Classifying Digital Colposcopy Image Quality Using Predictive Models with Explainable AI

Isabel B. Baclig
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
ibbaclig@up.edu.ph

Nicole Anne I. Balde
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
nibalde@up.edu.ph

David Raphael B. Bobis
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
dbbobis@up.edu.ph

James DC. Sablay
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
jdsablay@up.edu.ph

Ma Sheila A. Magboo
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
mamagboo@up.edu.ph
ORCID 0000-0002-6221-7892

Vincent Peter C. Magboo
Department of Physical Sciences and Mathematics
University of the Philippines Manila
Manila, Philippines
vcmagboo@up.edu.ph
ORCID 0000-0001-8301-9775

***Abstract*— Cervical cancer remains a significant global health challenge, ranking as the fourth most common cancer among women, with particularly high mortality rates in low and middle-income countries. Early and accurate diagnosis is essential for improving outcomes, yet traditional colposcopy methods face limitations due to operator dependency and variability in image quality. Digital colposcopy, coupled with machine learning, offers a promising solution to address these challenges by enhancing diagnostic accuracy through standardized image quality assessment. This study utilizes predictive models and Explainable Artificial Intelligence (XAI) to classify digital colposcopy images as "good" or "bad" quality. Key preprocessing steps included Synthetic Minority Oversampling Technique to address class imbalance and SelectKBest for a feature selection procedure. Logistic Regression without SMOTE and without feature selection achieved the highest F1-score (0.89) on the Hinselmann dataset, while XGBoost with SMOTE and without feature selection excelled on the Green dataset (F1-score: 0.87). Random Forest without SMOTE but with feature selection attained the highest F1-score (0.91) on the Schiller dataset. To ensure interpretability, Local Interpretable Model-Agnostic Explanations (LIME) identified influential features such as RGB intensity values and artifact-related metrics. These findings highlight the potential of combining preprocessing, robust machine learning algorithms, and XAI techniques to improve the reliability and transparency of automated colposcopic evaluations.**

Keywords—*Digital Colposcopy, Explainable AI, Image Quality Classification*

I. Introduction

Cervical cancer is the fourth most common type of cancer in women and associated with continuous infections with high-risk sexually transmitted human papillomaviruses (HPV) [1]. The delays in HPV vaccination and cervical screening, shortages in cervical cancer treatment and other cultural factors has been noted as some of the primary reasons in the high incidence and mortality of this disease particularly in low and middle income countries (LMIC) [2]. Compared to other types of malignancies, cervical cancer could be treated successfully given an early diagnosis with subsequent prompt treatment and effective management [3].

While primary cervical screening usually involves methods such as Pap smears or HPV testing, colposcopy serves as an important diagnostic follow-up for abnormal findings with much higher diagnostic accuracy in assessing cervical lesions as compared to the pap smear [4]. In the conventional colposcopy method, visual inspection of the cervix is done through magnification using acetic acid [5]. It is usually recommended only after a positive screening test and serves as the standard procedure for biopsy and treatment guidance [6]. This method, however, faces multiple constraints especially in the LMIC setting as colposcopy is largely dependent on the subjective professional experience of the operators leading to significant inter- and intra-operator variabilities. This is further compounded by the lack of skilled colposcopists in LMICs which subsequently aggravate diagnostic inaccuracies and/or discrepant reporting [7].

With the continuous transition of the health industry towards digitalization, digital colposcopy has become a promising technology for diagnostic analysis as it can provide high-definition cervical images to be used by medical practitioners and offers a potential solution to address the issues in performance of colposcopy [7]. It must be noted, however, that the quality of the colposcopy images is highly crucial for accuracy as poor quality can impede accurate diagnosis and can adversely impact the verification and generalizability of the diagnostic outcomes [8]. Such variability in the image quality can be due to various factors such as blur, poor focus, poor light, noise, obscured view of the cervix due to mucus and/or blood, improper position, and over- and/or under-exposure [9]. The integration of machine learning in digital colposcopy, in this case, can provide potential solutions to the bottleneck problems – one of which being the cervical image quality assessment.

This research focuses on the development of predictive models and incorporating XAI techniques to assess the quality of digital colposcopic images. The paper aims to classify the images as "good" or "bad" quality for the enhancement of the reliability of colposcopic evaluations. This study also

analyzed the significant contribution of incorporating a feature selection procedure and class imbalance handling in their effects on machine learning (ML) algorithms for assessing quality of colposcopy images to upgrade cervical cancer prediction. These ML-based models coupled with explainability could then be potentially integrated in the clinical workflow of the health professionals as a non-invasive method to assess image quality in colposcopy, thus enabling reliable diagnostic outcome and its subsequent prompt therapeutic intervention, thereby alleviating the public health impact of cervical cancer.

