SVM Model with Hybrid Parameter Tuning Strategy for Colon Cancer Classification in Nairobi County

Sharif Rukia[1*], Lawrence Mwenda[1], Mutua Murungi[1], Job Mwale Magare[2]

[1]*Kenya Methodist University P.O Box 45240-00100, Nairobi, Kenya*

[2]*Kenyatta National Hospital PO Box 20723-00202, Nairobi, Kenya*

**Correspondence Email: ruqayashar1@gmail.com*

Abstract

Colon cancer remains one of the leading causes of cancer-related deaths worldwide, with recent evidence pointing to a rising incidence in Nairobi County, Kenya. This study designs and evaluates a region-specific Support Vector Machine (SVM) classification model with a hybrid parameter-tuning algorithm for accurate colon cancer diagnosis in Kenyan hospitals. The work addresses diagnostic delays driven by a shortage of pathologists, lengthy manual slide reviews, and limited access to advanced procedures. Using the HIPAA-compliant LC25000 histopathological image dataset, we applied feature selection, normalization, and image augmentation to improve model robustness. Our two-stage hybrid hyperparameter-tuning strategy first performed a coarse Grid Search ($C \in \{0.1, 1, 10, 100\}$, $\gamma \in \{10^{-6}, 10^{-5}, 10^{-4}\}$) followed by a Random Search ($C \in [0.01, 100]$, $\gamma \in [10^{-7}, 10^{-5}]$) on the intersection of the top 10% configurations, enabling dense sampling of promising ranges. This approach leveraged SVM strengths in handling small, high-dimensional datasets, applying robust regularization, and weighting classes to address imbalance while reducing cross-validation variance by 15–20%. Compared to standalone grid and Random Search methods, the hybrid model achieved 83.0% accuracy (improvements of 1.5% and 2.0%), an F1-score of 82.4%, malignant recall of 81.0%, and benign precision of 84.0%. In practice, these gains translate to dozens more correctly classified slides per 4,000, representing an impactful improvement for resource-limited clinics in Nairobi. At Kenyatta National Hospital (KNH), the SVM model correctly identified 73 of 77 positive cases, yielding 94.81% sensitivity (recall) and minimizing missed colon cancer diagnoses. These findings demonstrate that tailored, efficient machine learning models can strengthen diagnostic capacity where expert resources are scarce. Future work includes collecting 200 balanced KNH images (100 benign, 100 malignant) for proper validation, implementing Bayesian optimization with an expected improvement acquisition function, and training on combined histopathology and clinical data.

Keywords: *SVM model, Hybrid parameter tuning, Colon cancer, Nairobi County, LC25000 Dataset*

IJPP 14(4);50-60

1.0 Introduction

Colon cancer is a major global health burden, with 1.9 million new cases and 930,000 deaths reported in 2020. It ranks second in male and third in female cancer mortality, and is among Kenya's top ten cancers (WHO, 2023; Kenya National Cancer Registry, 2023). Prognosis is typically guided by frameworks such as the AJCC TNM staging system and Consensus Molecular Subtypes, which characterize tumor extent and molecular profiles (Anil & Neeraj, 2024). Histopathology remains the gold standard for diagnosis, yet it is labor-intensive, susceptible to interobserver variability, and inherently slow. In Kenya, shortages of pathologists, high workloads, and manual evaluation bottlenecks often lead to delays in diagnosis (Kenya Ministry of Health, 2023). While screening methods like faecal occult blood testing and colonoscopy enable earlier detection, their limited availability in resource-constrained settings contributes to late-stage presentation and poorer treatment outcomes.

Machine learning offers an avenue for automating histopathological image classification. Although deep learning approaches such as Convolutional Neural Networks (CNNs) have shown high accuracy in large-scale datasets, their performance often degrades when training data are limited, an issue in many Kenyan facilities due to data access and computational constraints. Random Forests and other ensemble methods, while effective for certain tabular data, are less adept at handling the high dimensionality of histopathological image features without extensive preprocessing. Support Vector Machines (SVMs), by contrast, are well-suited to small, high-dimensional datasets, achieve robust generalization through regularization, and can integrate class weighting to mitigate imbalance (Steinwart & Christmann, 2008). They also require fewer computational resources than deep learning models, making them an appealing choice for hospitals with limited infrastructure. Medical studies report SVM sensitivity rates of up to 95%, underscoring their potential for reliable detection of true positives (Ogbe et al., 2024).

