Learning cancer grade in an automated manner

Abstract:

Owing to the molecular heterogeneity of bladder cancers in general, and urothelial carcinoma in particular, it often manifests differently in different patients based on the underlying genetic mutations. And this manifestation is often variable across stages and grades of the disease. The treatment strategies for early-stage tumors are often very different from late-stage ones; likewise, those of metastasized growths are different from localized ones. Muscle invasive tumors often require chemotherapy, whereas for non-muscle invasive ones the preferred course of treatment is surgical intervention, specifically resection. In this project, we employ a genetic algorithm-based approach to synthetically mimic tumor regions of interest from healthy tissue segments, in the process learning characteristic features pertinent with specific stages as they manifest on the CT scans. The proposed method detects cancer stage from images based on cancer and healthy segments. The results are <describe results>.

The proposed method performs four classification tasks: (1) differentiating non-muscle-invasive from muscle-invasive bladder cancer, (2) categorizing cancer by stage, (3) distinguishing early-stage from late-stage cases, and (4) identifying post-treatment changes versus muscle-invasive cancer.

Introduction:

Bladder cancer is a prevalent and biologically heterogeneous malignancy, exhibiting a wide spectrum of tumor characteristics that contribute to diverse clinical outcomes [1]. It is broadly classified into non-muscle invasive bladder cancer (NMIBC) and muscle invasive bladder cancer (MIBC), each with distinct prognostic and therapeutic implications. NMIBC, encompassing stages Ta, Tis, and T1, is generally associated with a lower risk of recurrence and is typically managed through transurethral resection of bladder tumor (TURBT) and intravesical therapies [2]. In contrast, MIBC, which includes stages T2, T3, and T4, is characterized by a higher propensity for metastasis and requires more aggressive treatment strategies, such as radical cystectomy and systemic chemotherapy [3]. Additionally, tumor grading plays a crucial role in clinical decision-making. Low-grade tumors exhibit slow growth and are usually managed with endoscopic surveillance and intervention, posing minimal immediate threat to patients. Conversely, high-grade tumors are more aggressive, demonstrating a higher likelihood of malignant progression and increased cancer-related mortality. Consequently, optimal treatment strategies are contingent upon both the stage and grade of the tumor, indicating the need for accurate and reliable diagnostic tools [4].

Various conventional machine learning and deep learning techniques have been employed to identify and classify bladder cancer using medical imaging. The following two papers examine different methodological approaches to this task. The first study employs texture-based feature extraction combined with traditional machine learning classifiers. The second study integrates pre-trained deep learning models for feature extraction, followed by conventional machine learning techniques for classification. In the first study [5], intensity and texture features derived from Local Binary Patterns (LBP) and the Grey Level Co-occurrence Matrix (GLCM) were extracted. A Support Vector Machine (SVM) classifier was then applied to preoperative T2-weighted MRI scans from 65 consecutive patients undergoing radical cystectomy. The algorithm achieved a sensitivity of 74.2%, a specificity of 82.4%, an accuracy of 78.5%, and an area under the curve (AUC) of 80.6% at the patient level. In the second study [7], the authors employed a pre-trained ResNet-18 model with 71 layers for feature extraction, followed by classification using five machine learning algorithms: k-nearest neighbor (KNN), support vector machine (SVM), linear discriminant analysis (LDA), decision tree (DT), and naïve Bayes (NB). These classifiers were applied to three distinct tasks: (1) distinguishing between bladder cancer tissue and normal tissue, (2) differentiating MIBC from NMIBC, and (3) detecting post-treatment changes (PTC) relative to MIBC. The proposed framework demonstrated robust performance, achieving high F1-scores across all classification tasks.

Accurate cancer staging through medical imaging plays a crucial role in guiding treatment decisions. However, distinguishing between different cancer stages and grades based on imaging features remains a challenging task. This study employs GLCM-based texture analysis to improve classification performance to extract features from cancerous and non-cancerous regions of interest (ROIs) in 100 bladder cancer CT scans. In addition to these baseline features, we incorporate three distinct outputs generated by a genetic algorithm, which are subsequently used as separate classification features. These features are then applied to four classification tasks using traditional machine learning techniques. By evaluating classification performance across these tasks, this study aims to provide insights into the effectiveness of texture-based features and hybrid feature generation methods in bladder cancer staging.

Methodology:  
  
This study follows a structured workflow, as illustrated in Figure 1, to analyze CT scans by identifying regions of interest (ROIs), extracting relevant features, and applying machine learning models for classification. Instead of examining the entire CT image, the focus is on the urinary bladder region, from which ROIs are extracted and analyzed. As a baseline, Gray-Level Co-Occurrence Matrix (GLCM) features are utilized for feature extraction. Furthermore, a genetic algorithm is used to generate synthetic cancer-like ROIs from healthy tissue sections, facilitating the identification of characteristic features associated with various cancer stages and grades. The genetic algorithm produces three key outputs: (1) the best distance between healthy and cancerous regions for a given number of iterations, (2) the mean distance between healthy and cancerous regions across iterations, and (3) the number of iterations required for a healthy region to converge to a cancerous one. These outputs are subsequently utilized to perform classification, with the aim of demonstrating that classification based on the genetic algorithm's outputs yields improved results compared to traditional feature-based classification.