The format of this research work is presented as follows: an introduction section is followed by literature review underscoring the recent directions as to colposcopy image enhancement as a means of detecting cervical cancer. The next section details the research methods employed including the particulars of the dataset, the needed pre-processing steps and the machine algorithms utilized in various model configurations in the study. The unveiling of the results and corresponding analysis are then detailed in the next section. Finally, the study ends with a synopsis of the research outcomes and recommendations for subsequent work.

II. Relevant Literature

Nikookar et al. developed an AI-based cervical cancer prediction model by merging features from three colposcopic imaging modalities—Greenlight, Hinselmann, and Schiller [10]. Using ensemble classifiers like Naïve Bayes, AdaBoost, Random Forest, and Support Vector Machine (SVM), along with aggregation strategies (e.g., majority voting, logical "AND") and classifier selection methods (e.g., Principal component analysis (PCA), Forward Search), the model achieved 96% sensitivity, 94% specificity, and a ROC-AUC of 0.94. Despite outperforming single-modality classifiers, the model's generalizability was limited by inconsistent image quality and reliance on predefined aggregation rules. In the study by Wu et al., researchers highlighted the use of an AI-based digital colposcopy diagnostic system that will provide assistance to junior trainees in accurately identifying the lesions where to conduct tissue biopsy [11]. Results showed the clinical utility of such AI-based system with its enhanced diagnostic accuracy and biopsy efficiency, showcasing it as an encouraging solution to upgrade the quality of cervical cancer screening in low-resource scenarios. The benefits of incorporating ML in cervical cancer diagnosis was also studied by Wang et al. [12]. Using variance threshold analysis, univariate feature selection (SelectKBest), and least absolute shrinkage and selection operator (LASSO), ML-based prediction model with a testing accuracy rate of 83.5% has led to a reduction in invasive examination prior to surgery, directing surgical intervention and adjuvant chemotherapy for cervical cancer. Likewise, Kim et al. elaborated the utility an AI-based system in cervical cancer screening. Authors believed that the AI interpretation can be harnessed as an auxiliary tool in combination with human colposcopic evaluation of exocervix [13].

Mehmood et al., emphasized preprocessing and feature selection in their ML-based system called CervDetect to predict cervical cancer using random forest feature selection technique to select significant features [14]. Results showed an impressive performance with an accuracy rate of 93.6%. Ileberi and Sun highlighted the challenging task of choosing the most relevant and informative features from a colposcopy dataset that shall have an impact on the predictive performance and interpretability of ML models [1]. Authors employed a particle swarm optimization technique as a feature selection procedure applied to a variety of ML models. Their findings showed Adaboost obtained the highest prediction capability with a 98% accuracy rate. In [15], authors employed Chi-square and Least Absolute Shrinkage and Selection Operator (LASSO) feature selection procedures together with Shapley Additive Explanations (SHAP) to create a diagnostic system for cervical cancer prediction. Results showed an impressive predictive capability of the decision tree with an 97.6% accuracy. SHAP enhanced the model transparency by recognizing key attributes such as Hormonal Contraceptives (years), Age, and Number of Pregnancies. Despite its robustness, the study primarily focused on cancer prediction, leaving image quality classification underexplored.

Ekem et al. highlighted the relevance of maximizing image quality in colposcopy for cervical cancer prediction [8]. Authors reviewed several deep learning algorithms to assess the colposcopic image quality and concluded that poor quality colposcopy images can adversely impact the generalizability of the deep learning models. Another study assessing colposcopic cervical images quality emphasized that quality classification is a pivotal task in guaranteeing reliable detection of cervical lesions [9]. Authors analyzed the image quality problem head on by developing a classifier to delineate colposcopic cervical images into into "low", "intermediate" and "high" quality categories. Finally, Jelen et al. studied optimizing features through advanced feature selection techniques to reduce feature dimensionality while preserving only relevant diagnostic information [16]. Results showed that optimized feature sets coupled with support vector machine classifier generated comparable classification capability with that of convolutional neural networks while prominently minimizing computational complexity.

