing overexposure and detail loss in bright areas. CLAHE (d) enhances local contrast through histogram equalization in small regions, preserving details and making subtle differences more visible. This method is preferred in medical imaging for its ability to improve feature visibility without excessive noise amplification. Through this normalization process, the model has become more sensitive to details in the images, facilitating the identification of tumorous regions and other pathological changes. This method has provided stability, generalization, and better differentiation of tissue characteristics in medical breast images, thereby contributing to overall improvement in segmentation processes and result quality.

C. Model Architecture

In this study, three deep learning architectures were selected: VGG19, InceptionV3, and SEResNet50. VGG19, known for its simplicity and effectiveness in hierarchical feature extraction, was selected for its well-established performance in image classification tasks [7]. InceptionV3, with its innovative Inception modules, enables the model to efficiently learn multiscale features, making it suitable for complex image analysis tasks [8]. SEResNet50, which incorporates Squeeze-and-Excitation (SE) blocks into the ResNet architecture, was chosen for its enhanced representational power by adaptively recalibrating feature responses [9].

The Dice coefficient, as represented by (2), is a robust metric for evaluating the similarity between predicted and actual image segments [10]. Particularly suited for breast cancer segmentation, it quantifies the overlap between predicted and actual tumor regions by accounting for true positive overlap while penalizing false positives and false negatives. This balanced evaluation is critical in medical imaging, where precise delineation of tumor boundaries significantly impacts diagnosis and treatment planning. By ensuring accurate capture of intricate tumor morphology, the Dice coefficient facilitates reliable model performance assessment and supports effective clinical decision-making.

$$Dice(y_{true}, y_{pred}) = 2 * \frac{\sum(y_{true} * y_{pred}) + pom}{\sum y_{true} + \sum y_{pred} + pom} \quad (2)$$

Due to the model's sigmoid activation function at the output, the intersection and union are mapped to multiplication and addition, while the purpose of the auxiliary variable set to 0.0001 is to avoid division by zero.

The Intersection over Union (IoU) coefficient [11], as described in (3), is a key measure in evaluating model performance, as it enables the quantification of tumor segmentation accuracy relative to actual contours in medical images.

$$IOU(y_{true}, y_{pred}) = \frac{\sum(y_{true} \cap y_{pred}) + pom}{\sum(y_{true} \cup y_{pred}) + pom} \quad (3)$$

The loss function is formulated as a fusion of the Dice coefficient and the Intersection over Union (IoU) coefficient. This composite function offers a holistic evaluation of the model's performance in image segmentation, accounting for both the precision of segmentation localization and the degree of overlap with ground truth image segments. The weighting assigned to the Dice coefficient is 0.4, whereas for the IoU coefficient, it is 0.6. By subtracting the resultant value from 1, the ultimate expression of the loss function is derived.

D. Training Methodology

In the initial phase of the research, a dynamic strategy was employed for model training, utilizing available hardware resources based on model size and computational needs. Initially, a single Nvidia Tesla P100 GPU was used for training smaller models. The training involved a batch size of 8 augmented image pairs, with data augmentation through rotation applied to increase model robustness. As the models became larger, two NVIDIA T4 graphics cards were employed, utilizing a distributed training strategy. This enabled the scaling of the batch size to 16 pairs, processing 32 augmented pairs per epoch. This flexible approach ensured efficient resource utilization and maintained optimal training times, allowing the models to be trained effectively while leveraging the hardware's capabilities. Data augmentation through rotation remained a key strategy, enhancing the dataset and improving model generalization.

III. Results

Neural Architecture Search (NAS) was used to explore various models and identify optimal architectures for performance and efficiency. Among the models tested, SEResNet50, InceptionV3, and VGG19 were selected based on their distinct characteristics and performance. SEResNet50, with its attention mechanisms and residual connections, outperformed the others, showing superior Dice and IoU scores while maintaining a balance in computational complexity. InceptionV3 demonstrated strong performance and was chosen for further training to evaluate its scalability. VGG19, despite its lower performance, was included to analyze the impact of architectural simplicity and parameter count on segmentation. This selection highlights the importance of varying model complexity in optimizing segmentation tasks.

