Electronic Journal of General Medicine
2026, 23(3), em732
e-ISSN: 2516-3507
https://www.ejgm.co.uk/

MODESTUM

Original Article

OPEN ACCESS

ML in breast cancer IHC: Pilot evaluation on non-ideal slides in Kazakhstan

Arshat Urazbayev [1], Bakytzhan Amangeldinovna Issakhanova [2,3], Zhanas Baimagambet [4,5], Zamart Ramazanova [1,6], Yeldar Baiken [1,7], Askhat Myngbay [1,8]*

[1] PI National Laboratory Astana, Nazarbayev University, Astana, KAZAKHSTAN
[2] Department of Anatomic Pathology, RSE "Medical Centre Hospital of the President's Affairs Administration of the Republic of Kazakhstan", Astana, KAZAKHSTAN
[3] QazGene LLP, Astana, KAZAKHSTAN
[4] School of Medicine, Nazarbayev University, Astana, KAZAKHSTAN
[5] University of Cape Town, Faculty of Health Sciences, SOUTH AFRICA
[6] Department of Electrical and Computer Engineering, School of Engineering and Digital Sciences, Nazarbayev University, Astana, KAZAKHSTAN
[7] Center for BioEnergy Research LLP, Astana, KAZAKHSTAN
[8] Department of Biology, K. Zhubanov Aktobe Regional University, Aktobe, KAZAKHSTAN
***Corresponding Author:** askhat.myngbay@zhubanov.edu.kz

Citation: Urazbayev A, Issakhanova BA, Baimagambet Z, Ramazanova Z, Baiken Y, Myngbay A. ML in breast cancer IHC: Pilot evaluation on non-ideal slides in Kazakhstan. Electron J Gen Med. 2026;23(3):em732. https://doi.org/10.29333/ejgm/18520

ARTICLE INFO

Received: 06 Jul. 2025

Accepted: 06 Mar. 2026

ABSTRACT

Objectives: Automation of quantitative analysis of breast cancer (BC) immunohistochemistry (IHC) specimens is important to optimize pathologists' workflow and improve diagnostic reproducibility. This is especially important in low- and middle-income countries where there is a shortage of highly trained pathologists. However, existing approaches face challenges in implementing fully automated quantitative IHC due to the difficulty of both delineating tumor areas, including discrete areas, especially in IHC slides with poor quality. Moreover, accurate identification of invasive carcinoma areas and accurate quantification of positive and negative cells in the specimen are critical for quantitative analysis.

Methods and results: This study presents a method to automatically identify types of carcinoma areas in whole slide IHC images of BC, focusing on quantifying IHC images on realms of Kazakhstan. The used model is a combination of morphological characteristics and boundary features, which provides high accuracy of segmentation of tumor zones of images of mild and low quality. We used several methods includes convolutional neural network based on the Keras framework, k-nearest neighbors machine learning methods, and self-developed image analysis methods. The developed model showed high accuracy, where the results corresponded to the diagnoses of pathologists. As expected, the method proved to be ineffective when applied to severely degraded slides, such as those with insufficient staining or inadequate washing. Slides of inferior quality were excluded from analysis, which negatively affected the statistical robustness. On slides of moderate quality, the reliability of nucleus segmentation dropped significantly.

Conclusions: The combination of models we used showed high accuracy in differentiating BC cells between the basal-like subtype of BC and its invasiveness and recurrence in Kazakhstan. However, IHC specimens with low DPI or low-quality IHC need further optimizations and improvements in algorithm design. The main issue can be considered methodological differences between the approaches of AI and humans: AI operates in a large number of cases (more than 10,000), yet its accuracy is relatively low. In contrast, humans work with a much smaller number of cases but achieve a level of precision that AI cannot currently match. This discrepancy necessitates a revision of the methodology of IHC analysis for AI, including the development of new requirements, methods, and thresholds from scratch. This approach provides analysis of the entire area of the slide, increases the speed of interpretation of IHC results, and reduces human errors in diagnosis, especially in low- and middle-income countries.

Keywords: machine learning, breast cancer, diagnostics, Kazakhstan, immunohistochemistry

INTRODUCTION

Breast cancer (BC) is still one of the fastest-growing global health challenges, affecting millions of individuals annually and claiming hundreds of thousands of lives. In 2022 alone, about 2.3 million new cases of BC were registered worldwide, with approximately 670,000 deaths recorded, underscoring its significant impact on global mortality. Although the 99% of the cases are related to women, the other 0.5-1% consisted of men [1]. Nowadays, there are certain delays in the diagnosis of BC, especially among young women [2]. It is more prevalent in low- and middle-income countries. This might be due to the lack of awareness of local people, limited resources in early diagnosis, or in appropriate treatment. Compared to high-income countries, the number of incidents in low-income countries is low, with a higher rate of mortality [3]. This could be the cause of being diagnosed late or misdiagnosed at early stages of the

disease. Therefore, it is important to have a tool for establishing a diagnosis in a timely manner with less pathologist involvement.

One of the most reliable methods of diagnosis of BC is immunohistochemistry (IHC) [4]. IHC is a cornerstone technique in histopathology, providing invaluable insights into the presence, abundance, and precise localization of specific proteins within tissue samples [4]. This analysis involves staining formalin-fixed and paraffin-embedded tissue slices with hematoxylin and specific markers. After staining, tissue cells are incubated with primary antibodies to distinguish between normal cells and cancer cells. Eventually, slides are examined manually by pathologists for intensity and quantity of positive cells. And it is worth mentioning that the whole slide consists of millions of cells; manual evaluation by a trained pathologist could sometimes be biased or incorrect [5]. Such factors can have an impact on further treatment effectiveness or designing an appropriate treatment strategy. The spatial distribution of specific markers, often in the context of the complex tumor microenvironment (TME), holds significant biological and clinical relevance [6]. Understanding these spatial relationships is crucial for unraveling disease mechanisms, predicting patient prognosis, and guiding treatment decisions.