Collectively, while existing research has made significant contributions to cervical cancer diagnostics through AI, ensemble classifiers, deep learning models, and feature selection techniques, the task of predicting colposcopic image quality remains underexplored. Most research focuses on diagnostic outcomes, often neglecting the importance of image quality as a foundation for accurate predictions. There were only limited research works that have focused on assessing quality of cervical images to enhance prediction of cervical cancer, most of which pertain to deep learning approaches. This study addresses this gap by focusing explicitly on predicting image quality applied to a publicly available tabular digital colposcopy dataset.

III. Materials and Methods

There are several phases in the framework for this research endeavor. First, the dataset used was extracted from a publicly available data depository followed by preprocessing methods (data cleaning). The cleaned dataset was split into a training set and a testing set. A feature-selection procedure and class imbalance handling were also incorporated in the ML models. Lastly, the models were appraised based on their classification performance. An XAI approach was then applied in the topmost models. The pipeline for this research endeavor is detailed in Fig. 1.

A. Dataset

The Quality Assessment of Digital Colposcopies dataset, sourced from Fernandes et al. explores the subjective quality assessment of digital colposcopies and includes 287 instances

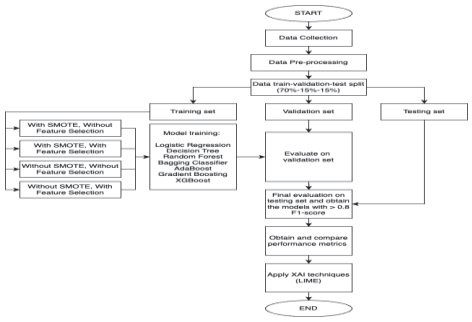

Fig. 1. Pipeline for Digital Colposcopy Image Quality Classification

with 62 predictive attributes and 7 target variables [17]. Acquired at Hospital Universitario de Caracas and annotated by professional physicians, the dataset provides three .csv files per imaging modality: Hinselmann, Green, and Schiller. The subjective quality judgments were initially ordinal (poor, fair, good, excellent) and were later discretized into two classes (bad, good). The consensus variable was used as the target variable, and individual expert assessments were excluded.

The dataset features a variety of attributes, including image area segmentation (e.g., cervix, speculum, artifacts), color properties in RGB and HSV spaces, and geometric features (e.g., convex hull, bounding box, ellipse fitting). Additionally, the dataset includes the original images and their manual segmentations. It supports classification tasks in the domain of health and medicine .

B. Pre-processing Steps

The preprocessing of the datasets was conducted with consensus designated as the target variable. A detailed inspection of the datasets revealed no missing values nor duplicate rows. All remaining features were numerical, so no categorical encoding was required. The target variable consensus, represented binary outcomes, with 1.0 indicative of a positive consensus while 0.0 represented a negative consensus.

Initial examination of the datasets revealed notable class imbalances, particularly in the Hinselmann dataset, where the positive class prominently outnumbered the negative class (82 positive cases and 15 negative cases). Correlation analysis was conducted to identify relationships between features and the target variable, highlighting potential predictors for model development. Identified outliers were likewise deleted from the analysis. To prepare the data for modeling, normalization was applied using MinMax scaling, which standardized all feature values to a range between 0 and 1. This step ensured consistency in the scale of features, facilitating better model convergence and improving overall performance. All of these preprocessing steps provided a robust foundation for machine learning tasks.

C. Feature Selection and Class Imbalance Handling

To improve model performance, SelectKBest with k=30 and f_classif test was used to select the most relevant features. The f_classif method evaluates the relationship between each continuous feature and the target variable using an ANOVA F-test. This test is suitable for continuous features as it assesses the variance between groups (positive and negative consensus) and selects features with the highest discrimination power. By retaining features with significant variance and discarding irrelevant ones, this approach ensured a more efficient and focused model. To address the significant imbalance in the all datasets, SMOTE (Synthetic Minority Over-sampling Technique) was applied. SMOTE generates synthetic samples for the minority class, balancing the dataset and preventing model bias toward the majority class.

D. Model Configuration and Machine Learning Models

The study explored the impact of SMOTE (Synthetic Minority Over-sampling Technique) and feature selection on the performance of various ML models on the three colposcopy datasets. Four configurations were evaluated: 1) Without SMOTE, Without Feature Selection, serving as a baseline; 2) Without SMOTE, With Feature Selection; 3) With SMOTE, Without Feature Selection; and 4) With SMOTE, With Feature Selection. Models include Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Bagging Classifier (BC), AdaBoost (Ada), Gradient Boosting (GB), and XGBoost (XGB). These configurations sum up to a total of 84 models in total (24 models for each of the three datasets) which allowed for a comprehensive analysis of how SMOTE and feature selection influence model performance in handling class imbalance and dimensionality. Lastly, performance indices consisting of F1-score, recall and precision were determined to evaluate ML classifier's detection capability with F1-score serving as the basis for determining the best performing model.

