Journal of Clinical and Nursing Research, 2026, Volume 10, Issue 2
http://ojs.bbwpublisher.com/index.php/JCNR
Online ISSN: 2208-3693
Print ISSN: 2208-3685

Research on Automatic Identification of Colorectal Cancer Cells Based on Machine Learning Strategies and Analysis of their Morphological Heterogeneity and Prognostic Value

Yanna Ding

Zhongshan City People's Hospital, Zhongshan 528403, Guangdong, China

Abstract: In the fast-paced living environment, changes in dietary patterns have led to a continuous increase in the incidence and mortality rates of colorectal cancer (CRC), making it a prevalent malignant tumor of the digestive system worldwide. Currently, CRC clinical diagnosis and treatment face challenges such as high costs and persistently high recurrence rates. Traditional quantification of tumor-infiltrating lymphocytes (TILs) relies on manual analysis and judgment, resulting in low diagnostic efficiency and susceptibility to subjective factors, leading to missed or misdiagnosed cases. To enhance the efficiency and quality of CRC clinical diagnosis and treatment, this study explores domestic and international research on the automatic identification of CRC cells using machine learning strategies. It analyzes the morphological heterogeneity and prognostic value in the application of this strategy, aiming to deepen the understanding of intelligent tool applications in precise diagnosis, treatment, and prognostic evaluation of colorectal cancer, comprehend the current research status and development trends, and provide references for addressing and addressing the gaps in related research.

Keywords: Machine learning; Colorectal cancer cells; Automatic identification; Morphological heterogeneity

Online publication: Mar 10, 2026

1. Introduction

In recent years, the number of colorectal cancer (CRC) patients in China has increased significantly, and the burden of this disease has continued to rise with the aging population and changes in lifestyle, posing a severe challenge to the public health system in terms of its diagnosis and treatment demands. The state of the tumor immune microenvironment, particularly the distribution and density of tumor-infiltrating lymphocytes (TILs), has emerged as a crucial biomarker for predicting CRC recurrence and survival outcomes [1]. However, traditional TIL

quantification methods suffer from poor diagnostic performance, making large-scale clinical application difficult. These methods also incur high detection costs, limiting their accessibility and preventing their widespread adoption in primary healthcare institutions. Against this backdrop, there is a need to focus on the application and analysis of advanced technologies. Machine learning-based automated analysis technology for Hematoxylin and Eosin (H&E)-stained tissue sections offers a scientific and precise new approach to address the inefficiency of traditional methods [2]. Building on this, the study reviews relevant literature, draws on previous research findings, and explores the association mechanism between morphological heterogeneity characteristics and the immune microenvironment, aiming to provide an intelligent technological solution for accurate CRC prognosis assessment.

2. Application of machine learning technology in colorectal cancer cell identification

In the clinical diagnosis of CRC patients, machine learning technology is applied to accurately identify colorectal cancer cells by enabling automated extraction and analysis of pathological features through algorithmic models. Traditional TIL detection primarily relies on the experience of clinicians, which is subject to subjective factors, resulting in low accuracy in pathological diagnosis and frequent occurrences of missed and misdiagnosed cases. By leveraging machine learning techniques, the model performs color space conversion from RGB to HSV, employs the Canny algorithm in morphological edge detection, and analyzes texture features through gray-level co-occurrence matrices to accurately capture the morphological differences between tumor-infiltrating lymphocytes and cancer cells. This method effectively integrates a multi-dimensional feature strategy, enhancing the objectivity and reproducibility of colorectal cancer cell identification. It provides a standardized technical framework for pathological analysis, suitable for large-scale clinical screening scenarios [3].