Traditional manual analysis of IHC images has significant limitations, as it is inherently time-consuming, demanding substantial expertise from trained pathologists, and is prone to inter-observer variability, leading to inconsistencies in diagnosis and prognosis. The subjectivity of manual scoring introduces biases that can affect the reliability and reproducibility of results across different researchers and laboratories. Machine learning (ML), with its capacity for automated image analysis and pattern recognition, offers a powerful solution to overcome these limitations. ML algorithms can automate the analysis of spatial patterns in IHC data, resulting in improved accuracy, efficiency, increased throughput, and the potential for the discovery of novel biomarkers that might be missed by traditional methods [7].

Nowadays, the impact of artificial intelligence (AI) in medicine is increasing. There are several studies showing algorithms used for the interpretation of BC scoring using HER2 marker and markers such as ER [8]. However, those algorithms are designed for perfect, clear images, which can be an obstacle in low- and middle-income countries.

In this study, we examined the role of AI in improving immunohistochemical interpretation in BC diagnosis in Kazakhstan according to its capabilities and IHC image quality. In this study, alongside other key markers of BC, we used AI/ML algorithms to interpret IHC-stained sections for CK5/6, p53, and p63. These markers are used to distinguish between basal like subtype of BC and its invasiveness and recurrence.

This paper demonstrates the application of ML techniques for spatial analysis of IHC images according to Kazakhstani capabilities, their applications, challenges, and the promising future directions in the field.

METHODS AND MATERIALS

Clinical Data

We retrospectively selected 307 immunohistochemical slides of patients with an established diagnosis of BC who underwent IHC examination from January 2023 to December 2024 at RSE “Medical Centre Hospital of the President’s Affairs Administration of the Republic of Kazakhstan,” Astana, Kazakhstan. The slides were scanned using an Aperio CS2 scanner. After excluding missing and very poor-quality slides, a combination of AI/ML algorithms was used to identify markers like ER, PR, HER2, Ki-67, CK5/6, p53, and p63. This study was approved by the institutional review board of RSE “Medical Centre Hospital of the President’s Affairs Administration of the Republic of Kazakhstan.” Due to the retrospective nature of this study, patient-informed consent was not required.

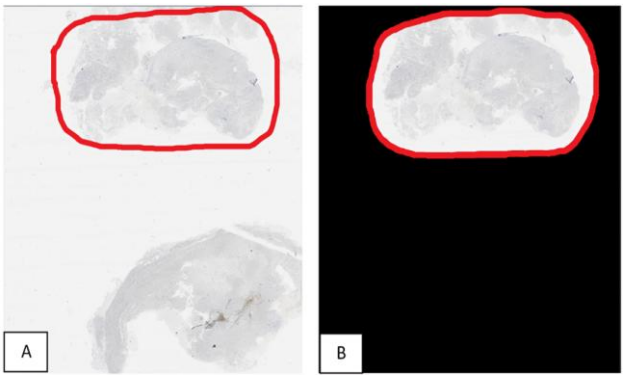

Figure 1. (A) General view of the slide, primary, and reference object & (B) after removal of the reference object (the main object, outlined in red by a pathologist, is shown, with the reference object located below it & on the right, the same image is displayed after the removal of the reference object) (Source: Authors’ own elaboration, using Python Plotly data visualization libraries)

Study Design

All obtained images were in scanned virtual slide (SVS) format. SVS is a designated, developed format for image data, typically featuring large image dimensions of up to 100,000 × 100,000 pixels. Our target object occupied only one portion of the whole slide IHC, as the other object was a reference (positive control) (part A in **Figure 1**). Pathologists delineated the regions of interest (ROI) using a red contour. At this stage, slides that were clearly unsuitable for further analysis were also excluded, including completely empty slides and those in which reliable annotation of the ROI was not possible. As a result, only **286 slides** were retained for subsequent analysis.

To identify areas of the image suitable for analysis, we divided the entire image into 500-pixel-by-500-pixel tiles. This dimension was chosen for two reasons: first, due to the limited amount of RAM (8 GB); at this scale, image quality (e.g., focus, staining quality, etc.) appeared relatively consistent. At this resolution, it was feasible to judge whether a tile was suitable for further analysis.

As part of preprocessing, we computed the Fourier spectrum of each tile. If high frequencies were absent in the spectrum, it indicated either that the image was out of focus or that the tile contained no tissue. This type of filtering reduced the number of tiles from 8 million to 414 thousand tiles. Still, even this number was too large for manual processing.

To address this, we needed a prepared data set to train an AI model for sorting our tiles. To minimize the time commitment required from medical professionals, we limited the classification to two categories: suitable and not suitable.

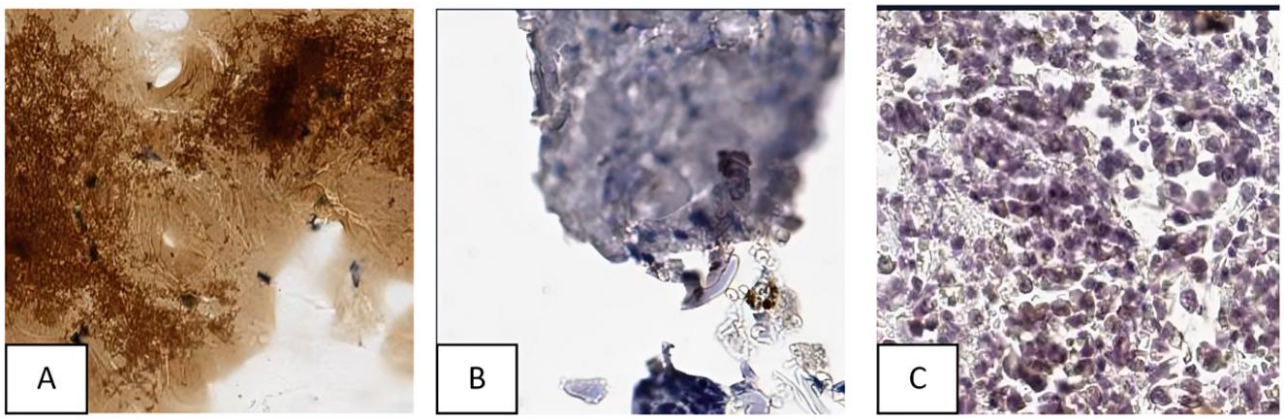

Figure 2. Examples of Keras output: (A) probability 0.0001 (unsuitable), (B) 0.06 (definitely unsuitable), & (C) 0.25 (unfit) (Source: Authors' own elaboration, using Python Plotly data visualization libraries)

