A Comparative Study on the Effectiveness of Data Augmentation Techniques for Cervical Cancer Detection

All-Marufi Rahaman Sajon[1*], Sadia Parvin Ripa[2], Sadat Iqbal Priom[2], and Shamima Afrin Sweety[3]

[1] University of Dhaka, Dhaka 1000, Bangladesh
[2] East West University, Dhaka 1212, Bangladesh
[3] Shaheed Suhrawardy Medical College Hospital, Dhaka 1207, Bangladesh
*Email ID- allmarufirahaman-2019415666@devs.du.ac.bd

Abstract. This paper presents a comparative study of traditional and generative data augmentation techniques for cervical cancer detection using the UCI Cervical Cancer Risk Factors dataset. Conventional oversampling methods, namely SMOTE and ADASYN, are evaluated alongside advanced generative approaches, including diffusion-based (ForestDiffusion) and adversarial (CTGAN) models. Three machine learning classifiers, namely XGBoost, CatBoost, and TabNet, are employed to assess predictive performance across multiple metrics including accuracy, precision, recall, F1-score, and AUC. Experimental results reveal that the optimal augmentation strategy varies across different diagnostic targets, with TabNet consistently outperforming gradient boosting methods. Specifically, the ForestDiffusion–TabNet model achieved a 42.5 percent improvement in F1-score for Hinselmann prediction, while the SMOTE–TabNet model yielded increases of 112.12 percent and 134.7 percent for Schiller and Cytology predictions, respectively. Furthermore, the ADASYN–TabNet model enhanced Biopsy prediction performance by 57.89 percent. Ten-fold cross-validation confirmed model stability, though the persistent challenge of severe class imbalance limits performance on imbalanced test sets. These findings indicate that distinct augmentation methods capture complementary data characteristics, underscoring the potential of tailored augmentation strategies for robust cervical cancer screening.

Keywords: cervical cancer, data augmentation, class imbalance, ForestDiffusion, TabNet

1 Introduction

Globally, cervical cancer continues to be among the most common and lethal forms of cancer, representing a significant public health problem, especially in low- and middle-income countries. It ranks fourth in the global cancer mortality burden for women, with approximately 604,000 new cases and 342,000 cases

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_19

ending in death in 2020 according to the World Health Organization [1]. The disease continues to be a source of mortality, despite the existence of HPV vaccines and screening programs, for the simple reason that low-resource settings lack access to value early detection and treatment [2]. Access to early detection methods such as Pap smears, HPV tests, and colposcopy is associated with a significant reduction in mortality. These methods, however, are underutilized in low-resource settings because of restrictive factors such as expense, insensitivity, and invasiveness.

Optimized early detection of cervical cancer is hampered by the widespread presence of imbalanced datasets. Medical datasets utilized to train machine learning models to predict cervical cancer outcomes often contain severe class imbalance. In this context, class imbalance refers to the presence of non-cancerous cases being in a much greater number compared to the cancer-positive ones. This leads to the development of skewed predictive models that gain little proficiency in identifying the less prevalent cancerous cases. This, in turn, results in the inability to establish robust predictive models. As highlighted in earlier works, conventional machine learning approaches that ignore the class imbalance problem result in very poor sensitivity towards the less prevalent classes, which in turn results in inadequate performance in the detection of cervical cancer.

Focusing on data-driven techniques to enhance predictive accuracy in detecting cervical cancer, this paper investigates specific challenges in the domain. This study primarily seeks to compare and contrast several augmentation techniques, namely, the Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), Forest Diffusion [3] and Conditional Generative Adversarial Networks (CTGAN) [4], alongside three advanced classifiers: XGBoost, CatBoost, and TabNet. SMOTE and ADASYN, two widely used classical oversampling methods, alleviate the disproportionate bias of imbalanced datasets by creating synthetic instances of the underrepresented class. Unlike ForestDiffusion and CTGAN, which are considered sophisticated generative methods for data imbalance, SMOTE [6] and ADASYN [5] simply provide classical approaches, i.e., creating instances based on the statistical properties of the class data to be imbalanced, instead of discernibly learning the internal ideation patterns of the data so as to provide a more flexible and advanced solution to the problem of imbalance.