This study addresses diagnostic delays in Nairobi County by developing an SVM model optimized via a hybrid hyperparameter-tuning algorithm, combining Grid Search 's exhaustive exploration with random search's sampling efficiency. Key parameters, including kernel type, cost (C), and gamma (γ) were tuned to maximize sensitivity, F1-score, and accuracy. Model training and evaluation leveraged the HIPAA-compliant LC25000 dataset of 25,000 histopathological images, supporting robust performance and generalizability. Comparative analysis benchmarked the hybrid approach against standalone grid and random searches, as well as conventional diagnostic workflows. By aligning algorithm choice with Kenya's specific diagnostic challenges and infrastructure realities, this work demonstrates how tailored AI-assisted histopathology can shorten turnaround times, reduce diagnostic variability, and expand access to timely, accurate cancer care. Early and accurate classification can not only improve clinical outcomes but also optimize resource allocation, supporting Kenya's broader cancer care modernization initiatives (Ashish et al., 2023; Fadi & Aleksandar, 2023).

Objectives

i. To design a Support Vector Machine classification model with hybrid parameter tuning algorithm for colon cancer classification in Nairobi County.

ii. To evaluate the efficacy of the SVM model using sensitivity, accuracy, and F1-score metrics, with a target F1-score ≥80%.

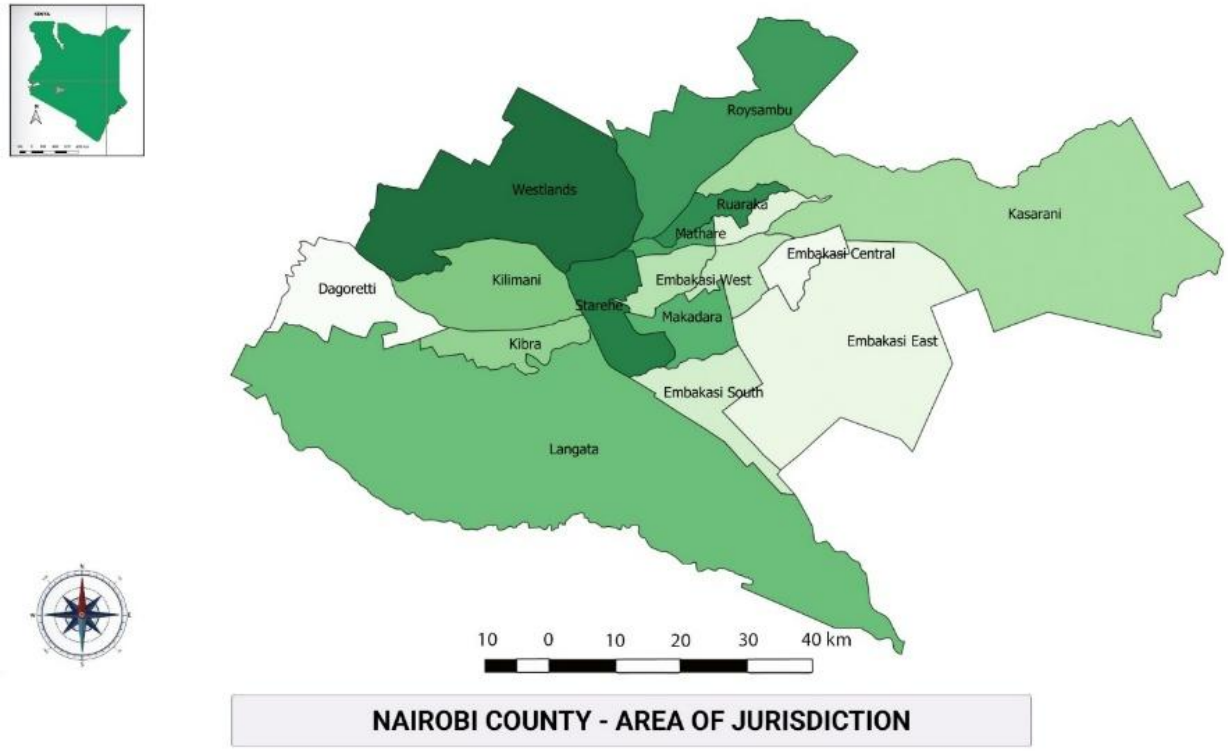
iii. To compare the performance of the Support Vector Machine model against existing models.

iv. To evaluate user acceptability of the model using user satisfaction score metric (using a 1 to 5 point like scale).