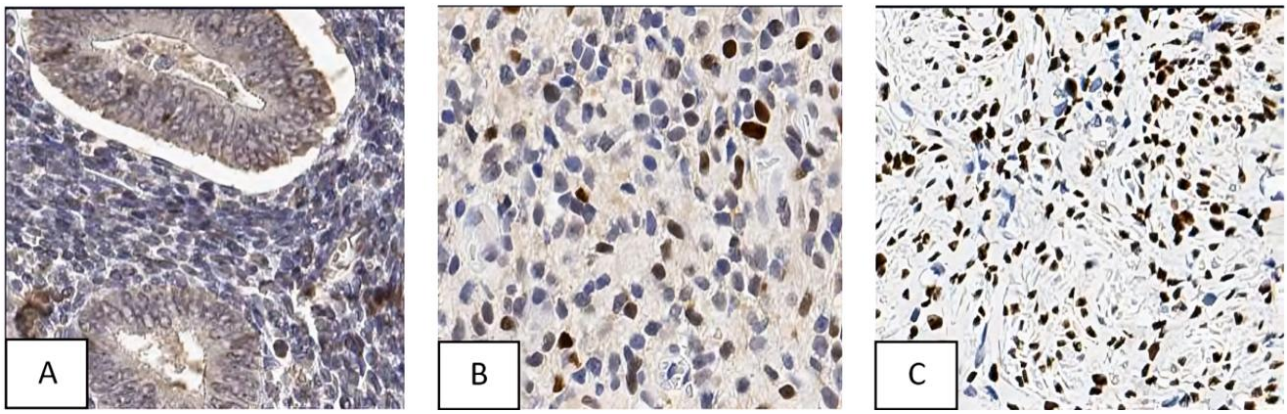

Figure 3. More CNN output examples: (A) probability 0.35 (most likely unsuitable), (B) 0.48 (uncertain), & (C) 0.79 (definitely suitable) (Source: Authors' own elaboration, using Python Plotly data visualization libraries)

Creating such an annotated dataset was relatively straightforward. In total, 12,000 tile images were labeled.

We used a convolutional neural network (CNN) based on the Keras framework [9]. For each image, the network provided the probability that the tile was suitable for further processing (**Figure 2** and **Figure 3**).

By default, a probability greater than 0.5 is considered sufficient to deem a tile suitable for further processing. According to our observations, tiles with a probability above this threshold were indeed almost always appropriate; however, they only represented approximately 12% of the dataset. It is 12% out of 1.8 million, which amounts to several hundred thousand tiles, providing a solid base; however, we aimed for better statistical coverage. Lowering the threshold to 0.4 increased the proportion of suitable tiles to 30% (more than double), but at the cost of reduced accuracy. From this point onward, a trade-off arises between statistical representativity and precision. While it is possible to achieve an accuracy of 99% with this method, the number of processed cases would be relatively small.

RESULTS

Processing of Individual Tiles

The identification of cell nuclei is challenging due to the large variety and abundance of different objects. Therefore, the task was to filter objects based on specific criteria:

1. **Color filtering:** Only brown (positive) and blue (negative) objects are considered for further analysis.
2. **Amplitude (intensity) filtering:** Cell nuclei typically exhibit well-defined boundaries, distinguishing them from other structures.
3. **Size and shape filtering:** Cell nuclei possess characteristic dimensions and morphological features that allow their discrimination from non-nuclear objects.

Color Model

Color is defined by three parameters—red, green, and blue (RGB)—each ranging from 0 to 255. One of the key issues with our images is the attenuation of the BLUE channel by approximately 40 units. As a result, the background color, which should ideally appear white, instead exhibits a yellowish tint. This distortion limits the applicability of standard color thresholding methods.

To address this, we developed a more generalizable approach. Several representative samples of blue and brown-stained cells were manually selected. In total, several image patches were extracted and categorized into three classes: blue (negative) nuclei, brown (positive) nuclei, and all other objects (**Figure 4**). Each object's RGB color vector (consisting of three numerical values) was used as input data for ML. Given the low dimensionality of the feature vector, for classification purposes k-nearest neighbors (k-NN) algorithm was employed [10].

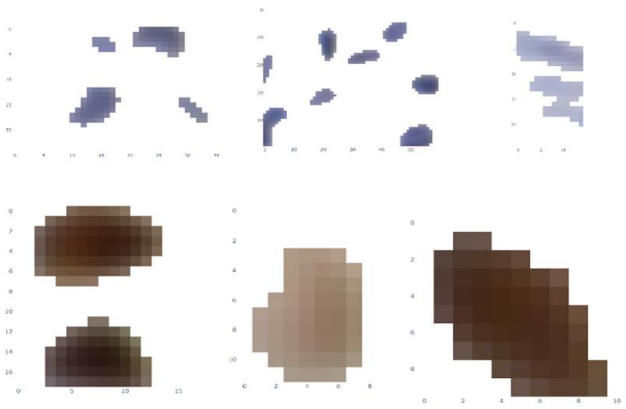

Figure 4. Examples of extracted nuclei that make up the data set based on color (Source: Authors' own elaboration, using Python Plotly data visualization libraries)

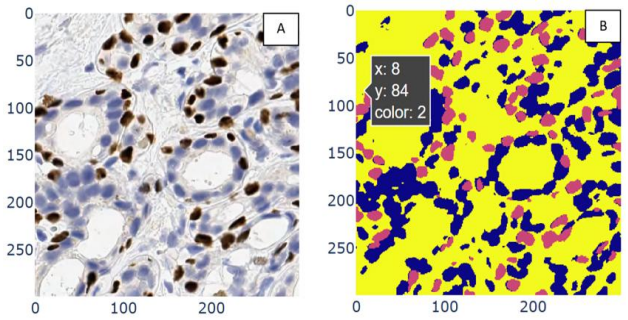

Figure 5. Result of applying the k-NN model to a tile's pixels: (A) original image & (B) classification into three classes (blue corresponds to class 1 [negative], red to class 2 [positive], and yellow to class 3 [other objects]) (Source: Authors' own elaboration, using Python Plotly data visualization libraries)

The k-NN algorithm is one of the ML techniques used for classification and regression purposes. k-NN is not designed to build explicit models during training. Instead, all the learning consists of is remembering all the data in the training set. In our case, these are three RGB channels. The algorithm performs the following steps: It determines the value of the parameter 'known,' which indicates the number of nearest neighbors. Then, it calculates the distances (e.g., Euclidean distance) between the new object and all objects from the training set. The k-NN algorithm is often used for its simplicity and efficiency on small datasets. However, its application can be complex on extensive data due to the computational complexity of determining the nearest neighbors.