The development of cervical cancer is explained medically by the pathophysiogenesis of precancerous lesions resulting from persistent infections of the high-risk HPV types, especially HPV 16 and 18. These infections may result in the development of cervical intraepithelial neoplasia (CIN) that may progress to malignant carcinoma if not recognized and managed early [2]. From a clinical perspective, the greatest value lies in the capacity to detect these lesions early during the pre-invasive phase, as this will markedly reduce the risk of the cancer becoming invasive. When this happens, the cancer treatment becomes even more challenging and may involve radical surgery and chemotherapy. This study highlights the expanding evidence that machine learning, and especially

the combination of these tools with data augmentation, may improve early cervical cancer detection.

This study contributes the first review of the literature on classical and generative augmentations and the first study to address the weaknesses of classical methods for dealing with imbalance within datasets. The use of advanced methods like ForestDiffusion and CTGAN contributes to model predictability and robustness, making the tool for cervical cancer detection more accurate. Moreover, this research is the only one to use an expert approach integrating the research with insights from the domain of pathology and medicine, which notes the clinical relevance of the findings to pathology.

This research combines novel approaches to data augmentation with state-of-the-art machine learning models (like TabNet) to improve the precision and the consistency of the models for detecting cervical cancer. As it is one of the first studies to address clinical decision making by machine learning augmentations, incorporating the findings into practice will transform the cervical cancer diagnostic process, enhance patient care, and lessen the global prevalence of cervical cancer. The study is the first of its kind to comprehensively assess the predictive performance of multiple models to derive actionable insights on how to construct machine learning models that are not only accurate and clinically relevant but also scalable to the demands of practice in the field of machine learning in medicine.

2 Related Work

Infection with high-risk strains of human papillomavirus (HPV) is the most common cause of cervical cancer. It is a growing public health issue, particularly in low-and middle-income countries. To this day, cervical cancer is the fourth most common cause of cancer mortality in women, with an estimated 604,000 new cases and 342,000 deaths globally in 2020. This is the case even with the implementation of screening programs and the HPV vaccine. More countries are beginning cervical cancer screening. Targeted approaches, including Pap smear and HPV testing, focus on cervical cancer prevention. Each has proven to be a crucial approach for early detection and diagnosis; however, their low sensitivity, expense, and necessity of specialized training and equipment are obstacles to more widespread application. There is rising interest, even more pronounced in developing countries with less access to cervical cancer screening and treatment, in the use of machine learning (ML) algorithms designed to process and classify datasets with unbalanced class distributions. Pap smear tests, HPV testing, and machine learning techniques in cancer detection and classification have proven to be crucial in early diagnosis and management of cervical cancer. However, most of these models suffer from class imbalance, particularly the underrepresentation of cancer cases in the healthy population.

To account for this issue, a number of augmentation strategies, particularly Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN), have been widely used to create synthetic instances of the

underrepresented class and thus, lessening the adverse consequences of imbalanced data on the accuracy of the model.

The application of Explainable Artificial Intelligence (XAI) in cervical cancer ML systems is rapidly gaining attention, especially in healthcare, where users ought to have confidence in the system's conclusions. XAI systems, e.g. SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), facilitate physician access to the rationale behind the models, which is critical in high-risk scenarios. Moreover, there is literature directed towards hybrid feature selection approaches bringing together classical approaches (e.g., correlation-based feature selection) and advanced techniques (e.g., Recursive Feature Elimination, RFE). These effectively streamline the surgical feature set of the cervical cancer prediction models to the clinical cancer prediction framework. Researchers recently integrated SHAP and LIME to RF, XGBoost, and Logistic Regression ensemble models and obtained 98 percent accuracy with 99.50 percent AUC for cancer detection, which, paired with expert validation, reinforces the credibility of the models' predictions[7].