In practical applications, the machine learning model utilizes a hierarchical training mechanism for intelligent identification of colorectal cancer cells. During this identification process, it employs lightweight models such as Support Vector Machine (SVM) and Random Forest algorithms, learns features based on TIL regions annotated in pathological diagnoses, and efficiently analyzes parameters such as cell area and circularity, as well as regional-level texture features like tissue contrast and entropy. Compared to traditional immunohistochemical staining, the application of machine learning technology avoids excessive reliance on specific antibodies, simplifies the operational process, enables analysis based on routine H&E-stained sections without additional staining steps, shortens detection time, and reduces detection costs. Meanwhile, the model ensures the stability of detection algorithms across different pathological scenarios through cross-validation and external dataset testing, providing a reliable basis for clinical decision-making. Overall, the machine learning model is characterized by automated feature extraction, significantly reducing the analysis time for single sections, and is compatible with existing digital pathology systems in hospitals without requiring complex hardware support [4]. Leveraging powerful technological transformation capabilities, we enable grassroots hospitals to conduct TIL quantitative analysis at a low cost, providing standardized tools for precise prognostic evaluation of colorectal cancer. This drives the progression of tumor immune assessment from scientific research exploration to routine clinical application, gradually establishing it as a conventional method for CRC disease screening. This further highlight the clinical value and social significance of colorectal cancer cell identification and the application of machine learning techniques.

3. Extraction method for morphological heterogeneity characteristics of colorectal cancer cells

To accurately extract the morphological heterogeneity characteristics of colorectal cancer cells, the machine learning-based strategy primarily utilizes algorithmic models to analyze cell morphology in pathological images. It delves into the complex information of texture and spatial distribution to precisely evaluate the state of the tumor microenvironment. In traditional pathological analysis, manual interpretation based on clinical experience fails to comprehensively capture the heterogeneity characteristics of cells. However, relying on the intelligent processing capabilities of machine learning technology, key indicators reflecting tumor immune infiltration can be systematically extracted through multi-dimensional feature fusion. In H&E-stained sections, the algorithm filters the color gamut thresholds for TILs and cancer cells through color space conversion. Combined with morphological edge detection, it segments individual cell contours and calculates parameters such as cell area, circularity, and perimeter, providing morphological evidence for distinguishing lymphocytes from cancer cells. Among them, texture feature extraction is the core method for revealing cellular heterogeneity. The machine learning model utilizes the gray-level co-occurrence matrix to analyze the contrast, correlation, and entropy values in densely populated cellular regions, clarifying the texture differences between TIL aggregation areas and tumor cell areas [5]. High-density TIL areas typically exhibit higher contrast and entropy values, reflecting the degree of active lymphocyte infiltration. Dense tumor cell areas show low contrast and high correlation, indicating the tightness of cell arrangement. Texture-based feature extraction methods effectively compensate for the limitations of single morphological parameters and enhance the accuracy of heterogeneity identification. Additionally, the model employs Z-score normalization to eliminate the influence of feature scales, ensuring the comparability of features across different sections, thereby enabling accurate extraction of morphological heterogeneity in colorectal cancer cells.

Furthermore, in the application and analysis of machine learning models, spatial distribution feature extraction methods can be employed to expand the analytical dimensions of morphological heterogeneity in colorectal cancer cells. Based on machine learning strategies, tumor areas and the junction areas between tumors and normal tissues are accurately labeled. Combined with a block-based feature extraction strategy, the density and proportion of TILs in different tissue regions are analyzed. Machine learning algorithms can output indicators such as TIL density within the tumor area and the proportion of TILs in the junction area, revealing the spatial patterns of immune cell infiltration. Feature extraction based on spatial distribution provides a scientific tool for evaluating the heterogeneity of the tumor immune microenvironment. By conducting correlation analyses with clinical pathological parameters such as TNM staging and differentiation degree, the potential associations between morphological heterogeneity in colorectal cancer cells and prognostic outcomes can be explored, thereby strengthening technical support for precision diagnosis and treatment of CRC patients [6].