Each pixel in the image was assigned a value from 0 to 2 by the model, corresponding to one of the three classes: "positive," "negative," or "non-cell nucleus." The classification results are shown in **Figure 5**.

Amplitude (Intensity) Filtering

The first stage involves localizing intensity maxima in the 2D image. Around each local maximum, we performed segmentation to form an object consisting of pixels with varying intensity levels. The object is constructed using the flood fill algorithm, seeded at the local maximum (**Figure 6**).

At the initial stage, the algorithm begins from a single pixel, which must be a local intensity maximum. From this starting point, the algorithm evaluates all four directly adjacent pixels—those located to the left, right, above, and below (i.e., the cross-shaped neighborhood, as shown in **Figure 3**).

Each of the four neighboring pixels is evaluated according to the following criteria:

1. Its intensity must be lower than that of the starting (seed) pixel.
2. Its intensity must be greater than zero.
3. It must not already belong to another segmented object.

If a pixel satisfies all three conditions, it is added to the current object. Otherwise, it is excluded. The process is then repeated recursively for every new pixel added to the object, applying the same rules to its four direct neighbors. This iterative procedure continues until no more eligible pixels are found—effectively segmenting the entire object.

When visualized in 3D (with pixel intensities as height), the algorithm performs a downhill traversal, where movement is only permitted to pixels of lower intensity, never upward (left part in **Figure 5**).

An example of the algorithm's operation is provided in **Figure 3**. Suppose the seed pixel has an intensity of 70. Among its neighbors, three are eligible for consideration. A pixel with an intensity of 90 is excluded because it has already been assigned to another object. Pixels with intensities 60 and 10 are added to the object, while the 90-intensity pixel is excluded due to the condition 70 < 90. The process is then repeated for the neighbors of the 60-intensity pixel, and so on.

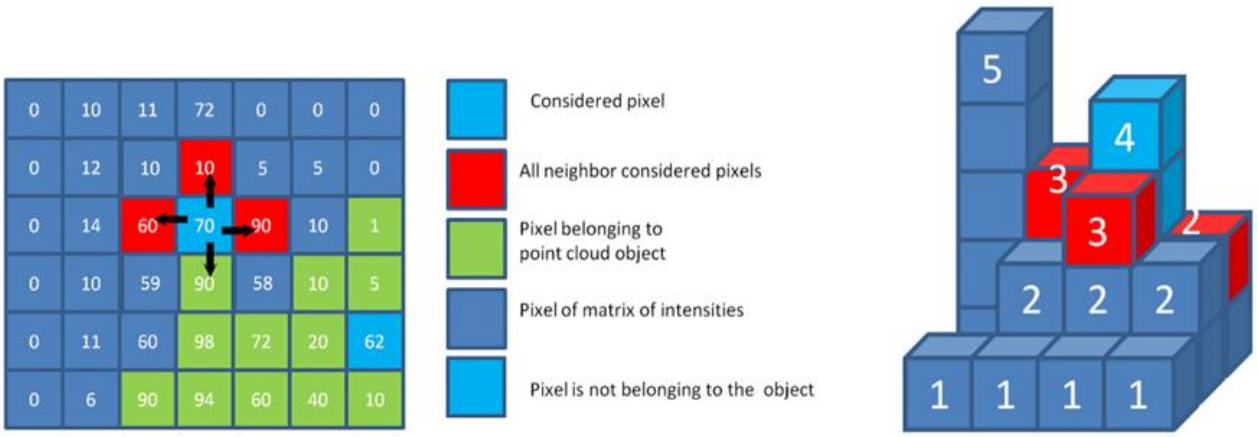

Figure 6. Schematic of the flood fill algorithm (left) & intensity matrix, and algorithmic flow (right) (Source: Authors' own elaboration)

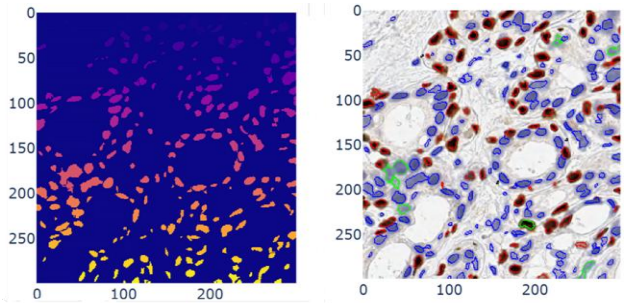

Figure 7. Result of the matrix segmentation (Source: Authors' own elaboration, using Python Plotly data visualization libraries)

A thresholding mechanism can also be introduced into the algorithm if the central maximum has an intensity *I*; only pixels with intensities greater than k × I/100% are allowed to join the object. This threshold is relative to the local maximum and thus serves as a form of adaptive thresholding. It will enable the inclusion of low-intensity maxima while restricting the spread of objects to only sufficiently similar regions.

This algorithm was applied to an intensity matrix from which all irrelevant color components had been removed. The result is shown in **Figure 7**: each object is assigned a unique ID number, and the output is a matrix where each pixel belonging to an object is labeled with its corresponding object number.

Size and Shape Filtering

To overcome overlapped nuclei, it was necessary to develop an algorithm capable of detecting such composite objects.

The most straightforward approach was to employ existing software solutions. We used StarDist [11], a widely adopted tool for nuclear segmentation. To make our data suitable, we applied a custom preprocessing pipeline to the input images. This approach yielded promising results; however, we found that StarDist struggled to handle large datasets in our specific case. Specifically, it consistently failed after processing approximately 50-60 tiles, producing a kernel-related error (**Figure 8**).

As an alternative to existing tools, we developed a custom method for detecting whether a given object consists of overlapping cell nuclei. The key idea was to compute a curvature-related parameter characterizing the tortuosity of the object's contour. To do this, we extracted the contour of the object and computed the angle of the vector connecting each pair of neighboring points along the contour.