1.1 Extracting spatially labeled ROIs using a sliding window implementation

In the initial phase of the study, the urinary bladder region was isolated from 100 CT scans utilizing ImageJ software. Within this delineated area, regions of interest (ROIs) were extracted, comprising 10, 20, 30, 40, and 50 healthy ROIs per scan alongside a single cancer ROI. The cancer ROIs were identified using the provided masks, while the healthy ROIs were generated through a sliding window approach. The sliding window technique involves moving a fixed-size window across the defined bladder region to extract multiple non-overlapping or slightly overlapping healthy ROIs. Figure X showcases sample images with annotated ROIs that correspond to various cancer stages.

* 1. Feature Extraction using GLCM  
       
     Six feature extraction techniques were evaluated: Histogram of Oriented Gradients (HOG), Gray-Level Co-occurrence Matrix (GLCM), Scale-Invariant Feature Transform (SIFT), Gradient Location and Orientation Histogram (GLOH), Fourier Transform, and Gabor Filters. Among these, GLCM was selected for two key reasons:
  2. Effective Texture Analysis – GLCM effectively captures the spatial relationships between pixel intensities, allowing for a detailed characterization of textural variations and directional patterns. This capability is particularly advantageous in medical imaging, where subtle texture differences can provide critical diagnostic insights.
  3. Consistent Feature Representation – Unlike some other texture analysis methods, GLCM produces a fixed number of features irrespective of the region of interest (ROI) size. This ensures a uniform feature representation across different ROIs, facilitating systematic comparisons and analyses of texture characteristics.

Using GLCM, 20 texture features were extracted from each ROI based on five key metrics: dissimilarity, correlation, energy, contrast, and homogeneity. To capture directional texture patterns, feature calculations were carried out at four angular orientations (0°, 45°, 90°, and 135°). Furthermore, for each orientation, features were computed across five pixel displacements (1, 2, 3, 4, and 5), representing varying distances between paired pixels. This structured approach provided a comprehensive analysis of texture characteristics across multiple spatial and directional configurations.

1.3Classification of Features  
  
To classify the extracted features, six supervised machine learning models were utilized: Support Vector Machine (SVM), Logistic Regression, k-nearest Neighbors (KNN), Decision Tree, Random Forest, and Linear Discriminant Analysis (LDA). Four distinct classification tasks were performed to evaluate the predictive capability of the extracted features:

* Cancer Invasion Classification: Differentiating between non-muscle-invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC).
* Cancer Stage Classification: Categorizing cancer into stages Ta, Tis, T1, T2, T3, and T4.
* Early-Stage vs. Late-Stage Classification: Grouping cases of early-stage bladder cancer (Ta, Tis) versus late-stage cases (T2, T3, T4).
* Post-Treatment Changes vs. MIBC Classification: Distinguishing between treatment-induced changes and muscle-invasive bladder cancer.

* 1. Classification using raw GLCM features.  
       
     The classification was independently performed for each of the five GLCM metrics: dissimilarity, correlation, energy, contrast, and homogeneity. In this phase, the extracted GLCM features were used as primary inputs for the machine learning models, establishing a baseline for comparison with classifications derived from genetic algorithm features. Unlike the genetic algorithm-based classification, this method focused solely on tumor ROIs, omitting the inclusion of healthy tissue regions. So the classification performance for each task remained the same across different ROIs.
  2. Classification based on the similarity between synthetic cancer tissue sections and real cancer Regions of Interest (ROIs)  
       
     This classification method employed two features derived from the genetic algorithm: best distance and mean distance, which were calculated over a fixed number of iterations. These features were assessed by analyzing the similarity between synthetic cancer tissue sections and tumor ROIs. For each GLCM metric—dissimilarity, correlation, energy, contrast, and homogeneity—the genetic algorithm was applied separately using varying sets of healthy ROIs (10, 20, 30, 40, and 50). For each set, the best distance and mean distance were computed independently, providing a comprehensive evaluation of how different texture descriptors contribute to classification performance.

* 1. Classification Based on Algorithm Convergence Time  
       
     A further classification feature was derived from the number of iterations—or maximum generations—required for the genetic algorithm to adjust a healthy region of interest (ROI) to resemble a tumor ROI. The total number of generations needed for this convergence was recorded as a classification feature. This metric was computed for different sets of healthy ROIs (10, 20, 30, 40, and 50) across all GLCM metrics.