Even though methods such as SMOTE and ADASYN have been used for quite some time, methods such as ForestDiffusion and Conditional Generative Adversarial Networks (CTGAN) have been starting to pick up use due to their ability to accurately and meaningfully replicate synthetic data for various scenarios. Specifically, CTGAN has been used for prediction of cervical cancer to resolve the data set issue by creating synthetic representations of the underrepresented classes modeled to increase the cancer cases within the training set. One of the recent studies which involve the application of CTGAN for the prediction of cervical cancer reported the prediction performance improvement to be additional to the other features. This study showed that using CTGAN to augment the training datasets improved the performance of Logistic Regression models to a record of 99 percent accuracy with also considerable precision, recall, and F1 scores. This points to CTGAN as one of the methods that helps resolve class imbalance for other methods along with improvement for the models' general performance and generalizability [8].

As noted above, CTGAN has been utilized in predicting cervical cancer, while ForestDiffusion is yet to be explored in the field. ForestDiffusion is a relatively new generative technique that synthesizes data by disseminating information through a forest of decision trees and diffusion, thereby in an effort to construct data that aligns closely with the actual distribution of the data. Although still to be applied to the field of cervical cancer, it has the ability to tackle both the issue of critical lack of data and the access to a wider range of synthetic instances. To the best of the researcher's knowledge, this is the first study that attempts to apply ForestDiffusion to cervical cancer datasets which in turn can be a new approach to enhancing the prediction accuracy as well as the robustness of the model in this critical field of medicine.

This paper aims to pioneer the integration of ForestDiffusion and CTGAN to develop more dependable, robust, and interpretable machine-learning models for cervical cancer prediction. The research aims to develop a transparent and

trustworthy system to provide healthcare practitioners with reliable assistance for early cancer detection by combining generative augmentation techniques with explanations via XAI and XAI-driven expert model refinement. The framework proposed enables high model fidelity while ensuring easy explainability, thus improving clinician's decision-making on patient management. This research marks a significant development in making the detection and diagnosis of cervical cancer more accurate and accessible with AI.

3 Methodology

The proposed methodology involves five main stages: data preprocessing, data augmentation, model training, evaluation, and comparison.

3.1 Dataset

The dataset used is the UCI Cervical Cancer Risk Factors dataset, containing 858 samples and 32 attributes with four binary target labels: Hinselmann, Schiller, Cytology, and Biopsy. The 32 attributes cover a diverse set of factors involving reproductive lifestyle, medical history and demography. The factors make the dataset particularly apt for medical data analysis. The small to moderate size and noisy nature are also reflective of real-world medical datasets, where positive diagnoses in tests like Hinselmann, Schiller etc. are relatively rare. After some exploratory data analysis, only 21 features were kept. The decision of dropping some variables hinged on the variables having high values of Variable Inflation Factor (VIF) and the authors' domain knowledge. However, a couple of variables were not dropped despite having values in excess of 10, as they were deemed necessary.

3.2 Data Preprocessing

Missing values were imputed using mean/mode techniques, and categorical features were encoded numerically. This approach helps preserve the overall distribution of the data while minimizing information loss. The dataset was later normalized before fitting to ensure uniform scaling of feature values, putting each feature on equal footing.

3.3 Data Augmentation Techniques

Several augmentation methods were applied for this study. SMOTE was utilized as it interpolates new samples among minority class points. ADASYN was employed which generates samples for harder-to-learn minority samples. Forest-Diffusion, which generates synthetic data through stochastic diffusion modelling, was also applied. Finally, CTGAN was also utilized which uses conditional Generative Adversarial Networks (GANs) to create realistic synthetic data in tabular format.

Table 1. Variable Inflation Factor (VIF) of the considered features

No.	Variable	VIF
0	Age	22.941
1	STDs: Number of diagnosis	10.203
2	Number of sexual partners	3.383
3	First sexual intercourse	17.693
4	Num of pregnancies	5.082
5	Smokes	2.632
6	Smokes (years)	3.994
7	Smokes (packs/year)	2.251
8	Hormonal Contraceptives	3.680
9	Hormonal Contraceptives (years)	1.705
10	IUD	2.699
11	IUD (years)	2.469
12	STDs:vaginal condylomatosis	1.159
13	STDs:vulvo-perineal condylomatosis	4.833
14	STDs:syphilis	2.427
15	STDs:pelvic inflammatory disease	1.132
16	STDs:genital herpes	1.125
17	STDs:molluscum contagiosum	1.127
18	STDs:HIV	3.222
19	STDs:Hepatitis B	1.078
20	STDs:HPV	1.080