By analogy with neural network assessment frameworks, we prepared annotated data for IHC images. To reduce pathologists' time involvement, annotation was organized in two stages: initial pre-labeling was performed by our program, followed by expert refinement into fully annotated datasets. In total, 1,070 cells were obtained.

Using these data, the algorithm was executed multiple times with adjustments to its parameters. Specifically, we modified the color model and varied certain filter parameters, such as size. In particular, the algorithm has demonstrated improved ability to distinguish debris, which had previously been misclassified as cells.

Furthermore, following the same logic as in neural network evaluation, an independent validation dataset was prepared.

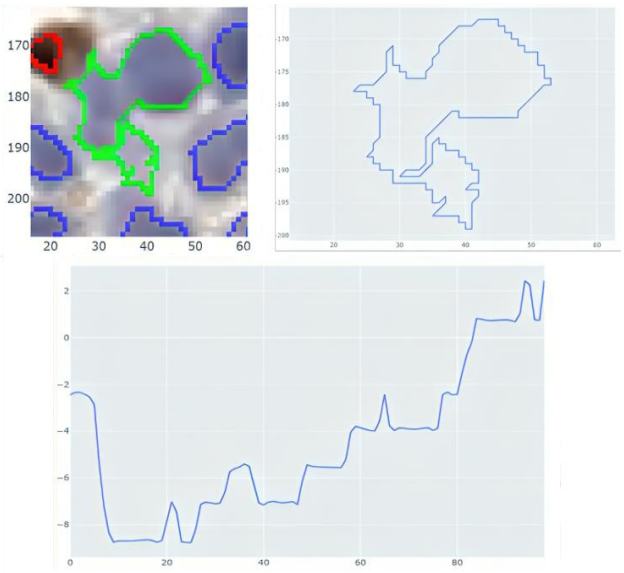

Figure 8. Example of overlapping cell nuclei, corresponding object contour, and (below) the graph of the angle trajectory along this contour (Source: Authors' own elaboration, using Python Plotly data visualization libraries)

Using the same annotation approach, 1,533 cells were labeled. To assess accuracy, we employed the dice similarity coefficient (dice metric). Since two object classes were considered—positive and negative cells—the dice metric was calculated separately for each class.

As in the previous stage, to facilitate the work of medical experts, we used tiles that had been preliminarily annotated by the computer. These pre-annotated data were then reviewed by the experts. To apply the dice metric, the errors were further classified into six categories: three classes for negative cells and three for positive cells.

For negative cells:

- **Class 1:** A negative cell that was missed by the computer (the blue cell was not outlined), corresponding to a *false negative* in dice terminology.
- **Class 2:** A false detection, where non-cellular structures (e.g., debris) were incorrectly outlined in blue, corresponding to a *false positive*.
- **Class 3:** A false detection due to color confusion, where a positive (brown) cell was incorrectly classified as negative and outlined in blue, also corresponding to a *false positive*.

In this case, the ground truth was derived from the computer output by adding class 1 errors and subtracting class 2 and class 3 errors.

An analogous classification was applied to positive cells:

- **Class 1:** A positive cell that was missed by the computer (a brown cell not outlined), corresponding to a *false negative*.
- **Class 2:** A false detection, where debris was incorrectly identified as a positive cell and outlined in red, corresponding to a *false positive*.
- **Class 3:** A false detection due to color confusion, where a negative (blue) cell was incorrectly classified as positive and outlined in red, also corresponding to a *false positive*.

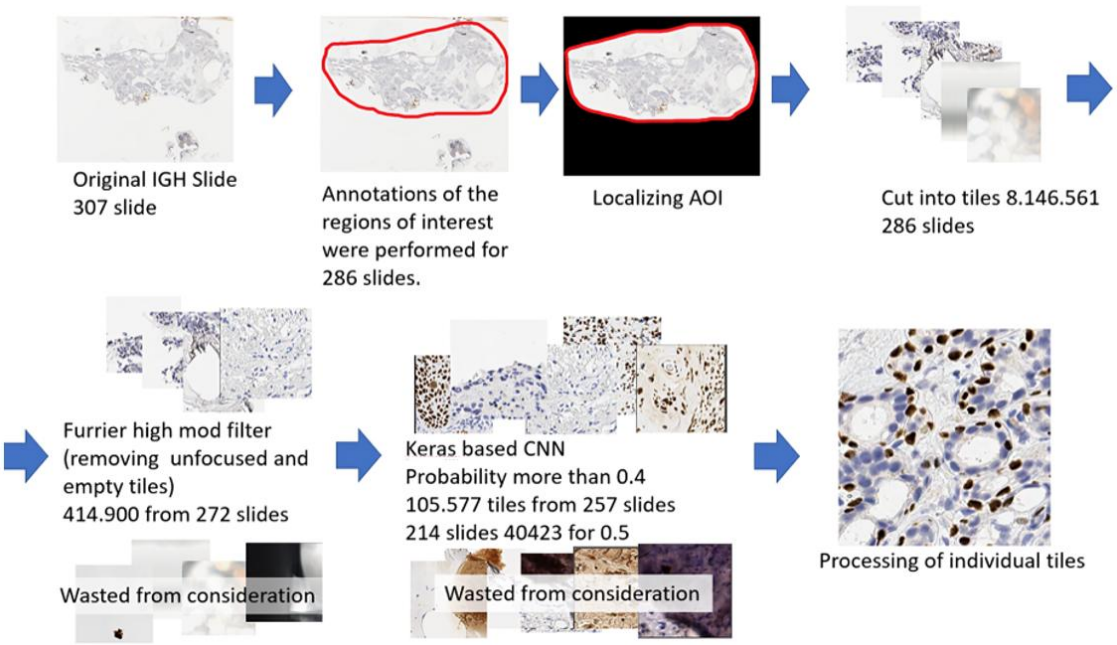

Figure 9. Tile-level workflow of the developed software (Source: Authors' own elaboration, using Python Plotly data visualization libraries)