Results:  
  
The performance of the proposed method and baseline approaches was evaluated using accuracy, F1-score, precision, recall and specificity, ensuring a rigorous assessment across all classification tasks. Given the limited dataset and potential class imbalances, F1-score was prioritized as the primary evaluation metric, as it provides a more reliable measure by balancing precision and recall.

For the classification of cancer stages, a 3-fold cross-validation strategy was employed to mitigate the effects of data scarcity while ensuring computational feasibility and result stability. For the remaining three classification tasks, 5-fold cross-validation was utilized, taking advantage of the larger sample sizes to enhance the reliability and robustness of performance estimates.

To assess the effectiveness of incorporating a genetic algorithm-based approach, its performance was compared against the baseline feature-based classification. While the genetic algorithm introduced novel features derived from synthetic cancer-like ROIs, its impact on classification varied across different tasks. The following sections provide a detailed breakdown of classification performance, analyzing the strengths and limitations of both approaches across different classification tasks.

Classification of cancer invasion (NMIBC vs. MIBC).

The experimental results demonstrated that GLCM features provided the best classification performance, with the correlation feature achieving the highest F1-score of 76.46\% (accuracy 78.46%, precision 79.17%, recall 76.75%, specificity 71.0%) using an SVM classifier across all ROIs, as illustrated in \textbf{\Cref{fig:cancer\_invasion}}. The best distance feature, derived through the genetic algorithm applied to the energy feature from 40 healthy ROIs per image, attained the second-highest F1-score of 75.36\% (accuracy 76.92%, precision 77.8%, recall 77.17%, specificity 76.0%) with a random forest classifier. The mean distance feature, obtained using the genetic algorithm on the contrast feature from 30 healthy ROIs per image, yielded an F1-score of 74.18\% (accuracy 75.38%, precision 75.0%, recall 75.14%, specificity 75.0%) with decision tree classifier. The maximum generation feature, computed via the genetic algorithm applied to the energy feature from 30 healthy ROIs per image, achieved an F1-score of 73.54% (accuracy 76.92%, precision 75.99%, recall 73.89%, specificity 65.0%) with a decision tree classifier.

(accuracy: %, precision: %, recall: %, specificity: %)

Cancer stages

The classification performance across all features was low due to class imbalance and the small size of the dataset. Among the features analyzed, the mean distance feature, derived through the genetic algorithm applied to the dissimilarity feature from 30 healthy ROIs per image, achieved the highest performance among all features, with an F1-score of 30.66\% (accuracy 25.04%, precision 34.32%, recall 30.66%, specificity 83.53%) using KNN, as illustrated in \textbf{\Cref{fig:cancer\_stage}}. The best distance feature, derived using the genetic algorithm with the energy feature from 10 ROIs as input, achieved an F1-score of 27.22\% (accuracy 29.03%, precision 36.7%, recall 30.95%, specificity 88.32%) with a decision tree classifier. For the maximum generation feature, the genetic algorithm applied to the correlation feature from 40 ROIs resulted in an F1-score of 26.31\% (accuracy 33.01%, precision 32.22%, recall 29.11%, specificity 88.23%) with decision tree classifier. Among the GLCM features, the correlation feature with KNN achieved an F1-score of 23.38\% (accuracy 40.05%, precision 25.05%, recall 24.89%, specificity 86.19%) across 10 and 50 ROIs.

Classification of early versus late-stage cancer

As shown in \textbf{\Cref{fig:cancer\_early\_vs\_late\_stage}}, the maximum generation feature, derived from the GA using the dissimilarity feature from 10 ROIs per image, achieved the highest classification performance with an F1 score of 76.27\% (accuracy 83.08%, precision 81.2%, recall 77.33%, specificity 66.67%) using random forest classifier. The GLCM energy feature attained an F1 score of 75.18\% (accuracy 80.0%, precision 74.42%, recall 77.67%, specificity 73.33%) across all the ROIs using SVM. The mean distance feature, obtained through GA using the dissimilarity feature from 10 ROIs, resulted in an F1 score of 73.95\% (accuracy 83.08%, precision 81.5%, recall 75.0%, specificity 60.0%) using KNN. The best distance feature, computed using GA with dissimilarity from 50 ROIs as input, achieved an F1 score of 70.41\% (accuracy 83.08%, precision 75.61%, recall 70.33%, specificity 46.67%) using KNN.