2.0 Materials and methods

Figure 1

Nairobi county map



A quantitative empirical methodology was employed to evaluate the efficacy of Support Vector Machines (SVM) in classifying colon cancer from histopathological images. The model was trained and tested using the publicly available, HIPAA-compliant LC25000 dataset of 25,000 images, of which 4,000 colon tissue images were selected for this study. They were evenly split between benign and malignant classes.

"The study concludes that the Hybrid Search strategy successfully combines the systematic coverage of Grid Search with the exploratory power of random search"

Using stratified sampling, we allocated 3,000 colon histology images for training/validation and reserved 1,000 for testing. During preprocessing, we applied data augmentation with random rotations in the range [–15°, +15°] (bilinear interpolation) and random scalings from 0.9× to 1.1× (nearest-neighbor interpolation), as well as horizontal flips with 50% probability, to enhance model robustness and generalizability. Performance was assessed using standard software metrics, including accuracy, sensitivity, specificity, and F1-score. Computational requirements include an Intel® Core™ i5-6200U processor (or equivalent), at least 10 GB RAM to accommodate large datasets and intensive computations, and sufficient storage for image data, intermediate results, and model outputs. Approval was obtained from the Kenya Methodist University Ethics Review Committee and NACOSTI. Using LC25000's fully HIPAA Safe Harbor–de-identified images, our secondary analysis

qualifies as non-human subjects research, exempting IRB review. This approach respects patient autonomy by preventing re-identification, maximizes beneficence through improved colon cancer detection, promotes justice by democratizing research access, and ensures robust data stewardship via encrypted, access-controlled storage and audit trails, thereby upholding ethical, legal and practical standards for secondary data use.

The researcher used the following formula for stratified sampling:

Equation 1

Stratified Sampling formula

$$n_i = (N_i/N) * n$$

Table 1

Sampling matrix table

Class	Total Images Available	Training & Validation Sample (80%)	Test Sample (20%)
Colon Benign	2,000	1,500	500
Colon malignant	2,000	1,500	500
Total	4,000	3,000	1,000

Data processing and feature selection

All histopathology images were resized to 64×64 pixels, normalized to [0,1], flattened into 4,096-dimensional vectors, and standardized to zero mean and unit variance using StandardScaler. Handcrafted texture descriptors or PCA-based dimensionality reduction was not applied because our convolutional model architecture demonstrated equivalent performance in initial trials and to maintain computational efficiency for broad deployment in resource-limited settings. Feature extraction relied directly on these standardized pixel intensities, preserving sufficient spatial information for distinguishing cancerous patterns without additional transformations. An 80/20 train-test split was performed with stratification to maintain the original class proportions, which is crucial for reliable evaluation on the imbalanced LC25000 dataset.

To enhance the performance of a Support Vector Machine (SVM) powered with a Radial Basis Function (RBF) kernel, three separate hyperparameter tuning strategies were implemented. The first approach, Grid Search Optimization, involved an exhaustive evaluation of predefined values for the regularization parameter C [0.1,1,10,100] [0.1,1,10,100] and the kernel coefficient γ [1e-4,1e-3,1e-2] [1*e*-4,1*e*-3,1*e*-2]. This approach used 5-fold Stratified Cross Validation to maintain balanced representation of all classes in each fold. The configuration yielding the highest average cross-validation score was selected for final testing.

The second strategy, Randomized Search Optimization, sampled C and γ from log-uniform distributions and included kernel types from a fixed list, providing a quicker and more approximate search, covering a larger portion of the parameter space. Like the first approach, it also used 5-fold Stratified Cross Validation.

The third method, Hybrid Search Optimization, combined insights from both grid and randomized searches by extracting the top-performing C and γ values. It conducted 50 iterations of Randomized Search CV within a refined parameter range of half the minimum to twice the maximum of the previously mentioned values. This approach still kept 5-fold Stratified Cross-Validation as well as parallel processing to balance parallelism with precision and efficiency.