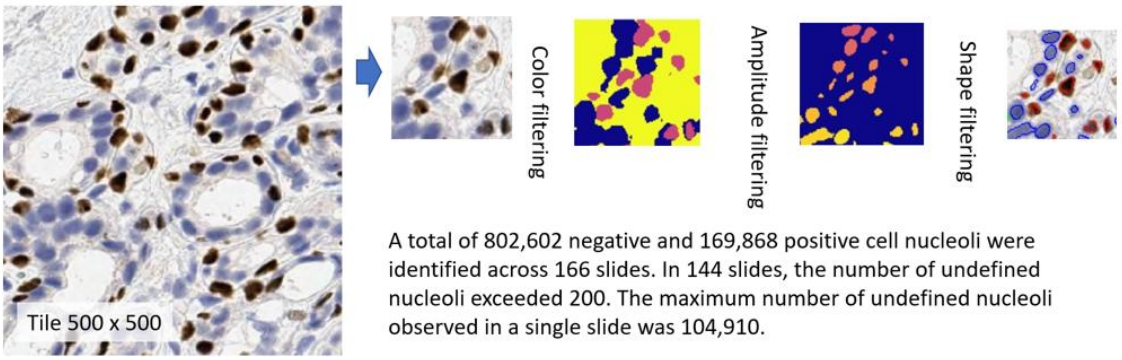

Figure 10. Individual tile processing pipeline (Source: Authors' own elaboration, using Python Plotly data visualization libraries)

The calculated dice coefficient was 94% for negative (blue) cells and 92% for positive (brown) cells. For negative cells, the most frequent source of error (9%) was class 1, primarily due to suboptimal staining, as some cells lacked a clearly defined blue coloration. For positive cells, the dominant error source was class 2 (11%), which can be attributed to insufficient washing of the samples, resulting in debris exhibiting a pronounced brown color.

Figure 9 and **Figure 10** illustrate the workflow of the software developed. As shown, the pipeline follows a linear processing scheme in which unsuitable cases are progressively filtered out at each stage.

Initially, the dataset consisted of 330 slides. At the first step, pathologists selected ROIs for each slide. During this stage, slides that were clearly unsuitable for further analysis were excluded, such as empty slides or those in which a ROI could not be reliably identified. Subsequently, all remaining slides were divided into tiles, and further processing was performed at the tile level.

In total, 802,602 negative cell nucleoli and 169,868 positive cell nucleoli were annotated across 166 slides. Based on the outcome, all slides can be divided into three major groups:

1. Slides unsuitable for analysis, which were excluded and did not enter the final set of 166 slides.
2. Moderately suitable slides, containing fewer than 1,000 segmented nuclei (41 slides in total). The threshold of 1,000 nuclei is conditional; approximately from this number onward, the results can be considered statistically meaningful.
3. High-quality, suitable slides (125 slides), containing on average about 10,000 segmented nuclei.

This type of quantitative assessment shows good agreement with the evaluations provided by pathologists.

The maximum number of detected cell nuclei on a single slide was 104,910; however, this value is primarily attributable to the large size of the original slide and the extensive tissue area. On average, approximately 10,000 nuclei per slide represents a robust result, supporting the conclusion that the obtained data are statistically reliable.

DISCUSSION

The application of AI to image processing has revolutionized numerous fields, achieving remarkable success

in general computer vision tasks such as object detection, classification, and segmentation [12]. Modern deep learning systems now routinely demonstrate human-level or superior performance on benchmark datasets of natural images [13].

However, these impressive results have not fully translated to specialized domains, particularly in medical imaging, where unique and complex challenges persist.

A fundamental obstacle in developing robust AI systems for specialized applications lies in the scarcity of large, high-quality annotated datasets [14]. While crowdsourcing platforms have enabled efficient labeling of general images, this approach proves inadequate for medical imaging domains such as IHC analysis [15]. The annotation of IHC slides presents distinct challenges that require specialized expertise. Accurate labeling of cellular structures demands trained pathologists who can recognize histological artifacts and apply diagnostic criteria with precision. Unlike natural images, IHC annotations often require pixel-level accuracy for tasks such as nuclear segmentation or membrane staining quantification [16]. The scale of the work presents another barrier, as a single whole-slide image may contain over a million nuclei, making comprehensive annotation impractical from both time and resource perspectives.

The contrast between general and medical image annotation becomes particularly evident when examining practical examples. Labeling thousands of everyday images for object detection can be accomplished quickly and cost-effectively by non-experts. In stark contrast, annotating medical images requires significantly more time and resources. The process demands specialized knowledge, with costs often orders of magnitude higher than those for general image annotation due to the need for expert involvement. Commercial AI annotation services, while effective for many applications, face limitations when applied to medical imaging. These services typically employ annotators who lack specific training in histopathology, leading to potentially significant error rates when dealing with complex medical images. Quality control becomes particularly challenging in this context, as studies have demonstrated substantial discrepancies between annotations by non-experts and those by qualified pathologists [17]. Furthermore, regulatory requirements for medical AI tools often mandate that annotations be performed or verified by board-certified medical professionals, creating additional barriers to scalable solutions [18].

The practical constraints of engaging medical professionals in large-scale annotation projects present multiple challenges. Clinicians already face substantial workloads, with the majority of their time dedicated to patient care responsibilities. The demanding nature of medical image annotation, requiring sustained concentration to distinguish subtle pathological features from artifacts, adds to the difficulty of securing expert participation. Institutional structures and economic factors further complicate matters, as healthcare systems typically do not allocate time or provide compensation for pathologists to engage in annotation work that falls outside direct clinical service [19].

Recent advances in ML have attempted to address these challenges through innovative approaches. Weakly supervised learning techniques aim to reduce reliance on exhaustive manual annotation by utilizing higher-level labels [20]. Self-supervised pretraining methods leverage unlabeled medical images to extract useful features without requiring extensive annotations [21]. Synthetic data generation offers another potential solution, creating artificial medical images with perfect annotations for training purposes [22]. However, these approaches continue to face limitations when dealing with the inherent variability of real-world medical images, rare morphological presentations, and regions dominated by technical artifacts.

