Cervical Cancer Detection Using Handcrafted Features and Machine Learning

Abdelrahman Sayed Nasr, Salah Mohamed Salah, Ali Younis, Yasmine Mahmoud
Team 4

***Abstract*—Cervical cancer is among the most common types of cancer in women worldwide. Early and accurate detection through Pap smear screening can significantly improve treatment outcomes. In this project, we aim to classify cervical cell images into five distinct categories using handcrafted features extracted from the images. The features used include Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) for capturing texture information, along with mean intensity and circularity for both the nucleus and cytoplasm to capture morphological characteristics. These features were extracted using the image masks provided for each region. After preprocessing and feature extraction, a variety of classifiers were applied to evaluate performance. The results highlight the effectiveness of combining texture and shape features for cervical cell classification.**

***Index Terms*—Cervical cell classification, feature extraction, HOG; SIFT, mean intensity, circularity, image processing, machine learning.**

I. Introduction

Medical images play a very important role in helping doctors detect and diagnose diseases at an early stage. With the rise of technology, machine learning has become a powerful tool for analyzing these images and supporting the decision-making process by making it faster, more consistent, and often more accurate. Many researchers have explored different machine learning models to assist with medical diagnosis, including both advanced deep learning techniques and simpler, more traditional models like Support Vector Machines (SVM). These tools are especially helpful in identifying patterns that may be difficult for the human eye to catch. In this project, we are focusing on a straightforward and practical method. We take cropped images of cervical cells, extract important features from them, and then use a machine learning model to classify the images to determine whether they show signs of cancerous or precancerous growths in the cervix. This method is simpler and more efficient than deep learning models, which usually require a lot of data and computational power. Our approach is suitable for cases where resources are limited but reliable results are still needed. To better understand how effective our approach might be, and to compare it with other possible methods, we reviewed a few recent research papers. Each paper used different techniques to solve similar classification problems, mostly related to cervical cancer or other types of medical image analysis. These studies helped guide our project and gave us useful ideas for feature extraction, preprocessing, and model selection.

II. Related Work

The integration of machine learning into medical diagnostics has opened new possibilities for early disease detection and decision support systems. In particular, the use of classical algorithms such as Support Vector Machines (SVM) alongside more modern deep learning techniques has been explored in a variety of studies. We examined three research articles published between 2021 and 2023, each contributing to our understanding of how machine learning models can assist in disease classification. These studies are relevant to our current project, which involves using cropped medical images, extracting meaningful features, and applying classification models.

Akinwunmi et al. (2021) *Machine Learning-Based Cervical Cancer Detection Using Clinical Data*

Akinwunmi et al. (2021) investigated cervical cancer detection by employing classical machine learning classifiers on a clinical dataset from the UCI repository. The dataset contained patient risk factors and clinical attributes rather than image data. Their approach utilized several algorithms including Support Vector Machines (SVM), Decision Trees, and Random Forests, with emphasis on preprocessing steps such as normalization and feature selection to improve model accuracy. Despite not using imaging data, this work underscores the effectiveness of classical classifiers for medical diagnosis tasks and highlights the importance of thorough data preparation, which is directly applicable when using extracted image features in our project.[1]

Khan et al. (2022) *Deep Learning-Based Pneumonia Classification from Chest X-Ray Images*

Khan et al. (2022) proposed a convolutional neural network (CNN) model for pneumonia classification using chest X-ray images. Their method involved image preprocessing, including normalization and resizing, and achieved high accuracy and sensitivity. This study illustrates the power of deep learning in medical imaging, emphasizing domain-specific architecture design and model tuning to address challenges such as class imbalance. Although our work focuses on feature extraction with classical classifiers rather than deep networks, their preprocessing pipeline and classification insights inform our image processing and evaluation strategies.[2]

Shah et al. (2023) *Addressing Class Imbalance in Cervical Cancer Prediction Using Machine Learning Techniques*

Shah et al. (2023) examined cervical cancer prediction with an emphasis on handling class imbalance in clinical datasets.