TABLE II. Neural Architecture Search (NAS) performance

Backbone	Params	Dice	IoU	Training Time
SEResNet50	35.05M	0.6642	0.4640	3.4
inceptionv3	29.90M	0.6544	0.4568	3.1
densenet121	12.06M	0.6409	0.4256	3.6
mobilenet	8.31M	0.6498	0.4106	1.6
vgg19	29.06M	0.4182	0.2515	2.9

InceptionV3 outperforms SE-ResNet50 despite its lower initial NAS ranking due to its architectural strengths. Its inception modules enable efficient multi-scale feature extraction, improving adaptability and generalization, particularly in tasks like medical image segmentation where capturing fine details is crucial. This flexibility, often overlooked in NAS evaluations prioritizing efficiency, becomes evident with further training and optimization. Additionally, task-specific hyperparameter tuning likely enhanced its performance, emphasizing the interplay of architecture and training in model effectiveness.

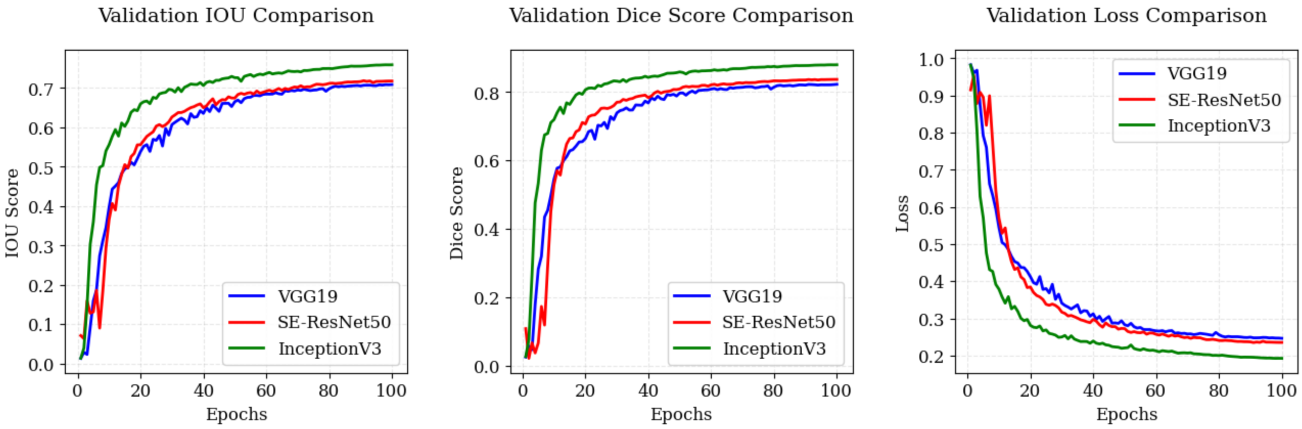


Figure 2. Validation Metrics Comparison

The performance metrics, as shown in Fig. 2, highlight the distinct capabilities of VGG19, InceptionV3, and SeResNet50 in breast cancer image segmentation. InceptionV3 stands out for its exceptional accuracy, attributed to its advanced architecture that effectively captures complex patterns, as evidenced by its superior Dice and IoU scores. SeResNet50 also demonstrates strong performance, benefiting from its design that enhances feature focus and precision, resulting in competitive validation metrics. In contrast, VGG19, despite a comparable parameter count, shows limitations in detail capture, as reflected in its lower performance metrics. These insights emphasize the importance of sophisticated network designs in achieving high precision in medical image analysis.

TABLE III. AVERAGE METRICS OF THE MODEL ON THE VALIDATION SET

	Dice coefficient	IOU	Parameter Count
VGG19	0.822	0.708	29.06M
inceptionv3	**0.879**	**0.759**	**29.90M**
Seresnet50	0.836	0.717	29.90M