3.4 Classification Models

Three machine learning models were trained on the original and the augmented version of the dataset. Extreme Gradient Boost or XGBoost, a highly efficient gradient boosting framework, was selected as it performs well on structured tabular data, and it better fits complex, non-linear interactions among the features. CatBoost was included due to its built-in handling of categorical variables and its being robust to overfitting, which makes it particularly suitable for heterogeneous, mixed-type datasets. Moreover, TabNet, a deep learning model that has sequential attention mechanisms, was used to investigate the possible benefits of representation learning in tabular classification tasks.

Initially, popular machine learning models such as Logistic Regression, Support Vector Machine (SVM) and Random Forest were taken into account. But, with low accuracy (approximately 50 percent), they were excluded from the study and the emphasis was completely shifted on the gradient boosting and attention-based methods.

3.5 Evaluation Metrics

A set of classification metrics was used to evaluate model performance. Accuracy offered an overall measure of correctly classified cases. Precision reflected how well the models reduced false alarms, while recall indicated their ability to capture true positives and limit false negatives. The F1-score provided a balanced view by combining precision and recall, particularly useful under class imbalance. Finally, the AUC assessed each model's ability to distinguish between positive and negative classes across different thresholds.

4 Results

Experimental results indicate that models trained on generative augmentation techniques (like CTGAN) often outperform classical ones in terms of accuracy. However, F1-score becomes a much more important metric where severe class imbalance is present. Cross-validation didn't serve to improve performance very much, if at all, as the test set was not augmented and held distribution similar to the original imbalance dataset. Only the training set was augmented to test if the trained model could distinguish patterns between a cervical cancer-positive patient and a healthy patient, even in imbalanced scenarios. The results of the models are as follows:

4.1 Hinselman Prediction

Table 2. Results for Hinselmann

Model	Augmentation technique	Accuracy	Precision	Recall	F1 Score	AUC
TabNet	Forest Diffusion	0.826	0.176	0.750	0.286	0.732
CatBoost	Original	0.953	0.500	0.125	0.200	0.518
TabNet	Original	0.890	0.133	0.250	0.174	0.582
TabNet	ADASYN	0.767	0.079	0.375	0.130	0.767

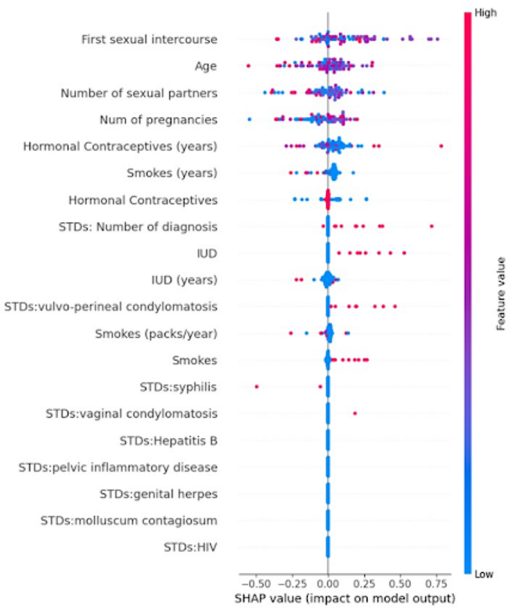

Fig. 1. SHAP for Hinselmann Predicting Model (Forest Diffusion-TabNet)

The SHAP Kernel Explainer indicates that the model's predictions for Hinselmann test positivity are primarily contingent upon variables reflecting sexual behavior, infection history, hormonal exposure, and smoking habits (see Fig. 1). Earlier age at first sexual intercourse and a higher number of sexual partners exhibit the strongest positive contributions, suggesting that increased exposure

to high risk types of (HPV 16 and HPV 18 viruses) remain a dominant risk factor.Older women show elevated predicted risk, potentially due to biological and environmental negative influences that accumulates over time. STD related indicators, including total STD count, number of diagnoses, and specific infections such as syphilis, herpes, and HPV, consistently push predictions upward. Smoking duration and intensity further raise risk estimates in line with established clinical evidence. Hormonal contraceptive use (eg. Oral Contraceptive Pills-OCP, intradermal implants), measured both in duration and usage status, demonstrates a moderate positive effect, while both IUD(eg. Copper IUD,hormonal IUD) use and years of IUD exposure show a mild but noticeable contribution. Additionally, infections such as vulvo-perineal condylomatosis and pelvic inflammatory disease(PID) exert a persistent positive influence on the model output.