The development of effective AI solutions for specialized medical imaging tasks like IHC analysis, therefore, requires careful consideration of these unique challenges. Balancing the need for expert-level accuracy with practical constraints on time and resources remains an ongoing challenge in the field. Future progress will likely depend on continued innovation in ML methodologies combined with thoughtful integration of clinical expertise and realistic assessment of implementation barriers in healthcare settings. The development of our AI-driven IHC analysis system successfully addressed two major challenges in computational pathology: reducing reliance on limited expert resources while maintaining diagnostic accuracy and overcoming significant data quality issues commonly encountered with clinical samples. Our hybrid approach combining classical image processing techniques with ML demonstrates that biologically relevant data can be extracted from suboptimal IHC slides, though several important limitations were identified that require further investigation.

A key achievement of this work was establishing an effective balance between automation and necessary expert oversight. The CNN substantially reduced the pathologist's workload by automatically filtering non-diagnostic tiles when using the 0.5 probability threshold. However, our analysis revealed an inherent trade-off between sensitivity and specificity when adjusting these thresholds, as lowering to 0.4 doubled usable tile yield but introduced false positives requiring algorithmic correction. This finding aligns with broader challenges in computational pathology, where excessive automation risks missing diagnostically critical features, while excessive manual review negates efficiency gains. Our implemented solution of targeted expert review at key decision points follows the emerging "human-in-the-loop" paradigm that is gaining acceptance in the field.

The study particularly highlighted data quality as a persistent and underappreciated bottleneck in digital pathology implementation. Despite standardized protocols, every batch contained poorly stained or washed slides exhibiting several problematic features. Most notably, we observed significant attenuation in the blue channel (approximately 40 units) that disrupted conventional RGB thresholding methods, necessitating the development of our specialized k-NN classifier. This finding correlates with previous reports of stain variability, though our cases demonstrated unusually severe channel-specific degradation. The modified Flood Fill algorithm proved particularly effective for challenging brown-on-brown staining scenarios where intensity-based methods failed, as shown in **Figure 6**. However, its performance degraded for weakly stained nuclei smaller than 10 pixels, suggesting potential benefits from incorporating emerging fluorescence-based IHC protocols.

Nuclear overlap represents another major challenge. We believe that overlapping nuclei constitute the majority of cases, indicating that nuclear crowding remains a significant obstacle for two-dimensional segmentation approaches. Recent advances in three-dimensional reconstruction

techniques may offer promising solutions to this persistent limitation.

The core issue lies in the inherently ambiguous nature of detecting and segmenting overlapping nuclei. As the number of intersecting cells increases, the ambiguity of the segmentation task grows accordingly. This presents a clear example where human analysis and AI-based processing may produce differing results. Moreover, even two human experts may interpret and process the same case differently, underscoring the subjective and complex nature of such tasks.

This poses a considerable challenge when attempting to compare data processed by AI and by human experts. In the case of human interpretation, it becomes necessary to introduce the concept of 'preference.' A human analyst may preferentially select specific data for processing based on convenience, familiarity, or interpretability. The same applies to AI: certain types of data may be inherently more difficult for the model to handle, leading it to avoid or misclassify such cases, even though they may not pose the same level of difficulty for a human.

Consequently, meaningful comparison between AI and human performance may ultimately require a detailed classification of all cell nuclei—not only in terms of standard categories such as positive/negative staining—but also concerning how they are 'preferred' or prioritized by both AI and human observers. This additional dimension of classification could help explain discrepancies in performance and provide a more nuanced understanding of the respective strengths and limitations of each approach.

The clinical and research implications of this work are noteworthy. The resulting dataset of 4 million annotated nuclei represents, to our knowledge, the largest available resource for IHC-specific CNN training for this particular cancer type. This resource shows strong potential for transfer learning applications, particularly for fine-tuning established models like HoVer-Net. Our findings regarding optimal tile size (500×500 pixels for 8GB RAM systems) may not generalize high-throughput clinical environments, suggesting the need for adaptive tiling strategies in future implementations. The collaborative framework we propose for multicenter validation addresses a critical need in the field for improving algorithm generalizability across institutions.

Several important limitations must be acknowledged. First, potential stain-specific biases mean the model's performance on non-NST subtypes requires additional validation. Second, hardware dependencies became apparent, with the Flood Fill algorithm requiring approximately three minutes per slide on our setup, suggesting GPU acceleration could provide significant improvements. Finally, the clinical impact of our 12-30% usable tile yield compared to traditional manual review must be rigorously evaluated in prospective trials.

This study makes several novel contributions to the field. We present the first systematic approach combining k-NN color correction with gradient-based nuclear segmentation specifically optimized for degraded IHC slides. Our work establishes a quantitative framework for balancing automation and accuracy in tile selection decisions. Furthermore, by openly sharing our extensive dataset, we enable community-driven development of improved analytical tools. While the pipeline successfully extracts valuable data from challenging samples, it also clearly identifies persistent obstacles that will require multidisciplinary solutions to overcome.

CONCLUSION

In conclusion, the AI/ML methods we used provide a powerful tool for automating and enhancing the spatial analysis of IHC images, except low quality images. They have the potential to revolutionize pathology by improving diagnostic accuracy, guiding treatment decisions, accelerating biomarker discovery, and providing a more comprehensive understanding of the TME.

Nevertheless, the primary challenge lies in a fundamental methodological discrepancy between how AI systems and human experts approach the analysis. AI relies on the processing of a very large number of cases—often exceeding 10,000 examples—to identify patterns and make decisions. Despite this extensive dataset, the accuracy of AI remains limited, and how to interpret this data is not clear yet. In contrast, human experts typically base their conclusions on a significantly smaller set of observations. However, they are able to achieve a level of diagnostic precision and contextual understanding that is currently unattainable for AI.

This contrast highlights a critical need to revise the methodological framework used in IHC when adapting it for AI-based analysis. Specifically, it is necessary to establish new standards, evaluation criteria, and decision thresholds that reflect the unique strengths and limitations of AI. These new methodological approaches should consider the statistical nature of AI decision-making, its dependence on large-scale pattern recognition, and its potential lack of nuanced understanding in complex or borderline cases.

However, overcoming the challenges related to data acquisition, annotation, model interpretability, generalizability, and standardization is crucial for widespread clinical adoption in low- and middle-income countries. Continued research and development in these areas will unlock the full potential of ML in transforming histopathology and improving patient care. The integration of clinical data and the development of more efficient and interpretable models will be key drivers of progress in this rapidly evolving field.