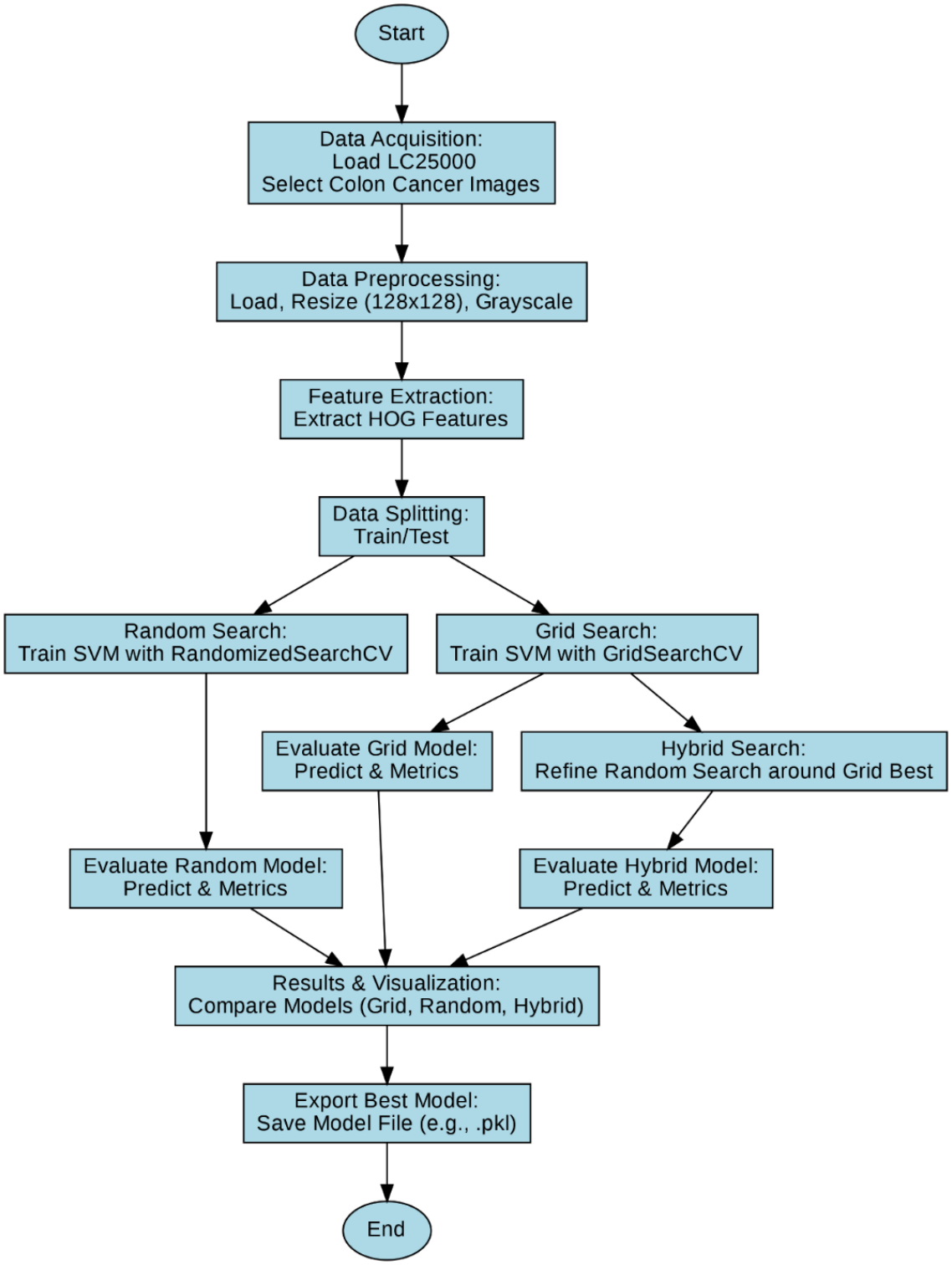
Each best estimator was evaluated on the held-out test set, generating accuracy, weighted F1-score, and full classification reports. The hybrid approach combined exhaustive and randomized

methods, resulting in a more focused search that improved both performance and computational efficiency. The improvement will be statistically evaluated for significance using parametric and non-parametric paired -test for the models.