As shown in \Cref{fig:cancer\_early\_vs\_late\_stage }, the GA-derived maximum generation feature (using dissimilarity from 10 ROIs per image) achieved the highest classification performance, with an F1 score of 76.27% (accuracy: 83.08%, precision: 81.2%, recall: 77.33%, specificity: 66.67%) using a random forest classifier. The GLCM energy feature attained an F1 score of 75.18% (accuracy: 80.0%, precision: 74.42%, recall: 77.67%, specificity: 73.33%) across all ROIs with an SVM classifier. The GA-generated mean distance feature (derived from dissimilarity in 10 ROIs) resulted in an F1 score of 73.95% (accuracy: 83.08%, precision: 81.5%, recall: 75.0%, specificity: 60.0%) using KNN. Finally, the best distance feature (computed via GA with dissimilarity from 50 ROIs) achieved an F1 score of 70.41% (accuracy: 83.08%, precision: 75.61%, recall: 70.33%, specificity: 46.67%) with KNN.

Classification of PTC versus MIBC

Classification using GA-derived features outperformed those based on GLCM features. The best performance was achieved with the best distance feature, derived from GA using the homogeneity feature from 20 ROIs per image as input, attaining an F1-score of 73.98% (accuracy: 74.83%, precision: 76.05%, recall: 74.56%, specificity: 68.57%) using a random forest classifier and SVM, as illustrated in \Cref{fig:ptc\_vs\_mibc . The maximum generations feature, obtained through GA with the dissimilarity feature from 10 ROIs per image, achieved an F1-score of 72.96% (accuracy: 73.67%, precision: 74.18%, recall: 73.13%, specificity: 65.71%) using a random forest classifier. Similarly, the mean distance feature, derived from GA using the correlation feature from 20 ROIs, resulted in an F1-score of 69.94% (accuracy: 71.0%, precision: 74.59%, recall: 72.06%, specificity: 88.57%) with a KNN classifier. In comparison, the highest F1-score obtained using the GLCM homogeneity feature was 66.56% (accuracy: 67.42%, precision: 68.63%, recall: 67.76%, specificity: 68.57%), achieved with SVM across all ROIs.

1.1 Baseline Performance (GLCM Features)

The baseline classification model relied on GLCM features extracted from the ROIs, including dissimilarity, correlation, energy, contrast, and homogeneity. Among these, correlation emerged as the most predictive feature, achieving the highest F1-score compared to other features.

1.2 Performance Using Genetic Algorithm Features

In the genetic algorithm-based approach, two features derived from the algorithm were used for classification:

* 1. Best Distance: Among the GLCM metrics evaluated (e.g., dissimilarity, correlation, energy, contrast, and homogeneity), best distance using correlation consistently achieved better results compared to other features.
  2. Mean Distance: Among the GLCM metrics evaluated (e.g., dissimilarity, correlation, energy, contrast, and homogeneity), mean distance using correlation consistently achieved better results compared to other features.
  3. Maximum Generations: When using the maximum number of iterations required by the genetic algorithm for a synthetic cancer region to converge into a cancer region, the energy-based feature yielded the best classification performance in terms of F1 score.

1. The classification performance across all features was low due to class imbalance and the small size of the dataset. Among the features analyzed, the maximum generations feature, derived through the genetic algorithm applied to the correlation feature from 40 healthy ROIs per image, achieved the highest performance among all features, with an F1-score of 36.81% using KNN. Among the GLCM features, the correlation feature with KNN achieved an F1-score of 34.24% across 10, 20, 30, and 50 ROIs. For the mean distance feature, the genetic algorithm applied to the dissimilarity feature from 10 ROIs resulted in an F1-score of 33.6% with KNN. The best distance feature, derived using the genetic algorithm with the energy feature from 30 ROIs as input, achieved an F1-score of 31.88% with a random forest classifier.

Discussions:

The findings of this study are influenced by several limitations, particularly the dataset size and class distribution. The dataset consisted of 100 CT images spanning seven cancer stages, with an imbalanced class distribution that affected stage-wise classification performance. A stratified k-fold cross-validation approach was implemented to ensure proportional representation across training sets. However, this method did not fully address the challenge posed by underrepresented classes (Ta, Tis, and T4), which likely impacted classification performance.

The use of Gray-Level Co-occurrence Matrix (GLCM) features introduced constraints due to their sensitivity to pixel relationships at predefined angular orientations (0°, 45°, 90°, and 135°). Conventional data augmentation methods, such as rotation and flipping, did not alter pixel-pair relationships, limiting their effectiveness in enhancing model generalization. Additionally, this study focused on five selected GLCM metrics, and the exclusion of other texture descriptors may have reduced the model’s ability to capture more complex texture patterns.

To address these limitations, expanding the dataset is essential to improve class balance and support deep learning-based methods. Handcrafted feature extraction methods, while useful, may not fully capture intricate texture variations. Deep learning approaches, particularly Convolutional Neural Networks (CNNs), offer a potential alternative for feature extraction. CNNs can learn hierarchical texture and structural representations, which could enhance classification performance. However, traditional CNN architectures often require fixed input sizes, presenting challenges when dealing with varying region of interest (ROI) dimensions across different cancer stages.