Author contributions: AU & AM: writing–original draft; **AU:** conceptualization and software; **BAI:** supervision; **ZB:** data curation and supervision; **ZR & YB:** data curation; & **AM:** writing–review & editing. All authors agreed with the results and conclusions.

Funding: This study was partially funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan grant AP19679717.

Ethical statement: The authors stated that no ethical approval was required for this study. The study was performed in accordance with the ethics standards of the participating institutions and the tenets of the Declaration of Helsinki. IHC images were shared by RSE "Medical Center Hospital of the President's affairs Administration of the Republic of Kazakhstan", and no patient data was collected or processed.

AI statement: The authors stated that no AI tool was used in writing this article

Declaration of interest: No conflict of interest is declared by the authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

1. WHO. Breast cancer. World Health Organization; 2025. Available at: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (Accessed: 22 June 2025).
2. Costa L, Kumar R, Villarreal-Garza C, et al. Diagnostic delays in breast cancer among young women: An emphasis on healthcare providers. Breast. 2024;73:103623. https://doi.org/10.1016/j.breast.2023.103623 PMid:38219460 PMCid:PMC10826418
3. Barrios CH. Global challenges in breast cancer detection and treatment. Breast. 2022;62 Suppl 1 (Suppl 1):S3-6. https://doi.org/10.1016/j.breast.2022.02.003 PMid:35219542 PMCid:PMC9097801
4. Zaha DC. Significance of immunohistochemistry in breast cancer. World J Clin Oncol. 2014;5(3):382-92. https://doi.org/10.5306/wjco.v5.i3.382 PMid:25114853 PMCid:PMC4127609
5. Clayton DA, Eguchi MM, Kerr KF, et al. Are pathologists self-aware of their diagnostic accuracy? Metacognition and the diagnostic process in pathology. Med Decis Making. 2023;43(2):164-74. https://doi.org/10.1177/0272989X221126528 PMid:36124966 PMCid:PMC9825636
6. Elhanani O, Ben-Uri R, Keren L. Spatial profiling technologies illuminate the tumor microenvironment. Cancer Cell. 2023;41(3):404-20. https://doi.org/10.1016/j.ccell.2023.01.010 PMid:36800999
7. Li M, Jiang Y, Zhang Y, Zhu H. Medical image analysis using deep learning algorithms. Front Public Health. 2023;11:1273253. https://doi.org/10.3389/fpubh.2023.1273253 PMid:38026291 PMCid:PMC10662291
8. Howat WJ, Blows FM, Provenzano E, et al. Performance of automated scoring of ER, PR, HER2, CK5/6 and EGFR in breast cancer tissue microarrays in the Breast Cancer Association Consortium. J Pathol Clin Res. 2014;1(1):18-32. https://doi.org/10.1002/cjp2.3 PMid:27499890 PMCid:PMC4858117
9. keras-team. keras-team/keras. Keras; 2025. Available at: https://github.com/keras-team/keras (Accessed: 22 June 2025).
10. Géron A. Hands-on machine learning with scikit-learn, keras, and tensorflow. O'Reilly; 2019.
11. Schmidt U, Weigert M, Broaddus C, Myers G. Cell detection with star-convex polygons. In: Frangi A, Schnabel J, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical image computing and computera intervention–MICCAI 2018. MICCAI 2018. Lecture notes in computer science(), vol 11071. Cham: Springer; 2018. p. 265-73. https://doi.org/10.1007/978-3-030-00934-2_30
12. Trigka M, Dritsas E. A comprehensive survey of deep learning approaches in image processing. Sensors (Basel). 2025;25(2):531. https://doi.org/10.3390/s25020531 PMid:39860903 PMCid:PMC11769216
13. Matsuo Y, LeCun Y, Sahani M, et al. Deep learning, reinforcement learning, and world models. Neural Netw. 2022;152:267-75. https://doi.org/10.1016/j.neunet.2022.03.037 PMid:35569196
14. Rao KN, Fernandez-Alvarez V, Guntinas-Lichius O, et al. The limitations of artificial intelligence in head and neck oncology. Adv Ther. 2025;42(6):2559-68. https://doi.org/10.1007/s12325-025-03198-4 PMid:40299277 PMCid:PMC12085315
15. Irshad H, Oh E-Y, Schmolze D, et al. Crowdsourcing scoring of immunohistochemistry images: Evaluating performance of the crowd and an automated computational method. Sci Rep. 2017;7:43286. https://doi.org/10.1038/srep43286 PMid:28230179 PMCid:PMC5322394
16. Kataria T, Rajamani S, Ayubi AB, et al. Automating ground truth annotations for gland segmentation through immunohistochemistry. Mod Pathol. 2023;36(12):100331. https://doi.org/10.1016/j.modpat.2023.100331 PMid:37716506
17. Sylolypavan A, Sleeman D, Wu H, Sim M. The impact of inconsistent human annotations on AI driven clinical decision making. NPJ Digit Med. 2023;6(1):26. https://doi.org/10.1038/s41746-023-00773-3 PMid:36810915 PMCid:PMC9944930
18. WHO. Regulatory considerations on artificial intelligence for health. World Health Organization; 2023. Available at: https://iris.who.int/bitstreams/ad62580f-540f-4e36-b957-e7f2946ae1fb/download (Accessed: 26 June 2025).
19. Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: A perspective for healthcare organizations. Artif Intell Med. 2024;151:102861. https://doi.org/10.1016/j.artmed.2024.102861 PMid:38555850
20. Otesteanu CF, Ugrinic M, Holzner G, et al. A weakly supervised deep learning approach for label-free imaging flow-cytometry-based blood diagnostics. Cell Rep Methods. 2021;1(6):100094. https://doi.org/10.1016/j.crmeth.2021.100094 PMid:35474892 PMCid:PMC9017143
21. Huang S-C, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: A systematic review and implementation guidelines. NPJ Digit Med. 2023;6(1):74. https://doi.org/10.1038/s41746-023-00811-0 PMid:37100953 PMCid:PMC10131505
22. Pezoulas VC, Zaridis DI, Mylona E, et al. Synthetic data generation methods in healthcare: A review on open-source tools and methods. Comput Struct Biotechnol J. 2024;23:2892-910. https://doi.org/10.1016/j.csbj.2024.07.005 PMid:39108677 PMCid:PMC11301073