4. Machine learning model construction and recognition performance optimization strategies

In the application of machine learning technology, it is necessary to select an algorithmic framework suitable for the characteristics of pathological images by considering the features of colorectal cancer cell identification and testing. Then, construct corresponding models and perform parameter tuning to maximize the performance of the technical model. Judging from the performance of traditional deep learning models, they excel in image

recognition but rely heavily on large-scale labeled data. Given the complex glandular structures and easily confused cell types in colorectal cancer pathology slides, the performance of these models cannot be fully leveraged, indicating their limited universality and certain constraints [7]. Therefore, this study employs lightweight models, such as Support Vector Machine (SVM) and Random Forest, combined with key pathological regions like tumor areas and junctional zones. Through feature learning, the recognition performance of machine learning models for tumor cells is optimized. These models utilize a highly interpretable feature selection mechanism, taking into full consideration different parameters such as color, morphology, and texture. A block-based feature extraction strategy is adopted to enhance the model's adaptability to complex tissue structures and improve its general applicability. To optimize the model's recognition performance, it is necessary to approach from both feature engineering and model training. Image texture features should undergo Z-score normalization to eliminate dimensional differences and avoid affecting the model's recognition performance. Model training should employ 5-fold cross-validation to adjust hyperparameters, such as the SVM penalty coefficient C and the tree depth in Random Forest. The generalization ability of the model is optimized through the division of sub-training and sub-validation sets. The Random Forest model selects the optimal feature subset using the Gini coefficient to mitigate the risk of overfitting, while the SVM model optimizes decision boundaries through the RBF kernel function to enhance classification accuracy in complex scenarios.

To ensure the recognition performance of machine learning models and achieve performance iteration, model validation is essential. It has been reported that selecting an independent test set to validate model accuracy, combined with Kappa consistency testing, can evaluate the alignment between AI quantitative results and manual scores from clinical diagnoses. To address stability issues in model application, external validation strategies are employed, incorporating multi-center data to test the model's adaptability across different pathological equipment and staining conditions. Meanwhile, through feature interpretability analysis, SHAP values are calculated to reveal the contribution of key features to recognition outcomes, providing direction for model optimization. For instance, if texture features contribute more to TIL recognition than color features, texture analysis dimensions can be refined by adding directional gray-level co-occurrence matrices to enhance the model's ability to distinguish heterogeneous cells, ensuring robust performance in complex pathological scenarios and providing reliable technical support for clinical applications [8].

5. Validation analysis of the prognostic value of morphological heterogeneity parameters

Accurate diagnosis of pathological changes and disease severity in CRC patients aids in the formulation and implementation of targeted treatments and facilitates scientific and rational clinical decision-making. By employing machine learning techniques for the automatic recognition of colorectal cancer cells and analysis of morphological heterogeneity, the distribution of cells within the tumor microenvironment can be clarified, morphological features can be extracted, and their potential associations with patient survival outcomes can be revealed. Traditional prognostic assessment relies on clinical indicators such as pathological staging and degree of differentiation, while the application of machine learning models enables the extraction of TIL density and texture heterogeneity, providing detailed information about the immune microenvironment. Among them, the indicators of TIL density within the tumor area and the proportion of TILs in the junction area between tumor and normal tissue can reflect the spatial pattern of lymphocyte infiltration, offering a scientific basis for evaluating tumor immune

escape mechanisms and thus predicting patient relapse risk and survival time [9].

To verify the prognostic value of machine learning technology application in CRC patients, the machine learning strategy employs multifactorial generative analysis and uses the Cox proportional hazards regression model to validate the independent predictive role of morphological heterogeneity parameters. Studies have shown that regions with high TIL density are often associated with lower tumor aggressiveness and lymph node metastasis rates. Among their corresponding morphological features, cell arrangement compactness and texture contrast can be accurately calculated through algorithmic models. Incorporating relevant parameters into prognostic assessment models improves the accuracy of predicting 3-year relapse-free survival and overall survival. For instance, a combined model integrating TIL density and TNM staging demonstrates superior prognostic stratification ability compared to single clinical indicators, providing a basis for formulating individualized treatment plans. The effectiveness of machine learning strategies in clinical application can also be reflected in the reproducibility and universality of prognostic assessment. Traditional manual scoring results in fluctuations due to inter-observer variability, whereas machine learning models ensure consistency of prognostic parameters across different pathological scenarios through standardized feature extraction and quantitative output. In addition, the model was validated using external datasets in a multi-center collaborative manner, confirming the stability of morphological heterogeneity parameters across different populations. The algorithm-based prognostic evaluation system provides new technical means and tools for precise diagnosis and treatment of colorectal cancer (CRC). It accurately assesses the dynamic changes in the tumor immune microenvironment, enabling prognostic model evaluation to be based not on clinical experience but on real data-driven accurate assessments [10].