Future work could focus on integrating adaptive CNN architectures that accommodate varying ROI sizes while leveraging both handcrafted and deep learning-based features. Additionally, alternative feature descriptors, such as wavelet-based texture analysis or frequency-domain representations, could be explored to enhance model robustness. Incorporating larger and more diverse datasets will be crucial for improving classification performance and ensuring the generalizability of the proposed approach.

This study was limited by the small dataset size, consisting of only 100 CT images distributed across seven cancer stages. The class distribution was imbalanced, affecting stage-wise classification. A stratified k-fold cross-validation approach was used to ensure proportional representation across training sets. However, this strategy could not fully compensate for the lack of sufficient training examples, particularly in underrepresented classes (Ta, Tis, and T4), which likely impacted overall classification performance.

GLCM features were also constrained by their sensitivity to pixel relationships at specific angular orientations (0°, 45°, 90°, and 135°). Conventional data augmentation techniques (e.g., rotation, flipping) were ineffective because they did not alter pixel-pair relationships, limiting the ability to improve model generalization through augmentation. Additionally, this study focused on five key GLCM metrics. The exclusion of other texture descriptors may have restricted the model’s ability to capture more complex texture variations.

To improve the proposed approach, expanding the dataset is necessary to enhance model generalization by providing better class balance and enabling deep learning-based methods. Given the limitations of handcrafted features, deep learning techniques, particularly Convolutional Neural Networks (CNNs), could be explored for feature extraction. CNNs learn hierarchical texture and structural features, which may lead to better classification performance. Traditional CNN architectures often require fixed input sizes, making them less adaptable to varying ROI dimensions across different cancer stages.

The objective of this study was to investigate whether GA-based features performed better than raw GLCM features and, if so, which GA-derived feature performed best for each classification task. Additionally, we examined the impact of the number of healthy ROIs used. Overall, GA-based features outperformed raw GLCM features in three out of four tasks. Among the raw GLCM features used as inputs for GA, dissimilarity, correlation, and energy tended to perform better compared to contrast and homogeneity. However, there was significant variability in the number of ROIs that gave the best performance. Since GA-derived features such as best distance, max generations, and mean distance were computed based on different raw GLCM inputs, it remains unclear whether a particular GLCM feature consistently led to superior performance. Given the limited and highly imbalanced dataset, forming a concrete conclusion regarding the impact of specific GLCM features or the number of ROIs remains challenging.

Conclusions:

References

1. Wong, V.K., et al., *Imaging and Management of Bladder Cancer.* Cancers, 2021. **13**(6): p. 1396.

2. Chang, S.S., et al., *Diagnosis and Treatment of Non-Muscle Invasive Bladder Cancer: AUA/SUO Guideline.* J Urol, 2016. **196**(4): p. 1021-9.

3. Chang, S.S., et al., *Treatment of Non-Metastatic Muscle-Invasive Bladder Cancer: AUA/ASCO/ASTRO/SUO Guideline.* J Urol, 2017. **198**(3): p. 552-559.

4. Kirkali, Z., et al., *Bladder cancer: epidemiology, staging and grading, and diagnosis.* Urology, 2005. **66**(6 Suppl 1): p. 4-34.

5. Tong, Y., et al., *Radiomics-guided therapy for bladder cancer: Using an optimal biomarker approach to determine extent of bladder cancer invasion from t2-weighted magnetic resonance images.* Adv Radiat Oncol, 2018. **3**(3): p. 331-338.

6. Yuhan Yang, X.Z., Yixi Wang,Xuelei Ma, *Application of deep learning as a noninvasive tool to differentiate muscle-invasive bladder cancer and non–muscle-invasive bladder cancer with CT* 2021.

7. Sarkar, S., et al., *Performing Automatic Identification and Staging of Urothelial Carcinoma in Bladder Cancer Patients Using a Hybrid Deep-Machine Learning Approach.* Cancers (Basel), 2023. **15**(6).

**5. Discussion**

**5.1 Interpretation of Results**

This study focused on the urinary bladder region in CT scans, selecting **10–50 control ROIs** to model tumor-like regions via a genetic algorithm. **GLCM-based classification** effectively captured texture variations, serving as a baseline. The genetic algorithm introduced complexity by simulating the transition from healthy to tumor tissue, with performance influenced by the similarity between selected ROIs.

**5.2 Limitations**

**5.2.1 Limited Dataset and Class Imbalance**

The dataset comprised **100 CT images across seven cancer stages**, with severe class imbalance (e.g., **Ta: 6, Tis: 9, T4: 4 samples**). Despite **stratified k-fold cross-validation**, minority classes remained challenging to classify.

**5.2.2 Feature Extraction and Augmentation Challenges**

GLCM features, computed at **0°, 45°, 90°, and 135°**, were invariant to common augmentation techniques like rotation and flipping, limiting data expansion strategies.