They evaluated multiple resampling methods like SMOTE and undersampling combined with classifiers including Logistic Regression, SVM, and XGBoost. Their findings stress that class imbalance, common in medical datasets, can substantially affect model performance. These insights are valuable for our project, where imbalanced datasets may arise between normal and abnormal image samples, making class balancing methods a potentially useful preprocessing step at the feature level.[3]

Insights

By combining elements from all three (such as using cropped images, extracting features manually or through simpler methods, and applying classical models), we aim to build a model that is computationally efficient yet effective for disease classification tasks. The reviewed literature confirms the validity of this approach and informs our methodology, particularly in areas of preprocessing, feature extraction, feature selection, and evaluation.

III. Methodology

A. Dataset and Preprocessing

We utilized the SIPaKMeD dataset, which contains cervical cell images categorized into five classes. Our preprocessing pipeline consisted of:

- Data organization into dictionary structure
- Contour extraction from .dat files
- Image standardization to 48×62 pixels with padding
- Binary mask generation

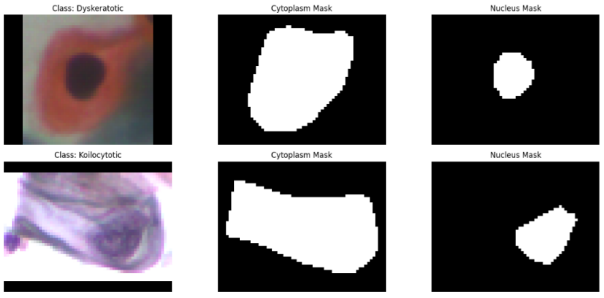

Fig. 1: Preprocessing pipeline showing original images, masks, and processed results

B. Feature Extraction

We extracted three categories of features:

1) Color Intensity Features:

- Cytoplasm mean R, G, B values
- Nucleus mean R, G, B values

2) Morphological Features: Circularity calculated as:

$$Circularity = \frac{Perimeter^2}{4\pi \times Area} \quad (1)$$

3) Texture Features:

- HOG from 48×48 grayscale images
- SIFT descriptors averaged to fixed-length vectors

C. Feature Selection

We applied:

- ANOVA F-test (SelectKBest)
- Mutual Information

IV. Results and Discussion

A. Model Training and Evaluation

We evaluated three classifiers with hyperparameter tuning using GridSearchCV:

TABLE I: Hyperparameter Search Spaces

Model	Parameters
Decision Tree	max_depth [2,5,10,15,20], min_samples_split [2,5,10,20,50]
Random Forest	n_estimators [50,100,200], max_depth [None,10,20]
SVC	kernel ['poly','rbf'], C [0.001,0.01,0.1,0.5,1,10]

Our experimental results demonstrated the effectiveness of the proposed approach:

TABLE II: Model Performance Comparison

Model	CV Accuracy	Test Accuracy
Decision Tree	0.797	0.796
Random Forest	0.853	0.845
SVC	0.856	0.856

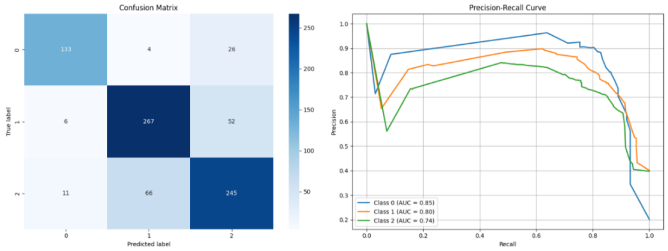

Fig. 2: DT Confusion matrix and Precision-Recall curve

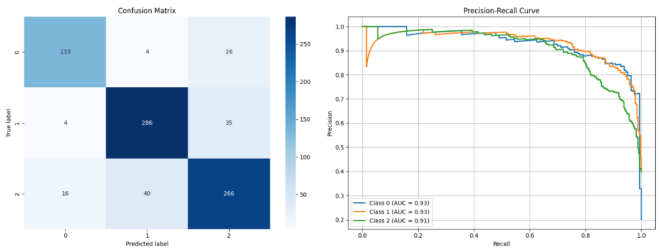

Fig. 3: RF Confusion matrix and Precision-Recall curve

The SVC achieved the highest performance with 85.6% accuracy. The success of our approach can be attributed to careful preprocessing preserving morphological features and comprehensive feature extraction.