Table 3. Hinselmann prediction results of ForestDiffusion-TabNet after Cross-Validation

Metric	Mean Across 10 Folds
Accuracy	76.981
Precision	0.067
Recall	0.300
F1-Score	0.108
AUC	0.659

The ForestDiffusion-TabNet model for Hinselmann prediction was further evaluated through a 10-fold cross-validation, but performance metrics remained comparable to the initial test results, as the validation maintained the original data distribution with severe class imbalance. The results on the test set are given in Table 3.

4.2 Schiller Prediction

Table 4. Results for Schiller

Model	Augmentation technique	Accuracy	Precision	Recall	F1 Score	AUC
TabNet	SMOTE	0.779	0.286	0.824	0.424	0.832
TabNet	ADASYN	0.709	0.200	0.647	0.306	0.735
TabNet	CTGAN	0.878	0.300	0.176	0.222	0.558
TabNet	Forest Diffusion	0.628	0.121	0.571	0.200	0.535

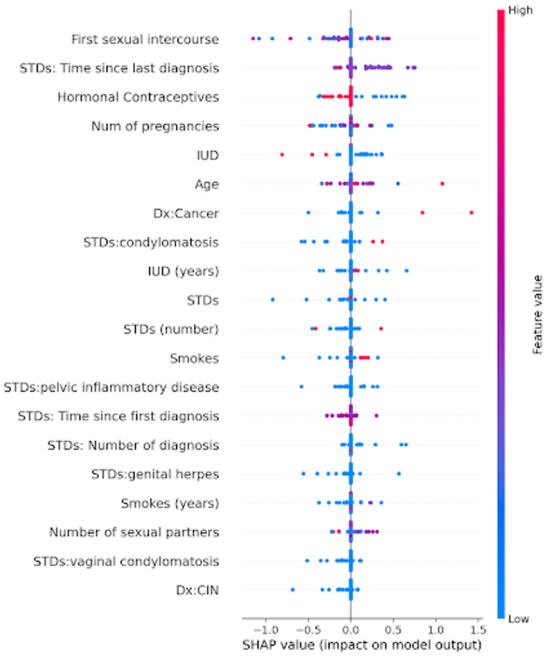

Fig. 2. SHAP for Schiller Predicting Model (SMOTE-TabNet)

Schiller prediction model is most strongly influenced by variables related to sexual behavior and recent infection history, as seen in Fig. 2. Early onset of sexual intercourse, and recency of STD diagnoses seem to substantially elevate risk measurements. Hormonal contraceptive use, number of pregnancies, IUD use, and age, collectively determine the model's intermediate-level predictions. General STD indicators and their counts also contribute to moderate positive effects.

Table 5. Schiller prediction results of SMOTE-TabNet after Cross-Validation

Metric	Mean Across 10 Folds
Accuracy	75.932
Precision	0.092
Recall	0.214
F1-Score	0.128
AUC	0.575

The SMOTE-TabNet model for Schiller prediction was further evaluated through a 10-fold cross-validation, but performance metrics remained comparable to the initial test results, as the validation maintained the original data distribution with severe class imbalance. The results on the test set are given in Table 5.