**5.2.3 Constraints in GLCM Feature Selection**

Only five GLCM metrics (**dissimilarity, correlation, energy, contrast, homogeneity**) were used, excluding others like **entropy and variance**, which could improve texture representation. Similarly, **five pixel displacements** may have restricted spatial analysis.

**5.3 Future Directions**

**5.3.1 Dataset Expansion**

A larger dataset would enhance class balance and enable **deep learning-based methods**, improving generalization.

**5.3.2 Deep Learning for Feature Extraction**

CNNs could replace handcrafted features, learning hierarchical representations. However, **ROI size variability** poses a challenge, which can be addressed through:

* **Fully Convolutional Networks (FCNs):** Replace fully connected layers with convolutions, allowing flexible input sizes.
* **Global Average Pooling (GAP):** Summarizes feature maps spatially, making CNNs adaptable to varying ROI dimensions.
* These results indicate that **GLCM features, particularly correlation, were the most effective for cancer invasion classification**. However, distance-based features derived using the genetic algorithm also demonstrated competitive performance, highlighting the potential of embedding-based similarity measurements in distinguishing cancerous regions from healthy tissue.

The objective of this study was to investigate if GA based features performed beteer compared to raw glcm features, if so which GA feature performed well for each of the tasks, also investigate the impact of number of healthy rois used. For cancer invasion (dataset included 24 non-muscle invasive bladder cancer (NMIBC) and 41 muscle-invasive bladder cancer (MIBC) images), raw correlation features performed better where for other 3 tasks GA based fatured performed better, mean distance performing better fro cancer stage (with Ta (6), Tis (9), T1 (9), T2 (13), T3 (24), and T4 (4) images), early versus later (15 early-stage and 50 late-stage images), max generations performed better and best distance performed better for ptc vs mibc classification(35 PTC and 41 MIBC images). There was a lot of variability in the number of rois considered and the raw glcm feature used for gGA across which was the best, with this limited dataset and highly unbalanced dataset its hard to form a concrete conclusion

for PTC vs. MIBC classification, the dataset contained 35 PTC and 41 MIBC images, which is slightly balanced compared to other 3 classification task, here ga featues outperformed the rwa glcm one, which is good thing, but we might need more data to confirm on the exact effect, also number of rois didn’t play a significant effect as the best performance varied across rois.

This study evaluated the effectiveness of genetic algorithm (GA)-based features compared to Gray-Level Co-occurrence Matrix (GLCM) features across multiple cancer classification tasks. The results demonstrated that GA-derived features generally outperformed traditional texture-based GLCM features, particularly in early vs. late-stage cancer classification and PTC vs. MIBC classification. Notably, the best distance and maximum generations features achieved the highest F1-scores in several tasks. However, the classification of cancer stages exhibited lower performance, likely due to class imbalance and the small dataset size.

The findings of this study are influenced by several limitations, particularly the dataset size and imbalanced class distribution. For cancer invasion classification, the dataset included 24 non-muscle invasive bladder cancer (NMIBC) and 41 muscle-invasive bladder cancer (MIBC) images. For cancer stage classification, the distribution was highly imbalanced, with Ta (6), Tis (9), T1 (9), T2 (13), T3 (24), and T4 (4) images. Similarly, for early vs. late-stage classification, there were 15 early-stage and 50 late-stage images, while for PTC vs. MIBC classification, the dataset contained 35 PTC and 41 MIBC images. A stratified k-fold cross-validation approach was implemented to ensure proportional representation across training sets. However, this method did not fully address the challenge posed by underrepresented classes (Ta, Tis, and T4), which likely impacted classification performance.

Additionally, the use of GLCM features introduced constraints due to their sensitivity to pixel relationships at predefined angular orientations (0°, 45°, 90°, and 135°). Conventional data augmentation methods, such as rotation and flipping, did not alter pixel-pair relationships, limiting their effectiveness in enhancing model generalization. Moreover, this study focused on five selected GLCM metrics, and the exclusion of other texture descriptors may have reduced the model’s ability to capture more complex texture patterns.

To address these limitations, expanding the dataset is essential to improve class balance and support deep learning-based methods. Handcrafted feature extraction methods, while useful, may not fully capture intricate texture variations. Deep learning approaches, particularly Convolutional Neural Networks (CNNs), offer a potential alternative for feature extraction. CNNs can learn hierarchical texture and structural representations, which could enhance classification performance. However, traditional CNN architectures often require fixed input sizes, presenting challenges when dealing with varying region of interest (ROI) dimensions across different cancer stages.