E. Explainable AI

This study used Explainable AI (XAI) which is essential for understanding and trusting machine learning models, especially in healthcare applications. Local Interpretable Model-agnostic Explanations (LIME) was employed to clarify the model's predictions. LIME provides interpretable explanations by approximating the behavior of complex models with simpler, locally accurate surrogate models around specific predictions. By perturbing the input data and analyzing the corresponding model outputs, LIME identifies the features most influential in the prediction process. This enables healthcare professionals to determine whether the model's decision-making aligns with domain knowledge and clinical guidelines. Through LIME, the study ensures that the models not only achieve high performance but also delivers transparent and actionable insights, fostering trust and improving usability in medical contexts.

IV. Results and Discussion

There were four model configurations evaluated in this study. With seven ML algorithms in each of the three colposcopy datasets, a total of 84 models were analyzed. Table 1 showcases the performance of the various models on the Hinselmann dataset. In the baseline configuration (without SMOTE and feature selection), LR appeared to have the best performance with the highest indices (F1-score = 0.89, recall = 1.00 and precision = 0.80). RF, BC and XGB have comparable performance with F1-score of 0.85) while DT, Ada and GB have lower predictive capability. To assess the impact of incorporating a feature selection step as seen in the second configuration (without SMOTE but with feature selection), performance indices were compared to the baseline configuration. There was an improvement in the performance with an increase of 5 and 4 percentage points in the F1-score of GB (0.80) and BC (0.89), respectively. The rest of the models did not elicit any improvement in the predictive capability. Likewise, to evaluate the impact of utilizing SMOTE to address imbalance as shown in the third

configuration (with SMOTE, without feature selection), the indices were compared to the baseline configuration. Generally, all models did not show any improvement in the performance indices. The incorporation of both SMOTE and a feature selection step as seen in the fourth configuration, marked improvement in the performance was evident for DT and GB having an increase of 10 percentage points in their F1-scores at 0.85. The other models did not show any significant improvement in the performance. Overall for the Hinselmann dataset, the top performing models were LR under the baseline configuration (no SMOTE and without feature selection) and LR and BC under the second configuration of adding purely a feature selection with no SMOTE.

Table II highlights the performance indices of various ML models for the Green dataset. In the baseline configuration, both LR and XGB obtained the highest indices (F1-score = 0.80, recall = 1.00 and precision = 0.67). As compared to the second configuration evaluating the impact of solely adding a feature selection step, significant improvement with a 9 and 7 percentage point increase in the F1-scores was seen for BC (F1-score = 0.82) and DT (F1-scores = 0.80), respectively. On the other hand, incorporating SMOTE only as shown in the third model configuration elicited a significant improvement only in XGB showing a 7 percentage point increase in F1-score at 0.87. The rest of the models did not show any increase in the predictive capability. With the incorporation of both SMOTE and a feature selection, only DT showed an impressive improvement with an increase of 13 percentage point with an F1-score (0.86). Overall for the Green dataset, XGB obtained the highest predictive capability with F1-score (0.87), recall (1.00) and precision (0.77) with the addition of only SMOTE without feature selection.

Table III shows the performance indices of the various ML models for the Schiller dataset. The baseline configuration showed RF with the highest indices at F1-score (0.87), recall (1.00) and precision (0.77). As compared to the baseline, the second configuration with an addition singly of a feature selection step, several models (DT, RF, GB and XGB) showed improvement in the predictive capability with an increase in their F1-scores from 4 to 19 percentage points. Likewise, an increase of 8 to 13 percentage point in the F1-scores were generated only by DT, BC and GB in the third configuration where SMOTE only was incorporated. Lastly, when both SMOTE and a feature selection were added as seen in the fourth configuration, BC, Ada, GB and XGB obtained a percentage point increase from 4 to 17 in the F1- scores. The best performing model for the Schiller dataset was obtained by RF with an impressive F1-score (0.91), recall(1.00) and precision (0.83) under the second configuration of having incorporated only a feature selection, without SMOTE.