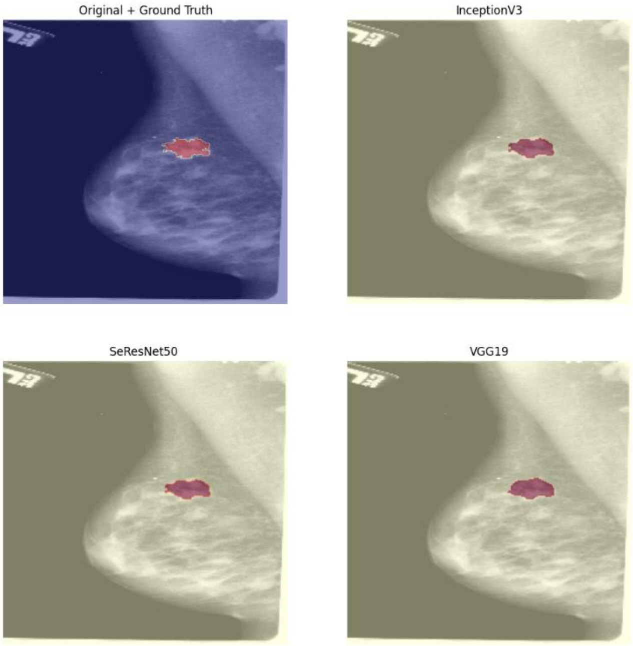


Figure 3. Image Segmentation Comparison

The segmentation results of breast cancer images using InceptionV3, SeResNet50, and VGG19 are compared against ground truth annotations, which highlight the precise region of interest. SeResNet50 achieves the most accurate boundary delineation, closely followed by InceptionV3, while VGG19 captures the general region but with less precision. This comparison highlights the effectiveness of NAS in selecting architectures that optimize segmentation accuracy, with

IV. DISCUSSION

The results of this study highlight the potential of advanced neural network architectures like InceptionV3 and SeResNet50 in addressing real-world challenges in breast cancer diagnosis. These models demonstrated superior segmentation capabilities, with InceptionV3 achieving the highest Dice coefficient (0.879) and IOU (0.759), followed closely by SeResNet50 (Dice coefficient: 0.836, IOU: 0.717). These architectures excel at identifying regions of interest (ROIs) with high precision, which could significantly enhance clinical workflows in hospital settings.

One practical application of these findings is the development of automated ROI selection tools. By leveraging segmentation models such as InceptionV3 or SeResNet50, clinicians could rapidly identify critical regions in medical scans, reducing the manual effort required for diagnosis. This could be particularly valuable in breast cancer diagnosis, where identifying the tumor boundary and extent is essential for treatment planning. Automating these tasks not only saves time but also minimizes the risk of human error, ensuring consistent and accurate results across cases.

Moreover, the adoption of such models could alleviate the reliance on high-performance computational infrastructure. By pre-processing scans to segment ROIs before storage, hospitals could significantly reduce the data volume required for long-term archival. For instance, rather than storing full-resolution images, only the segmented ROIs—augmented with relevant metadata—could be saved. This approach could optimize storage resources while ensuring that critical diagnostic information remains accessible.

In resource-constrained settings, deploying slightly less complex models like SeResNet50 or even optimized variants of InceptionV3 might allow hospitals to perform segmentation locally, avoiding the need for cloud-based solutions or powerful centralized hardware. Additionally, integrating these models into edge devices, such as portable imaging systems or workstation-level machines, could make high-accuracy diagnostics feasible in remote or under-equipped facilities.

These advances also have implications for patient-centered care. Faster segmentation and ROI identification could streamline workflows for radiologists, enabling quicker diagnostic reports and reducing patient wait times. Furthermore, accurate segmentation could guide treatment decisions, such as biopsy locations or radiation therapy planning, thereby improving outcomes.

To fully realize these benefits, further efforts are needed to optimize these models for computational efficiency. Techniques such as model pruning, quantization, and transfer learning could enable their deployment on standard hospital systems without compromising accuracy. Additionally, integrating these models into existing radiology software and SeResNet50 and InceptionV3 outperforming VGG19. ensuring compliance with data privacy regulations would be essential for practical adoption.

V. CONCLUSION

This research demonstrates the effectiveness of advanced neural network architectures, particularly SeResNet50 and InceptionV3, in improving breast cancer segmentation in medical imaging. These models exhibited strong performance in accurately delineating tumor boundaries, which is essential for precise diagnosis and treatment planning. By enhancing segmentation accuracy, these models can serve as valuable tools for doctors, providing additional support in tumor detection and assessment.