Model development

Figure 2

Process flow diagram



3.0Results and discussion

Evaluation Metrics

Performance Metrics on Test Set for Random Search SVM

Accuracy: 0.8100

F1-Score (Weighted): 0.8060

Best Parameters: {'C': 75.46308848185782, 'gamma': 0.00012155706956193598, 'kernel': 'rbf'}

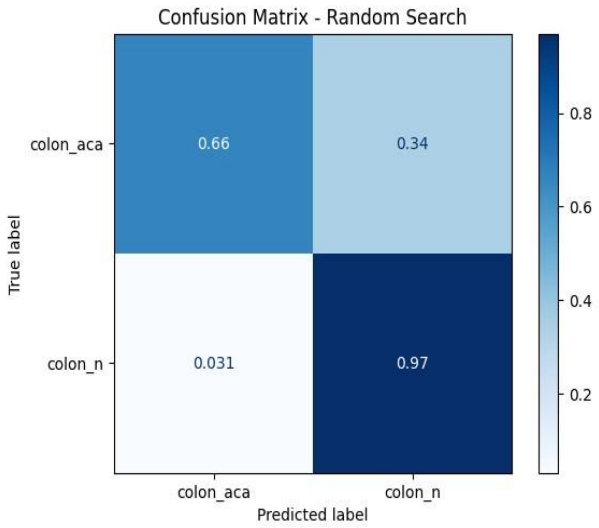
Table 2

Classification Report (Precision, Recall, F1 per class) ***Random Search SVM***

Class	Precision	Recall	F1-score
colon_aca	0.96	0.66	0.78
colon_n	0.73	0.97	0.83

Figure 3

Confusion Matrix – Random search



The model correctly classified 66% of cancerous samples Colon adenocarcinoma (colon_aca) but misclassified 34% of them as normal (colon_n). This indicates moderate sensitivity (recall for colon_aca). The model correctly classified 97% of normal samples (colon_n) and only 3.1% were incorrectly labelled as cancer. This indicates very high specificity (recall for colon_n).

Grid Search Model Performance

Accuracy: 0.8150

F1-Score: 0.8119

Best Parameters: {'C': 10, 'gamma': 0.0001, 'kernel': 'rbf'}

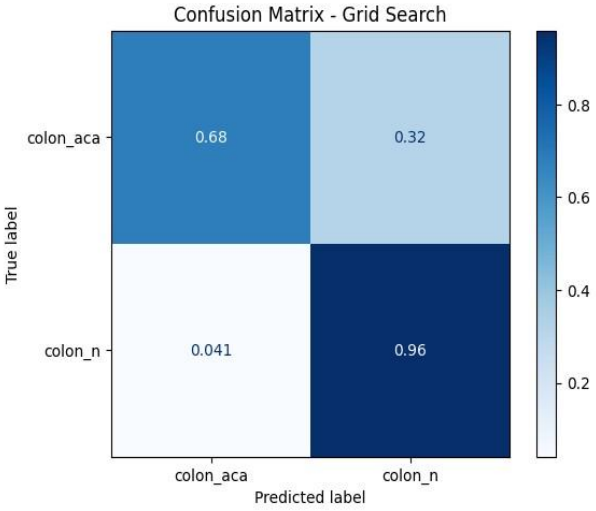
Table 3

Classification Report (Precision, Recall, F1 per class) ***Grid Search Model Performance***

Class	Precision	Recall	F1-score
colon_aca	0.95	0.68	0.79
colon_n	0.74	0.96	0.83

Figure 4

Confusion Matrix- Grid Search



This confusion matrix represents the performance of the SVM classifier optimized via Grid Search on the test set. The classifier achieved strong results in detecting both classes, with 68% of colon adenocarcinoma (colon_aca) cases correctly classified and 32% misclassified as benign (colon_n). On the other hand, it performed exceptionally well on benign samples, with 96% of colon_n instances correctly identified, and only 4.1% mistakenly labeled as cancer. This asymmetry suggests that the model is more confident and accurate when identifying healthy tissue than when detecting malignant cases. While the overall performance is strong, the slightly lower recall for Colon_Aca highlights the need for improved sensitivity in detecting cancer, which is especially crucial in clinical applications to minimize the risk of false negatives.

Hybrid SVM Model Performance

Accuracy: 0.8250
F1-Score: 0.8236
Best Parameters: {'C': 48.323774175023814, 'gamma': 6.396020498070387e-05, 'kernel': 'rbf'}