The underlying principle in harnessing a feature selection procedure in a machine learning study is to choose only relevant data and discarding the unnecessary noise in the dataset. In [18], authors have reported that datasets with huge number of attributes can give rise the following: (1) overfitting, where the ML model becomes very complicated and thus, unable to render generalizations, (2) soaring computational complexity, (3) enlarging risk of bias as the dataset may contain inappropriate features, (4) shrunken interpretability due to a variety of attributes and (5) data redundancy because of highly correlated features with each other [18]. Feature selection steps generally picks out a subset of relevant attributes for enhanced performance and interpretability coupled with decreased computational costs and complexity [1]. Basically, feature selection procedures are labelled as filter-based methods and wrapper-based methods. Filter-based feature selection employed intrinsic features such as statistics or correlation in relation to how the feature is associated with the target variable. The features are ranked and analyzed based on the importance of attributes with the target while irrelevant attributes are then discarded [19].

TABLE I. PERFORMANCE INDICES FOR THE HINSELMANN DATASET

Configuration	Models	F1-score	Recall	Precision
Without SMOTE Without Feature Selection	LR	0.89	1.00	0.80
	DT	0.75	0.75	0.75
	RF	0.85	0.92	0.79
	BC	0.85	0.92	0.79
	Ada	0.80	0.83	0.77
	GB	0.75	0.75	0.75
	XGB	0.85	0.92	0.79
Without SMOTE With Feature Selection	LR	0.89	1.00	0.80
	DT	0.73	0.67	0.80
	RF	0.85	0.92	0.79
	BC	0.89	1.00	0.80
	Ada	0.80	0.83	0.77
	GB	0.80	0.83	0.77
	XGB	0.85	0.92	0.79
With SMOTE Without Feature Selection	LR	0.78	0.82	0.75
	DT	0.70	0.73	0.67
	RF	0.85	0.79	0.92
	BC	0.80	0.77	0.83
	Ada	0.75	0.75	0.75
	GB	0.75	0.75	0.75
	XGB	0.75	0.75	0.75
With SMOTE With Feature Selection	LR	0.78	0.75	0.82
	DT	0.85	0.92	0.79
	RF	0.85	0.92	0.79
	BC	0.75	0.75	0.75
	Ada	0.83	0.83	0.83
	GB	0.85	0.92	0.79
	XGB	0.85	0.92	0.79

TABLE II. PERFORMANCE INDICES FOR THE GREEN DATASET

Configuration	Models	F1-score	Recall	Precision
Without SMOTE Without Feature Selection	LR	0.80	1.00	0.67
	DT	0.73	0.80	0.67
	RF	0.75	0.90	0.64
	BC	0.73	0.80	0.67
	Ada	0.75	0.90	0.64
	GB	0.78	0.90	0.69
	XGB	0.80	1.00	0.67
Without SMOTE With Feature Selection	LR	0.80	1.00	0.67
	DT	0.80	0.80	0.80
	RF	0.75	0.90	0.64
	BC	0.82	0.90	0.75
	Ada	0.70	0.80	0.62
	GB	0.73	0.80	0.67
	XGB	0.75	0.90	0.64
With SMOTE Without Feature Selection	LR	0.75	0.90	0.64
	DT	0.74	0.70	0.78
	RF	0.78	0.90	0.69
	BC	0.74	0.70	0.78
	Ada	0.64	0.70	0.58
	GB	0.73	0.80	0.67
	XGB	0.87	1.00	0.77
With SMOTE With Feature Selection	LR	0.73	0.80	0.67
	DT	0.86	0.90	0.82
	RF	0.75	0.80	0.67
	BC	0.74	0.70	0.78
	Ada	0.64	0.70	0.58
	GB	0.50	0.50	0.50
	XGB	0.57	0.60	0.55

TABLE III. PERFORMANCE INDICES FOR THE SCHILLER DATASET

Configuration	Models	F1-score	Recall	Precision
Without SMOTE Without Feature Selection	LR	0.83	1.00	0.71
	DT	0.63	0.60	0.67
	RF	0.87	1.00	0.77
	BC	0.70	0.70	0.70
	Ada	0.67	0.60	0.75
	GB	0.63	0.60	0.67
	XGB	0.63	0.60	0.67
Without SMOTE With Feature Selection	LR	0.83	1.00	0.71
	DT	0.78	0.70	0.88
	RF	0.91	1.00	0.83
	BC	0.84	0.80	0.89
	Ada	0.63	0.60	0.67
	GB	0.82	0.90	0.75
	XGB	0.80	0.80	0.80
With SMOTE Without Feature Selection	LR	0.74	0.70	0.78
	DT	0.71	0.60	0.86
	RF	0.76	0.80	0.73
	BC	0.78	0.70	0.88
	Ada	0.59	0.50	0.71
	GB	0.76	0.80	0.73
	XGB	0.38	0.30	0.50
With SMOTE With Feature Selection	LR	0.74	0.70	0.78
	DT	0.63	0.50	0.83
	RF	0.84	0.80	0.89
	BC	0.78	0.70	0.88
	Ada	0.78	0.70	0.88
	GB	0.67	0.60	0.75
	XGB	0.80	0.80	0.80