Future work could focus on integrating adaptive CNN architectures that accommodate varying ROI sizes while leveraging both handcrafted and deep learning-based features. Additionally, alternative feature descriptors, such as wavelet-based texture analysis or frequency-domain representations, could be explored to enhance model robustness. Incorporating larger and more diverse datasets will be crucial for improving classification performance and ensuring the generalizability of the proposed approach.

[GLCM](https://chatgpt.com/c/67b9f020-890c-800e-8a3f-069b34cca214)

This study investigated whether GA-based features outperformed raw GLCM features and assessed the impact of the number of healthy ROIs on classification performance. Overall, GA-based features performed better in three out of four tasks. Among the raw GLCM features used as inputs for GA, dissimilarity, correlation, and energy generally performed better than contrast and homogeneity. However, the best-performing GA-derived feature varied by task: mean distance for cancer stage classification, max generations for early vs. late-stage classification, and best distance for PTC vs. MIBC classification. Despite these findings, variability in the number of ROIs and the specific GLCM feature that yielded the best performance for GA made it difficult to establish a consistent pattern.

The study was constrained by dataset limitations, particularly class imbalance and the small bladder region available for analysis, which restricted the number of ROIs to 10, 20, 30, 40, and 50. Similarly, for GLCM computation, only displacement values of 1, 2, 3, 4, and 5 pixels were considered. While a stratified k-fold cross-validation approach was implemented to mitigate the effects of imbalance, underrepresented classes (such as Ta, Tis, and T4) remained a challenge.

The findings of this study are influenced by several limitations, particularly the dataset size and imbalanced class distribution. For cancer invasion classification, the dataset included 24 non-muscle invasive bladder cancer (NMIBC) and 41 muscle-invasive bladder cancer (MIBC) images. For cancer stage classification, the distribution was highly imbalanced, with Ta (6), Tis (9), T1 (9), T2 (13), T3 (24), and T4 (4) images. Similarly, for early vs. late-stage classification, there were 15 early-stage and 50 late-stage images, while for PTC vs. MIBC classification, the dataset contained 35 PTC and 41 MIBC images. A stratified k-fold cross-validation approach was implemented to ensure proportional representation across training sets. However, this method did not fully address the challenge posed by underrepresented classes (Ta, Tis, and T4), which likely impacted classification performance.

Additionally, the GLCM-based features also posed challenges due to their sensitivity to predefined angular orientations (0°, 45°, 90°, and 135°). Conventional data augmentation techniques, such as rotation and flipping, failed to alter pixel-pair relationships, limiting their effectiveness.

To address these challenges, future work should focus on expanding the dataset to improve class balance and support the development of more robust models. While handcrafted features like GLCM and GA-based attributes offer valuable insights, deep learning approaches, such as CNNs, could further enhance classification performance by capturing hierarchical texture and structural representations. Additionally, exploring adaptive models that accommodate varying ROI sizes may improve robustness and generalizability.

**Conclusion**

This study demonstrated the potential of GA-based features in bladder cancer classification, outperforming raw GLCM features in most tasks. However, dataset limitations, class imbalance, and constraints on ROI selection influenced the results. Future research should focus on larger datasets and advanced feature extraction techniques to improve classification performance and ensure model reliability.

The experimental results demonstrated that GLCM features provided the best classification performance. Specifically, the correlation feature achieved the highest F1-score of 76.46\% (accuracy: 78.46\%, precision: 79.17\%, recall: 76.75\%, specificity: 71.0\%) using an SVM classifier across all ROIs, as illustrated in \Cref{fig:cancer\_invasion}. The next highest performance was achieved by the \texttt{average\\_best\\_distance} feature with an F1-score of 75.36\%, followed by the \texttt{average\\_mean\\_distance} feature at 74.18\% and the \texttt{average\\_generation} feature at 73.54\%. A comparison of F1-scores for each feature is provided in the corresponding figure and table.

GA-derived features demonstrated superior performance compared to raw GLCM features. However, the overall classification performance across all features was limited due to class imbalance and the small size of the dataset. Among the features analyzed, the mean distance feature, derived through the GA applied to the dissimilarity feature from 30 healthy ROIs per image, achieved the highest performance, with an F1-score of 30.66\% (accuracy: 25.04\%, precision: 34.32\%, recall: 30.66\%, specificity: 83.53\%) using a KNN classifier, as illustrated in \Cref{fig:cancer\_stage}. The next highest performance was achieved by the  
\texttt{average\\_best\\_distance} feature with an F1-score of 27.22\%, followed by the \texttt{average\\_generation} feature at 26.31\% and the GLCM correlation feature at 23.38\%. A comparison of F1-scores for each feature is provided in the corresponding figure and table.