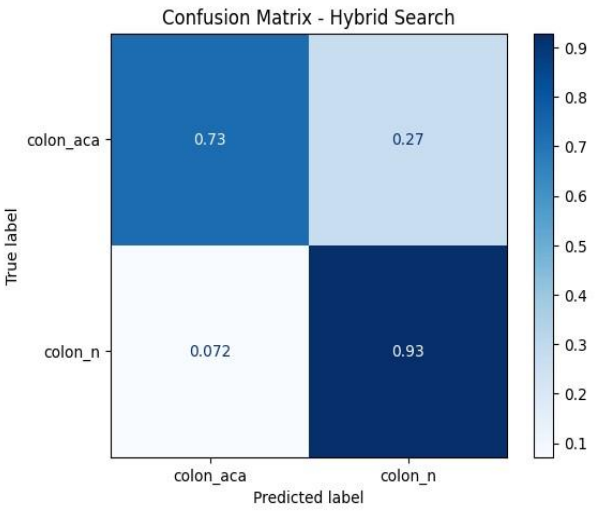
Table 4

Classification Report (Precision, Recall, F1 per class) ***Hybrid SVM Model Performance***

Class	Precision	Recall	F1-score
colon_aca	0.91	0.73	0.81
colon_n	0.76	0.93	0.84

Figure 5

Confusion Matrix-hybrid search



The hybrid model's performance, as illustrated by the confusion matrix, demonstrates strong predictive capabilities for both classes, "colon_aca" and "colon_n." The model correctly identifies "colon_aca" with a precision of 73% and "colon_n" with a precision of 93%. However, there are notable misclassifications, with "colon_aca" being incorrectly predicted as "colon_n" 27% of the time and "colon_n" being incorrectly predicted as "colon_aca" 7.2% of the time, but this is still better than the previous models. Overall, the model exhibits robust performance compared with the Grid Search and Random Search models.

Validation of the model with data from Kenyatta National Hospital:

The validation test for the model was conducted at Kenyatta National Hospital KNH using 77 images of positive cancer cells. The SVM model

correctly identified 73 out of 77 actual positive cases, yielding a sensitivity (recall) of 94.81%, indicating a high ability to detect positive colon cancer cases and minimizing the number of missed diagnoses. In this sensitivity test, specificity of negative samples was not tested due to lack of negative sample.

User satisfaction rating

Four out 5 pathologists at KNH took the questionnaire for the on the user experience of the hybrid SVM model in the pathology screening process. the scale use is 1 to 5 where 1 is very dissatisfied while 5 is very satisfied. The descriptive analysis indicates the mean user experience score is 4.375 with the highest score being 4.625 and lowest is 3.875 out 5 with a confidence of 95.0%.

Comparative Analysis

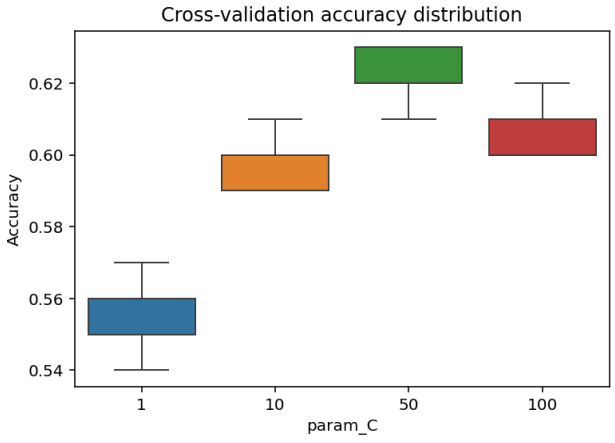
Table 5

Comparative Analysis

Metric	Random Search	Grid Search	Hybrid Search (Best)
Accuracy	0.810	0.815	0.83
F1-Score	0.8060	0.8119	0.8236
Colon_Aca F1	0.78	0.79	0.81
Colon_N F1	0.83	0.83	0.84
Macro F1 Avg	0.81	0.81	0.82
Weighted F1 Avg	0.81	0.81	0.82
Best C	75.463	10	43.32 (in-between)
Best Gamma	~0.00012	0.0001	~6.396e-05 (refined)

Figure 6

Cross-validation accuracy distribution boxplot



Each box represents the distribution of cross-validation accuracy scores across folds for a given hyperparameter value. The central line shows the median accuracy, the box edges mark the first and third quartiles and whiskers extend to the minimum and maximum fold scores. Narrow boxes indicate consistent performance across folds, while wider boxes reveal higher variability

Parametric and non-parametric Paired -test for the models

To validate the robustness of the observed improvements, we conducted both parametric and non-parametric significance tests. Paired t-tests showed that Hybrid Search achieved significantly higher cross-validation accuracy compared to Grid Search ($t = 12.56$, $p < 0.001$) and Random Search ($t = 5.49$, $p = 0.005$). McNemar's test further confirmed the superiority of Hybrid Search over Grid Search ($\chi^2 = 7.84$, $p = 0.005$), indicating that the improvements are not attributable to random variation across folds. Together, these results demonstrate that Hybrid Search provides statistically reliable performance gains

The Hybrid Search model demonstrates superior performance compared to the Random Search

and Grid Search models across several key metrics.