4.3 Cytology Prediction

Table 6. Results for Cytology

Model	Augmentation technique	Accuracy	Precision	Recall	F1 Score	AUC
TabNet	SMOTE	0.773	0.114	1.000	0.204	0.824
TabNet	CTGAN	0.837	0.103	0.600	0.176	0.661
TabNet	Forest Diffusion	0.756	0.095	0.500	0.160	0.613
TabNet	Original	0.878	0.056	0.200	0.087	0.614

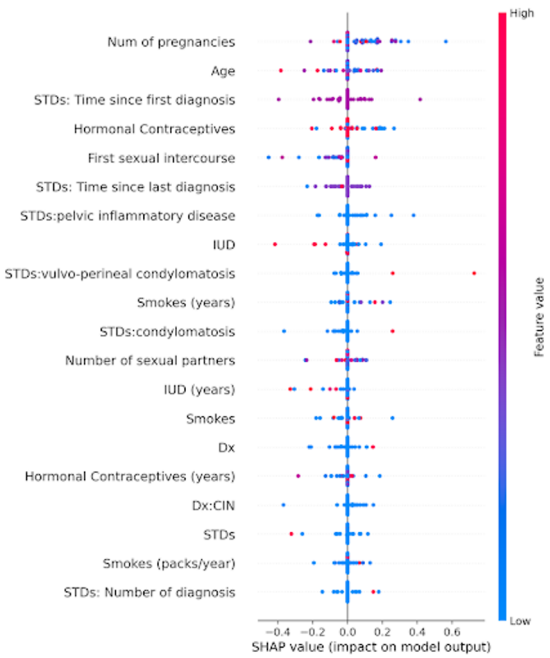

Fig. 3. SHAP for Cytology Predicting Model (SMOTE-TabNet)

Cytology prediction model indicates that reproductive history and STD timing are the most influential predictors, as seen in Fig. 3. Higher number of pregnancies, and age are the biggest contributors to risk measurements, along with first or last diagnoses of STDs. Duration of hormonal contraceptives and early onset of sexual activity exhibit moderate yet consistent contributions toward higher prediction scores. Additionally, pelvic inflammatory disease and vulvo-perineal condylomatosis further increases the model output.

Table 7. Cytology prediction results of SMOTE-TabNet after Cross-Validation

Metric	Mean Across 10 Folds
Accuracy	77.211
Precision	0.041
Recall	0.175
F1-Score	0.066
AUC	0.531

The SMOTE-TabNet model for Cytology prediction was further evaluated through a 10-fold cross-validation. The validation results remained consistent with initial test performance, reflecting the inherent limitations imposed by severe class imbalance in the unaugmented test set, which was intentionally preserved to simulate real-world screening conditions. The results on the test set are given in Table 7.

4.4 Biopsy Prediction

Table 8. Results for Biopsy

Model	Augmentation technique	Accuracy	Precision	Recall	F1 Score	AUC
TabNet	ADASYN	0.837	0.185	0.455	0.263	0.682
TabNet	Forest Diffusion	0.767	0.150	0.500	0.231	0.631
TabNet	CTGAN	0.581	0.114	0.818	0.200	0.770
XGBoost	CTGAN	0.942	1.000	0.091	0.167	0.657

The most influential predictors of biopsy outcomes are features related to sexual history and infection status (see Fig. 4). Earlier age at first intercourse, number of

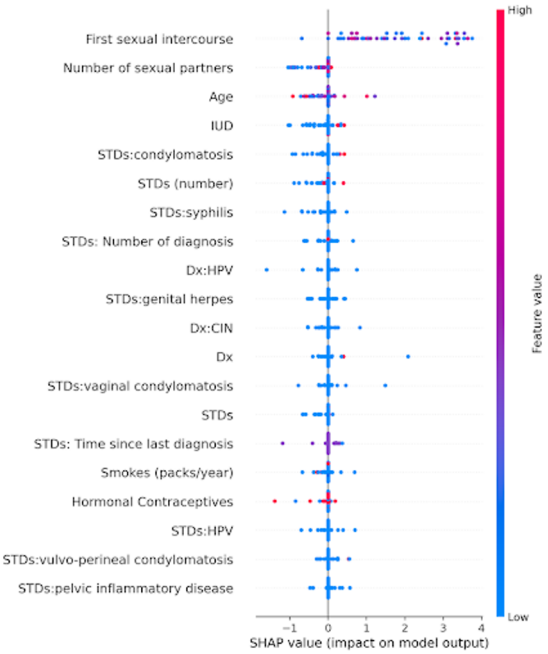

Fig. 4. SHAP for Biopsy Predicting Model (ADASYN-TabNet)

sexual partners, and age are the biggest contributors. IUD usage exhibits a mixed pattern, with both low and high usage levels influencing predictions in either direction, which suggests that it has heterogeneous clinical associations. Presence of sexually transmitted diseases (STDs), particularly condylomatosis, syphilis, and HPV, keep shifting the predictions toward a greater risk level. Diagnosis-related variables, such as HPV and CIN (Cervical Intraepithelial Neoplasia), also show moderate but biologically consistent influence.