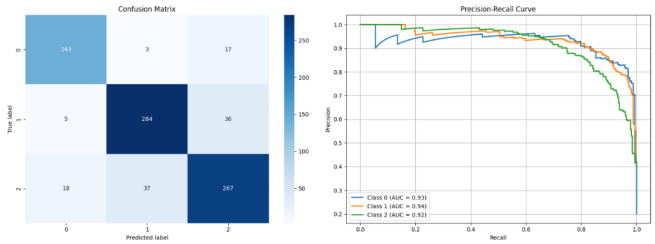

Fig. 4: SVC Confusion matrix and Precision-Recall curve

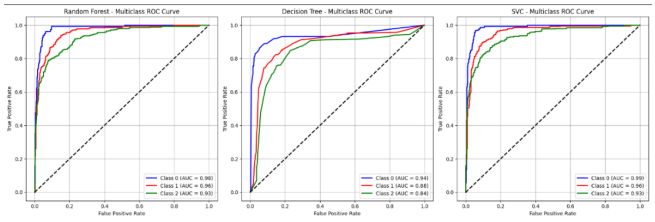

Fig. 5: ROC curves with AUC values

V. Conclusion and Future Work

In this project, we developed an effective approach for cervical cell classification using handcrafted features such as HOG, SIFT, mean intensities, and circularity measurements for both nucleus and cytoplasm. Our evaluation of classical machine learning models, like the Decision Tree, Random Forest, and Support Vector Classifier, showed that the Support Vector Classifier achieved the highest accuracy (85.6%) and multiclass ROC AUC, demonstrating that carefully selected features combined with optimized classical models can provide reliable classification results. The morphological and texture-based features successfully captured critical variations in cell structure and appearance, which are key to distinguishing normal from abnormal cells. Preprocessing steps like contour extraction and image padding preserved important cellular details, enhancing the quality of feature extraction and contributing to the overall performance of the models. This highlights that classical machine learning pipelines remain a viable and practical solution in medical image analysis. For future work, exploring additional feature extraction methods such as *Local Binary Patterns (LBP)* or *Gray-Level Co-occurrence Matrix (GLCM)*, which we mentioned in phase I, could further improve the representation of cellular textures and boost classification accuracy. Additionally, investigating ensemble or hybrid models that combine deep learning feature extraction with classical classifiers may offer a good trade-off between computational efficiency and predictive performance. Overall, this study demonstrates the potential of feature-based classical machine learning approaches as accessible alternatives to deep learning, particularly in resource-limited settings. Continued refinement of these methods can contribute to improved early detection and diagnosis of cervical cancer, supporting better clinical outcomes.

Contributions

-Salah Mohamed: Dataset Preprocessing
-Yasmine Mahmoud: Features Extraction
-Ali Younis: Features Selection
-Abdelrahman Sayed: Model training and Evaluation

Acknowledgments

We thank the maintainers of the SIPaKMeD dataset for making this valuable resource available to the research community.

References

[1] A. Akinwunmi, J. O. Folorunso, and T. M. Oyedotun, ”Machine Learning-Based Cervical Cancer Detection Using Clinical Data,” *Frontiers in Public Health*, vol. 9, p. 788376, Sep. 2021. doi:10.3389/fpubh.2021.788376. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpubh.2021.788376/full

[2] S. Khan, A. Iqal, and M. Khan, ”Deep Learning-Based Pneumonia Classification from Chest X-Ray Images,” *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 4688327, 2022. doi: 10.1155/2022/4688327. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1155/2022/4688327

[3] S. Shah, R. Khan, and M. Ali, ”Addressing Class Imbalance in Cervical Cancer Prediction Using Machine Learning Techniques,” *Applied Sciences*, vol. 13, no. 9, p. 5761, Apr. 2023. doi: 10.3390/app13095761. [Online]. Available: https://www.mdpi.com/2076-3417/13/9/5761