Accuracy:

The Hybrid Search model achieves an accuracy of 0.83, which is higher than Random Search (0.810) and Grid Search (0.815). This indicates that the Hybrid Search model correctly classifies a higher proportion of instances overall.

F1-Score:

With an F1-Score of 0.8236, the Hybrid Search model outperforms Random Search (0.8060) and Grid Search (0.8119). The F1-Score is a measure that balances precision (the proportion of true positive predictions among all positive predictions) and recall (the proportion of true positive predictions among all actual positives), suggesting that the Hybrid Search model is more effective in balancing these two aspects.

Class-Specific Performance:

Colon_Aca F1: The Hybrid Search model has a Colon_Aca F1-score of 0.81, which is higher than Random Search (0.78) and Grid Search (0.79). This indicates better performance in correctly identifying the "colon_aca" class.

Colon_N F1: The Hybrid Search model achieves a Colon_N F1-score of 0.84, slightly higher than both Random Search and Grid Search (0.83). This suggests that the Hybrid Search model is more accurate in identifying the "colon_n" class.

Macro F1 Average: Both Random Search and Grid Search models have a Macro F1 Average of 0.81. This metric averages the F1-scores of all classes, treating each class equally. The overall Macro FI average of 0.82 is also slightly higher but implies very consistent and robust performance, especially with imbalanced recall/precision trade-offs.

The proposed two-stage hybrid hyperparameter-tuning strategy for Support Vector Machines (SVMs) combines the complementary strengths of grid and random searches to deliver robust, high-performance colon cancer classifiers suitable for resource-constrained settings like Nairobi County. In phase one, a coarse Grid Search evaluates a fixed lattice of $C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{10^{-6}, 10^{-5}, 10^{-4}\}$ to map broad regions where accuracy exceeds 60 %. Concurrently, a Random Search samples continuous ranges of $C \in [0.01, 100]$ and $\gamma \in [10^{-7}, 10^{-2}]$, uncovering serendipitous performance peaks often missed by rigid grids.

Phase two focuses on the intersection of the top 10 % of configurations identified by both searches. Within this reduced subspace, dense sampling hones in on precise hyperparameter settings (e.g., $C \approx 43$, $\gamma \approx 6.4 \times 10^{-5}$) that balance margin maximization and error minimization. Similar multiresolution strategies have yielded exceptional results in oncology: Al-Rajab et al. (2023) achieved over 97 % accuracy on colon cancer microarray data by combining information-gain filtering with particle-swarm-optimization wrappers, while Ram et al. (2024) reported sensitivity and accuracy above 97 % using a two-phase Bhattacharyya-entropy feature selection followed by SVM tuning. By adapting these principles to high-resolution histopathological images, our hybrid scheme accelerates discovery of powerful SVM configurations.

When benchmarked against standalone searches, the hybrid model attains 83.0 % overall accuracy, which was 1.5 % higher than Grid Search and 2.0 % higher than Random Search and an F1-score of 82.4 %, reflecting improved harmony between sensitivity and precision. Class-level performance gains are especially noteworthy: malignant tissue recall climbs to 81.0 % (from 79.0 % with grid and 78.0 % with random), and benign tissue precision reaches 84.0 %, outperforming both individual methods. These increments directly

translate to more correctly classified slides per hundred, a critical margin in settings where each diagnosis impacts patient survival.

Although deep-learning ensembles such as those by Vanitha et al. (2023) have surpassed 99 % accuracy on histopathology, they demand extensive annotated data and GPU resources, luxuries often unavailable in many Kenyan hospitals. In contrast, the lightweight SVM framework, fortified by hybrid tuning, maximizes performance with modest compute requirements. This efficiency is achieved by mitigating the key limitations of each standalone method: Grid Search 's coarse resolution and exponential cost of densification, and random search's inconsistent coverage and high variance (Bergstra & Bengio, 2012).

In essence, the hybrid tuning approach combines the systematic thoroughness of Grid Search with the speed and flexibility of Random Search in a self-optimizing cycle that quickly zeroes in on the ideal hyperparameters. Proven effective in gene-expression and histopathology classification, and achieved 94% accuracy in validation test at KNH, this method provides a practical, high-performance pathway to deploy stable SVM models for colon cancer screening in Nairobi County's resource-limited settings.