Table 9. Biopsy prediction results of Adasyn-TabNet after Cross-Validation

Metric	Mean Across 10 Folds
Accuracy	83.711
Precision	0.175
Recall	0.367
F1-Score	0.235
AUC	0.737

The Adasyn-TabNet model for Biopsy prediction was further evaluated through a 10-fold cross-validation, but performance metrics remained comparable to the initial test results, as the validation maintained the original data distribution with severe class imbalance. However, the AUC did improve. The results on the test set are given in Table 9.

5 Discussion

5.1 Model Performance and Augmentation Strategy Selection

Our experimental results demonstrate that no single augmentation technique universally outperforms others across all cervical cancer screening targets. Traditional methods, namely SMOTE and ADASYN, proved most effective for three out of the four predictions (Schiller, Cytology, and Biopsy), while a more recent generative approach like Forest Diffusion emerged as the optimal augmentation strategy for predicting Hinselmann. These findings challenge the notion that newer generative methods generally surpass classical techniques, perhaps especially for medical datasets in which class imbalance is very high. Stronger performance of Forest Diffusion in enhancing Hinselmann prediction (F1-score: 0.286, AUC: 0.732) likely reflects the ability of it in modelling complex non-linear interactions among features relevant to this particular response. In contrast, better performance of SMOTE and ADASYN on other targets likely indicate that simpler, interpolation-based augmentation works well in cases with less complex class boundaries.

5.2 TabNet's Performance on Imbalanced Tabular Data

Although it is well established that gradient boosting methods (like XGBoost and CatBoost) generally dominate in small-medium sized tabular datasets, in this study, TabNet consistently outperformed them in predicting all the response variables despite the data having severe class imbalance. Attention-based architectures like TabNet are able to focus discriminative salient features even when the data is tabular. Its built-in attention mechanism allowed it to focus on patterns in the data and better identify positive cases.

5.3 Cross-Validation and Model Generalization

A ten-fold cross-validation was performed to assess the stability and reproducibility of the results obtained. The results on the test set appeared to be consistent with best performing models (without cross-validation), matching the initial results well enough (Hinselmann: 76.98 percent, Schiller: 75.93 percent, Cytology: 77.21 percent, Biopsy: 83.71 percent). However, results obtained from cross-validation were not free from the effect of severe class imbalance evident by the low precision scores (0.04-0.18) and moderate recall (0.18-0.37). Nonetheless, this is expected as the test set (which was intentionally not augmented to mirror

the real world) had very few cases where the responses were positive. Notably, the Biopsy prediction model showed improved AUC after cross-validation (from 0.682 to 0.737), suggesting elevated risk-ranking capability despite classification threshold limitations. This finding implies that the model learned robust probabilistic risk estimates that may be more clinically useful for patient stratification than binary classification alone.

5.4 Clinical Interpretability and Feature Importance

SHAP analysis revealed that best-performing models consistently prioritized clinically established risk factors: sexual behavior variables (age at first intercourse, number of partners), STD history (timing, specific infections), and reproductive factors (pregnancies, hormonal contraceptive use). Prediction-specific patterns aligned with known cervical cancer etiology: Hinselmann emphasized early sexual activity and STD history (consistent with HPV transmission patterns); Schiller focused on recent STD diagnoses (reflecting active infection impact); Cytology prioritized pregnancies and age (hormonal and cumulative exposure); and Biopsy depended most on cumulative sexual exposure and high-risk STDs (HPV, syphilis). The consistency of these patterns across augmentation techniques suggests genuine biological signals rather than artifacts. In cervical cancer screening, low recall (missed diagnoses) poses greater clinical risk than low precision (unnecessary follow-ups). The moderate-to-high recall scores obtained support the models' potential utility in identifying high-risk patients, while F1-scores provide balanced performance assessment under severe class imbalance.