With respect to feature selection techniques for cervical cancer diagnosis, Jelen et al., have concluded that utilizing feature selection optimized the handcrafted features for cervical cancer diagnosis which has shown to be a cost-efficient option in balancing accuracy with interpretability [16]. In [12], authors have also concluded that the incorporation of a feature selection could result to an optimum performance in predicting cervical cancer using only the optimal features in the dataset. Feature selection can upgrade the predictive capacity and interpretability of ML models by the removal of less important attributes which ultimately leads to a more accurate and reliable diagnosis of cervical cancer [15]. Recapitulating the advantages of incorporating a feature selection procedure, the simplification of models for easy interpretability coupled with faster training times and consequently decreased computational expense, leads to enhanced generalizability and improved predictive capability of ML models [18].

To interpret the predictions of our classification model, we utilized Local Interpretable Model-Agnostic Explanations (LIME). The visualizations generated for the Hinselmann, Green, and Schiller datasets provided insights into the feature contributions for individual predictions. The choice of classifier and training set for each dataset was based on the best-performing machine learning model for that particular dataset. An instance of ground truth 0 and 1 was chosen to be interpreted for each. In Figure 2 for the Hinselmann dataset, the instance with ground truth 0 had only 17% prediction probability for class 0, or 83% prediction probability for class 1, thus was misclassified. This can be attributed to the heavy class imbalance between the classes. The percentage contribution of the top features that led to a class 1 prediction for this instance were as follows: os_area=0 (2%), walls_area =0 (2%), rgb_cervix_r_std=0.33 (2%), rgb_cervix_g_std= (2%), and rgb_total_r_std=0.18 (1%). The instance with ground truth 1, the model correctly predicted a class 1 with an 82% prediction probability for class 1. The most influential features for class 1 for this instance were os_area=0.07 (2%), walls_area=0 (2%), and fit_cervix_bbox_total=1 (1%).

In Figure 3 for the Green dataset, the instance with ground truth 0 has 0.99 prediction probability for class 0, hence, was a correct classification. The most influential features were os_specularities_area=0(19%), rgb_total_g_mean_minus_std =0 (19%), fit_circle_total=0.2 (11%), cervix_area=0.2 (9%), fit_ellipse_rate=0.83 (7%), rgb_total_b_mean_plus_sd=0.33 (7%). The instance with ground truth 1 has a 99% prediction probability for class 1 and thus, was a correct classification. The most influential features were rgb_total_g_mean_minus_std=0.58 (16%), dist_to_center_os =0.44 (12%), rgb_cervix_g_mean_minus_std=0.40 (11%), fit_circle_total=0.8 (11%), rgb_total_b_mean_plus_std=0.53 (7%), fit_cervix_bbox_total=0.86 (5%). On the other hand for Figure 4 for the Schiller dataset, the instance with ground truth 0 has a 78% prediction probability for class 0 and was a correct classification. The top three most influential features were rgb_total_r_mean=0.49 (9%), rgb_cervix_r_mean=0.32 (4%), os_artifacts_area=0 (4%), rgb_cervix_b_mean=0.37 (4%), rgb_total_r_std=0.38 (4%). The instance with ground truth 1 has 93% prediction probability for class 1 and was a correct classification. The top three most influential features were rgb_total_r_mean_minus_std=0.62 (5%), os_area=0.31 (4%) rgb_total_r_mean=0.87 (3%), rgv_cervix_r_mean= 0.92 (3%), walls_area=0.17 (3%).

Overall, the LIME visualizations for the Hinselmann, Green, and Schiller datasets provided valuable insights into the key features influencing the model's predictions. While certain features consistently appeared as strong contributors to class "1" predictions, the model's decision-making process was also influenced by complex feature interactions and nonlinearities that may not always align with individual feature importance. These findings emphasized the need for a comprehensive approach to interpreting model predictions, considering both the local feature contributions and the broader behavior of the model across different instances. The integration of XAI further provided unprecedented insights into feature contributions, making this the first comprehensive approach to improving both the interpretability and reliability of models for colposcopy image quality assessment.