In practical applications, these models could be integrated into clinical workflows, aiding radiologists in making more informed decisions and streamlining the diagnostic process. Future work will focus on optimizing these models for real-time deployment and exploring their potential across other areas of medical imaging. Additionally, cutting-edge architectures such as Vision Transformers (ViTs) and ConvNeXT, which have demonstrated superior performance in capturing both global and local features, offer exciting possibilities for further improving segmentation accuracy and efficiency in medical imaging [12][13].

VI. REFERENCES

[1] S. Nasser, M., & Yusof, U. K. (2022). Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction. Diagnostics, 13(1). https://doi.org/10.3390/diagnostics13010161

[2] Adam, R., Dell'Aquila, K., Hodges, L. et al. Deep learning applications to breast cancer detection by magnetic resonance imaging: a literature review. Breast Cancer Res 25, 87 (2023). https://doi.org/10.1186/s13058-023-01687-4

[3] Litjens, G., et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88. https://doi.org/10.1016/j.media.2017.07.005

[4] Müjdat Tiryaki V. *Mass segmentation and classification from film mammograms using cascaded deep transfer learning.* Biomedical Signal Processing and Control. 2023;84:104819. doi: 10.1016/j.bspc.2023.104819.

[5] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy & Daniel L. Rubin. (2017) *A curated mammography data set for use in computer-aided detection and diagnosis research.* Scientific Data volume 4, Article number: 170177 DOI: https://doi.org/10.1038/sdata.2017.177

[6] John L. Hennessy, David A. Patterson. (2006) *Computer Architecture: A Quantitative Approach, 4th Edition* Morgan Kaufmann

[7] Simonyan, K., & Zisserman, A. (2014). Very deep

convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[8] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).

[9] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

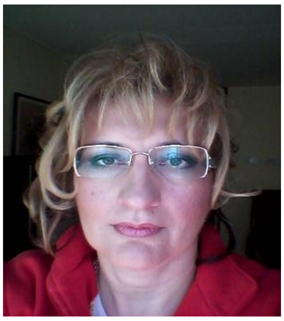
[10] Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. Ecology, 26(3), 297–302. doi:10.2307/1932409

[11] Jaccard, P. (1901). *Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles*, 37,w 547–579.

[12] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. arXiv:2201.03545. https://doi.org/10.48550/arXiv.2201.03545.

Luka Glišić, Born in 2004 in Novi Sad, currently a third-year student majoring in Software Engineering at the Tehnical Faculty "Mihajlo Pupin" in Zrenjanin, Serbia. Recognized for his outstanding academic performance, Luka's research interests encompass artificial intelligence, software engineering, and embedded systems, with a focus on optimizing system performance and enhancing user experiences. His dedication to innovation and academic excellence underscores his commitment to advancing technology and contributing to computer science.

Ivana Berković is a Full Professor at the Technical Faculty "Mihajlo Pupin" in Zrenjanin, Serbia. She obtained her Bachelor's degree from the Faculty of Sciences in Novi Sad and completed her Master's and Ph.D. degrees at the Technical Faculty "Mihajlo Pupin" in Zrenjanin, specializing in Logic Programming and Automated Reasoning. Since 1987, Prof. Berković has been teaching at the Technical Faculty "Mihajlo Pupin." In 2008, she was appointed as a Full Professor at the University of Novi Sad. Her research interests include Artificial Intelligence, Automated Reasoning, Logic Programming Languages, and Computer Graphics. She has authored numerous scientific papers, textbooks, and software products.



Biljana Radulovic is a Full Professor at the University of Novi Sad, Serbia. She earned her Bachelor's and Master's degrees in Informatics in Business from the University of Novi Sad in 1988 and obtained her Ph.D. from the same university in 1998, specializing in Databases. Prof. Radulovic teaches undergraduate courses in Data Base and Information Systems, and at the Master's level, she instructs on Business Intelligence and Distributed Information Systems. She has authored or co-authored over 180 papers in international and national conferences, journals, and books. Her research interests include Data Base, Business Intelligence, Information Systems, Software Engineering, and Artificial Intelligence..