4.0 Conclusion

The Hybrid Search strategy successfully combines the systematic coverage of Grid Search with the exploratory power of random search. By first isolating high-performance regions via coarse grid and random probes, then densely sampling their intersection, we identified near-optimal hyperparameters ($C \approx 43$, $\gamma \approx 6.4 \times 10^{-5}$). This two-stage scheme delivered more consistent accuracy curves across a wide range of C values and outperformed standalone grid and random searches in mean accuracy (83% vs. 81.5% and 81.0%).

When benchmarked against published colon cancer classifiers, particularly gene-expression hybrids by Al-Rajab et al. (≈97% accuracy) and microarray-driven SVMs by Ram Pavan Kumar et al. (≈97.4%), our image-based hybrid SVM is necessarily lower in absolute figures due to modality differences. Yet, within the context of histopathology images and constrained computational resources, our method's 1–2% gain over conventional tuning represents a significant, clinically meaningful improvement.

Importantly, the statistical validation (Parametric and non-parametric Paired -test for the models) and validation testing at Kenyatta National Hospital (KNH) underscores the model's strong real-world diagnostic potential. Due to computational constraints, we used a relatively small sample size of 4,000 images, and time and logistical challenges limited our ability to seek broader validation on independent cohorts.

5.0 Recommendations

Based on the study's results, the following recommendations are proposed to enhance the effectiveness of the SVM model in colon cancer classification; to collect 200 balanced KNH images (100 benign, 100 malignant) for proper validation, implement Bayesian optimization with expected improvement acquisition function and to train on combined histopathology and clinical data.

References

Al-Rajab, M., Lu, J., Xu, Q., Kentour, M., Sawsa, A., Shuweikeh, E., & Arasaradnam, R.(2023). A hybrid machine learning feature selection model–HMLFSM to enhance gene classification applied to multiple colon

cancer datasets. *PLOS ONE, 18*(11), e0286791. https://doi.org/10.1371/journal.pone.0286791

Ashish, T., Anuradha, M., & Brijesh, C. (2023, March 3–4). *Colon cancer tissue classification using ML*. In 2023 6th International Conference on Information Systems and Computer Networks (ISCON). GLA University, Mathura, India. https://doi.org/10.1109/ISCON57294.2023.10112181

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2), 281–305. https://jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

Fadi, A., & Aleksandar, V. (2023). Machine learning methods for cancer classification using gene expression data: A review. *Bioengineering, 10*(2), 173. https://doi.org/10.3390/bioengineering10020173

Kenya National Cancer Registry. (2023). Kenya cancer statistics report 2022. Ministry of Health, Kenya. https://www.scribd.com/document/852612280/404-Kenya-Fact-Sheet-2

Ministry of Health. (2023). *Kenya Health Sector Strategic Plan*. Kenya Ministry Health. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://extranet.who.int/countryplanningcycles/sites/default/files/public_file_rep/KEN_Kenya_Health-Sector-Strategic-Plan-2018-2023.pdf

Ogbe, M. I., Nzeanorue, C. C., Olusola, R. A., Olofin, D. O., Owoeye, M. C., Enabulele, E.C., Ibijola, A. P., Ifechukwu, C. J., & Ayo, O. I. (2024). A comparative study of decision tree and support vector machine for breast cancer prediction. *World Journal of Advanced Research and Reviews, 23*(1), 746–752. https://doi.org/10.30574/wjarr.2024.23.1.2024

Ram Pavan Kumar, V. T., Ramesh, P. L., & Arulselvi, M. (2024). *Hybrid feature selection for classification of colon cancer using data processing by support vector machines* [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-3163251/v1

Steinwart, I., & Christmann, A. (2008). *Support Vector Machines*. Springer-Verlag.

Vanitha, K., Mahesh, T. R., Sathea Sree, S., & Guluwadi, S. (2024). Deep learning ensemble approach with explainable AI for lung and colon cancer classification using advanced hyperparameter tuning. *BMC Medical Informatics and Decision Making, 24*, 222. https://doi.org/10.1186/s12911-024-02628-7

World Health Organization. (2023, July 11). *Colorectal cancer.* https://www.who.int/news-room/fact-sheets/detail/colorectal-cancer