V. CONCLUSIONS AND RECOMMENDATION

The study evaluated the predictive capability of machine learning models under four configurations assessing the impact of feature selection and class imbalance handling for the binary classification task of digital colposcopy image quality applied to three colposcopy datasets. For the Hinselmann dataset, logistic regression and bagging classifer outperformed the other models with an F1-score of 0.89. In the Green dataset, the XGBoost model achieved the highest F1-score of 0.87 when SMOTE was applied to address the class imbalance. For the Schiller dataset, Random Forest was the best performing model obtaining an F1-score of 0.91 with an adoption of a feature selection.

Additionally, Explainable AI using LIME had identified some features as significant contributors in the predictions made by the models such as RGB intensity values and artifact-related metrics which hold valuable insights in understanding the models' decision-making processes. Through the evaluations performed on the selected models, a clearer and more nuanced understanding of the trade-offs between the performance of the models and the preprocessing techniques

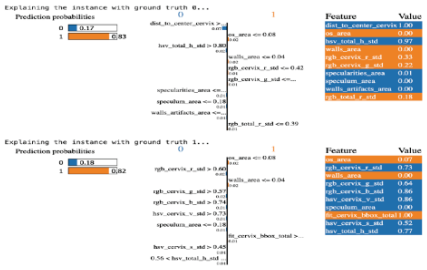

Fig. 2. LIME explanation for Hinselmann dataset using Logistic Regression

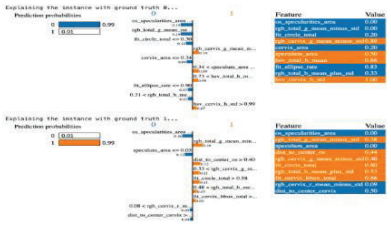

Fig. 3. LIME explanation for Green dataset using XGBoost

was also presented which highlighted how different configurations can significantly affect the performance results and emphasized the inherent necessity for balance in the optimization of various factors such as interpretability and model accuracy. However, in diverse clinical environments, extensive validation is necessary as the utilized dataset is challenged by its limited size and potential inherent biases.

Given these findings, it is recommended that future research place higher priority on the incorporation or utilization of a larger dataset containing more diverse image samples to improve the generalizability. Furthermore, the exploration of hyperparameter optimization techniques is also recommended to maximize the potential of the models for image quality classification such as in the case of boosting models. Integrating SHAP in future studies can also potentially augment and complement the insights gained from LIME, thus offering a deeper understanding of the models' behavior and feature importance. Finally, future studies may consider extending the feature engineering to include more nuanced image quality characteristics and include iterative refinement of preprocessing strategies, such as advanced augmentation techniques for further improvement in the robustness and adaptability.

REFERENCES

[1] E. Ileberi and Y. Sun, "Machine Learning-Assisted Cervical Cancer Prediction Using Particle Swarm Optimization for Improved Feature Selection and Prediction," in IEEE Access, vol. 12, pp. 152684-152695, 2024, doi: 10.1109/ACCESS.2024.3469869.

[2] H. Borda, P. Bloem, H. Akaba, et al., "Status of HPV disease and vaccination programmes in LMICs: Introduction to special issue," Vaccine, 42 Suppl 2, S1–S8, 2024. https://doi.org/10.1016/j.vaccine.2023.10.062.

[3] E. E. L. Jansen, N. Zielonke, A. Gini, et al., "Effect of organised cervical cancer screening on cervical cancer mortality in Europe: A systematic review," European Journal of Cancer, 127, 207–223, 2020. https://doi.org/10.1016/j.ejca.2019.10.018.

[4] F.S. Najib, M. Hashemi, Z. Shiravani, T. Poordast, S. Sharifi, and E. Askary, "Diagnostic Accuracy of Cervical Pap Smear and Colposcopy in Detecting Premalignant and Malignant Lesions of Cervix," Indian journal of surgical oncology, 11(3), 453–458, 2020. https://doi.org/10.1007/s13193-020-01118-2.