The best distance feature, derived using the GA with the energy feature from 10 ROIs as input, achieved an F1-score of 27.22\% (accuracy: 29.03\%, precision: 36.7\%, recall: 30.95\%, specificity: 88.32\%) with a decision tree classifier. Similarly, the maximum generation feature, obtained by applying the GA to the correlation feature from 40 ROIs, resulted in an F1-score of 26.31\% (accuracy: 33.01\%, precision: 32.22\%, recall: 29.11\%, specificity: 88.23\%) using a decision tree classifier. In contrast, among the GLCM features, the correlation feature achieved an F1-score of 23.38\% (accuracy: 40.05\%, precision: 25.05\%, recall: 24.89\%, specificity: 86.19\%) with a KNN classifier across 10 and 50 ROIs.

As shown in \Cref{fig:cancer\_early\_vs\_late\_stage}, the GA-derived maximum generation feature (using dissimilarity from 10 ROIs per image) achieved the highest classification performance, with an F1 score of 76.27\% (accuracy: 83.08\%, precision: 81.2\%, recall: 77.33\%, specificity: 66.67\%) using a random forest classifier. The GLCM energy feature achieved an F1 score of 75.18\% (accuracy: 80.0\%, precision: 74.42\%, recall: 77.67\%, specificity: 73.33\%) across all ROIs with an SVM classifier. The GA-generated mean distance feature (derived from dissimilarity in 10 ROIs) resulted in an F1 score of 73.95\% (accuracy: 83.08\%, precision: 81.5\%, recall: 75.0\%, specificity: 60.0\%) using KNN. Finally, the best distance feature (computed by GA with dissimilarity from 50 ROIs) achieved an F1 score of 70.41\% (accuracy: 83.08\%, precision: 75.61\%, recall: 70.33\%, specificity: 46.67\%) with KNN.

As shown in \Cref{fig:cancer\_early\_vs\_late\_stage}, the GA-derived \texttt{average\\_generation} feature (using dissimilarity from 10 ROIs per image) achieved the highest classification performance, with an F1 score of 76.27\% (accuracy: 83.08\%, precision: 81.2\%, recall: 77.33\%, specificity: 66.67\%) using a RF classifier. The next highest performance was achieved by the GLCM energy feature with an F1-score of 75.18\%, followed by the \texttt{average\\_mean\\_distance} feature at 73.95\% and the \texttt{average\\_best\\_distance} feature at 70.41\%. A comparison of F1-scores for each feature is provided in the corresponding figure and table.

Classification using GA-derived features outperformed those based on GLCM features. The best performance was achieved with the best distance feature, derived from GA using the homogeneity feature from 20 ROIs, attaining an F1-score of 73.98\% (accuracy: 74.83\%, precision: 76.05\%, recall: 74.56\%, specificity: 68.57\%) using a random forest classifier and SVM, as illustrated in \Cref{fig:ptc\_vs\_mibc}. The maximum generations feature, obtained through GA with the dissimilarity feature from 10 ROIs, achieved an F1-score of 72.96\% (accuracy: 73.67\%, precision: 74.18\%, recall: 73.13\%, specificity: 65.71\%) using a random forest classifier. Similarly, the mean distance feature, derived from GA using the correlation feature from 20 ROIs, resulted in an F1-score of 69.94\% (accuracy: 71.0\%, precision: 74.59\%, recall: 72.06\%, specificity: 88.57\%) with a KNN classifier. In comparison, the highest F1-score obtained using the GLCM homogeneity feature was 66.56\% (accuracy: 67.42\%, precision: 68.63\%, recall: 67.76\%, specificity: 68.57\%), achieved with SVM across all ROIs.

Classification using GA-derived features outperformed those based on GLCM features. The best performance was achieved with the best distance feature, derived from GA using the homogeneity feature from 20 ROIs, attaining an F1-score of 73.98\% (accuracy: 74.83\%, precision: 76.05\%, recall: 74.56\%, specificity: 68.57\%) using a random forest classifier and SVM, as illustrated in \Cref{fig:ptc\_vs\_mibc}. The next highest performance was achieved by the \texttt{average\\_generation} feature with an F1-score of 72.96\%, followed by the \texttt{average\\_mean\\_distance} feature at 69.94\% and the GLCM homogeneity feature at 66.56\%. A comparison of F1-scores for each feature is provided in the corresponding figure and table.

This study examined the effectiveness of GA-derived features compared to raw GLCM features and evaluated the impact of the number of healthy ROIs on classification performance. Overall, GA-based features outperformed raw GLCM features in three out of four classification tasks. Among the GLCM features used as inputs for GA, dissimilarity and homogeneity generally yielded better results than correlation, contrast, and energy. However, the best-performing GA-derived feature varied depending on the classification task: *average\_mean\_distance* for cancer stage classification, *average\_generation* for early vs. late-stage classification, and *average\_best\_distance* for PTC vs. MIBC classification. Despite these findings, the variability in the number of ROIs and the specific GLCM feature that produced optimal GA-based performance made it challenging to establish a consistent pattern across the classification tasks.