5.5 Implications for Cervical Cancer Screening

Our findings suggest several practical implications for machine learning-based cervical cancer screening. Firstly, rather than defaulting to the newest generative methods, practitioners should empirically evaluate multiple augmentation strategies for each response variable. Traditional methods remain competitive and may even offer better results occasionally while suffering from a stability-complexity trade-off. Secondly, given the consistently low precision across all tasks, instead of making a final diagnosis, these models may be more useful in corroborating doctors' endeavours in identifying higher-risk patients. The moderate-to-good AUC values (0.53-0.74) indicate potential efficacy in identifying high-risk patients warranting prioritized follow-up screening. Lastly, while augmentation improved model learning on training data, severe class imbalance related fundamental challenge persists during real-world application. Future work should explore complementary approaches, including cost-sensitive learning, optimized decision thresholds, or hybrid ensemble methods.

5.6 Limitations and Future Directions

This study has several limitations which need to be acknowledged. The dataset remains quite small even after augmentation, which may limit the models' capa-

bility to learn. Cross-validation confirmed model stability but could not improve performance on the imbalanced test, mirroring distributions found in the real world. Future research should investigate more recent augmentation techniques like TabSyn, while attention-based models like TabTransformer, FT-Transformer, SAINT and TabNet++ may further reveal efficacy of attention in cervical cancer screening. Additionally, while SHAP analysis provides valuable insights into feature importance, causal inference methods would be needed to distinguish true causal risk factors from correlated biomarkers. Future clinical trials will be needed to determine whether using the model to guide screening actually leads to better patient outcomes than current standard practices.

6 Conclusion

This study investigated the application of various data augmentation techniques combined with machine learning models with the purpose to screen cervical cancer across four diagnostic tests. Results of this study demonstrate that model performance is significantly impacted by task-specific augmentation techniques, with SMOTE-TabNet, Forest Diffusion-TabNet, and ADASYN-TabNet achieving the best results for Schiller, Hinselmann, and Biopsy tests respectively. The models utilized in this study show promise in identifying high-risk patients with moderate-to-good AUC values (0.53-0.74). However, challenges with respect to severe class imbalance and low precision persist. These findings contribute to the growing body of literature on applying machine learning for medical diagnostics, and emphasizes that empirical evaluation is necessary in selecting appropriate augmentation strategies for medical datasets with high imbalance.

References

1. World Health Organization: Cervical Cancer. https://www.who.int/health-topics/cervical-cancer (2020). Accessed 10 Oct 2025
2. González-Rodríguez, J.C., Cruz-Valdez, A., Madrid-Marina, V.: Cervical cancer prevention by vaccination: review. Front. Oncol. 14, 1386167 (2024). https://doi.org/10.3389/fonc.2024.1386167
3. Jolicoeur-Martineau, A., Fatras, K., Kachman, T.: Generating and Imputing Tabular Data via Diffusion and Flow-based Gradient-Boosted Trees. arXiv preprint arXiv:2309.09968v3 (2023)
4. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling Tabular Data using Conditional GAN. In: Advances in Neural Information Processing Systems, vol. 32, pp. 1–11 (2019)
5. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328. Hong Kong (2008). https://doi.org/10.1109/IJCNN.2008.4633969
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16, 321–357 (2002). https://doi.org/10.1613/jair.953
7. Roy, P., Hasan, M., Islam, M.R., Uddin, M.P.: Interpretable artificial intelligence (AI) for cervical cancer risk analysis leveraging stacking ensemble and expert knowledge. Digit. Health 11, 20552076251327945 (2025). https://doi.org/10.1177/20552076251327945
8. Tang, M., Chen, H., Lv, Z., Cai, G.: Diagnosis of Cervical Cancer Based on a Hybrid Strategy with CTGAN. Electronics 14(6), 1140 (2025). https://doi.org/10.3390/electronics14061140