[5] M. Darwish, M.Z. Altabel, and R.H. Abiyev, "Enhancing Cervical Pre-Cancerous Classification Using Advanced Vision Transformer,"

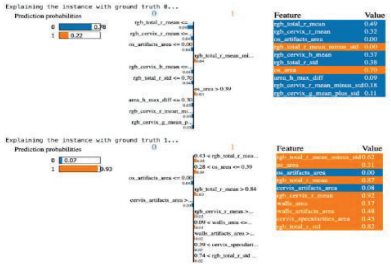

Fig. 4. LIME explanation for Schiller using Random Forest

Diagnostics 2023, 13, 2884. https://doi.org/10.3390/diagnostics13182884.

[6] J. Valls, A. Baena, G. Venegas, et al., "Performance of standardised colposcopy to detect cervical precancer and cancer for triage of women testing positive for human papillomavirus: Results from the ESTAMPA multicentric screening study," The Lancet Global Health, 11(3), e350–e360, 2023. https://doi.org/10.1016/S2214-109X(22)00545-9.

[7] P. Xue, M.T.A. Ng, and Y. Qiao, "The challenges of colposcopy for cervical cancer screening in LMICs and solutions by artificial intelligence," BMC Medicine, 18, Article number: 169, 2020. https://doi.org/10.1186/s12916-020-01613-x.

[8] L, Ekem, E. Skerrett, M.J. Huchko, and N. Ramanujam, "Automated image clarity detection for the improvement of colposcopy imaging with multiple devices," Biomedical Signal Processing and Control, Volume 100, Part B, 106948, 2025. https://doi.org/10.1016/j.bspc.2024.106948.

[9] S.R. Ahmed, B. Befano, D. Egemen, et al., "Generalizable deep neural networks for image quality classification of cervical images," Sci Rep 15, 6312, 2025. https://doi.org/10.1038/s41598-025-90024-0.

[10] E. Nikookar, E. Naderi, and A. Rahnavard, "Cervical Cancer Prediction by Merging Features of Different Colposcopic Images and Using Ensemble Classifier," J. Med. Signals Sens. 11(2), 67–78, 2021. https://doi.org/10.4103/jmss.JMSS_16_20.

[11] A. Wu, P. Xue , G. Abulisi, et al., "Artificial intelligence in colposcopic examination: A promising tool to assist junior colposcopists," Frontiers in Medicine, 10, 2023. https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2023.1060451.

[12] Y. Wang, L. Shen, JL. Jin, and G. Wang, "Application and Clinical Value of Machine Learning-Based Cervical Cancer Diagnosis and Prediction Model in Adjuvant Chemotherapy for Cervical Cancer: A Single-Center, Controlled, Non-Arbitrary Size Case-Control Study," Contrast media & molecular imaging, 2022, 2432291. https://doi.org/10.1155/2022/2432291.

[13] S. Kim, H. Lee, S. Lee, J.Y. Song, J.K. Lee, and N.W. Lee, "Role of Artificial Intelligence Interpretation of Colposcopic Images in Cervical Cancer Screening," Healthcare (Basel, Switzerland), 10(3), 468, 2022. https://doi.org/10.3390/healthcare10030468.

[14] M. Mehmood, M. Rizwan, M. Gregus Ml, and S. Abbas, "Machine Learning Assisted Cervical Cancer Detection," Front. Public Health 9, 788376, 2021. https://doi.org/10.3389/fpubh.2021.788376.

[15] R. Shakil, S. Islam, and B. Akter, "A Precise Machine Learning Model: Detecting Cervical Cancer Using Feature Selection and Explainable AI. J. Pathol. Inform. 15, 100398, 2024. https://doi.org/10.1016/j.jpi.2024.100398.

[16] L. Jeleń, I. Stankiewicz-Antosz, M. Chosia, and M. Jeleń, " Optimizing Cervical Cancer Diagnosis with Feature Selection and Deep Learning," Appl. Sci. 2025, 15, 1458. https://doi.org/10.3390/app15031458.

[17] K. Fernandes, J. Cardoso, and J. Fernandes, "Quality assessment of digital colposcopies [Dataset],". UCI Machine Learning Repository. https://doi.org/10.24432/C5C022.

[18] H.M. Farghaly, and T. A. El-Hafeez, "A high-quality feature selection method based on frequent and correlated items for text classification," Soft Comput 27, 11259–11274, 2023. https://doi.org/10.1007/s00500-023-08587-x.

[19] A. Moslemi, "A tutorial-based survey on feature selection: Recent advancements on feature selection," Engineering Applications of Artificial Intelligence, Volume 126, Part D, 2023, 107136. https://doi.org/10.1016/j.engappai.2023.107136.