6. Conclusion

In summary, due to the low efficiency, inadequate diagnostic accuracy, and poor clinical utility of traditional tumor-infiltrating lymphocyte (TIL) quantification methods, they struggle to meet the demands for efficient and precise diagnostic identification in CRC clinical practice. To achieve rapid automatic quantification, the promotion and application of machine learning technology hold significant and far-reaching practical importance. By leveraging its intelligent algorithm model performance, machine learning strategies can automate the extraction of pathological features, accurately capture morphological differences between cancer cells and TILs, enhance the objectivity and reproducibility of colorectal cancer cell identification, reduce detection costs and time, and provide standardized tools for clinical screening. In the extraction of morphological heterogeneity features, multidimensional fusion feature analysis can compensate for the limitations of single-parameter detection and expand the analytical dimensions of CRC pathological diagnosis through innovative approaches and methods based on spatial distribution feature extraction. Additionally, by combining the prognostic stratification capabilities of the model and relying on its universality and accuracy in prognostic evaluation, new ideas and methods for precise diagnosis and treatment of CRC can be explored.

Disclosure statement

The author declares no conflict of interest.

References

[1] Huang S, Chen G, 2025, Construction of an Interpretable Risk Prediction Model for Colorectal Sessile Serrated Lesions Based on Machine Learning Algorithms. Journal of Xuzhou Medical University, 45(10): 753–759.

[2] Shi Y, Ma C, Tang X, et al., 2025, Ursolic Acid Induces Ferroptosis in Colorectal Cancer Cells via Regulation of System xc− and the miR-214-3p/Stat3/GPX4 Axis. Frontiers in Immunology, 16: 1674321–1674321.

[3] Li R, Qin W, 2025, From Pathological Images to Molecular Prediction: A New Paradigm for MSI Detection in Colorectal Cancer Driven by Multimodal Artificial Intelligence. Chinese Journal of General Surgery, 34(10): 2232–2242.

[4] Xu S, Zhang H, 2025, Advances in the Application of Artificial Intelligence in Immunotherapy for Colorectal Cancer. Journal of Colorectal and Anal Surgery, 31(4): 303–307.

[5] Liu J, Li L, Zhang L, et al., 2025, Epigenetic Regulation of NR5A2 Influences Colorectal Cancer Cell Stemness via a Stemness-Related Transcription Factor NANOG. Cell Communication and Signaling, 23(1): 484–484.

[6] Li J, Li S, Sun W, et al., 2025, Construction and Evaluation of a CT-Based Differential Diagnosis Model for Ulcerative Colitis-Associated Dysplasia and Colorectal Cancer Using Machine Learning Algorithms. Clinical Medicine, 45(8): 74–79.

[7] Li B, Liu G, 2025, Predictive Value of a Machine Learning Model Combining MRI Radiomics With Clinical Features for Liver Metastasis in Colorectal Cancer. Chinese Journal of General Surgery, 34(7): 1410–1420.

[8] Xiang J, Xu J, Fan H, et al., 2025, PARP Inhibitor BMN673 Triggers PARylation-Mediated ATF4-GDF15 Pathway to Drive Autophagy and Ferroptosis in ATM-Deficient Colorectal Cancer Cells. Molecular Biomedicine, 6(1): 113–113.

[9] Wang Z, Wang W, Wang D, et al., 2025, Comparison of Deep Learning and Machine Learning Models in Preoperative Prediction of Mismatch Repair System Status in Colorectal Cancer Based on CT Radiomics. Journal of Molecular Imaging, 48(3): 315–322.

[10] Li Z, Ma R, Wang Y, et al., 2025, Application of a Radiomics Machine Learning Model and Nomogram Based on Spectral CT for Preoperative Differentiation of KRAS Gene Status in Colorectal Cancer. International Journal of Medical Radiology, 48(2): 151–158.