Jetwyn Wilson
CSC 5800

Cervical Cancer Prediction using Data Mining

Overview of the Problem

Cervical cancer is one of the leading causes of cancer-related deaths among women worldwide. Cervical cancer is caused by the growth of the human papillomavirus on the cervix. The effect is passed through sexual contact; however, most patients' immune systems are able to fight it. For patients who are not able to fight off the virus, it can turn into cancer cells.

Machine learning techniques are helpful for early detection in women to predict cervical cancer. They use patient data and behavioral risk factors to guide cancer detection. Both early detection and diagnosis are essential to help manage cervical cancer. By providing early detection, patient outcomes will improve. Providing accurate diagnosis will help prevent overtreatment and stress for patients.

Thus, the goal of the project is to create a model that can accurately predict cervical cancer patients for screening and diagnosis.

Dataset Overview

The study uses a patient's behavior and demographic attributes that could possibly cause cervical cancer. The data were obtained from the UCI Machine Learning Repository (specifically, the “Cervical Cancer (Risk Factors)” dataset) and consist of 858 instances and 36 features encompassing demographic, behavioral, and medical test categories. A small portion of values is missing across some features and must be carefully handled. Among these 36 features, several are binary while others are numerical, creating additional complexity for classification tasks. There is also a class imbalance, with the majority class representing non-cancer patients.The challenge for this dataset is to create a model that can accurately identify positive cancer patients through health records and behavioral attributes that can be relied upon for screening.

Data Preprocessing

After loading the data, the dataset had to be cleaned because of missing values, categorical feature encoding, and severe class imbalance. To address missing values, the median was used to preserve the dataset’s statistical distribution while being more robust to outliers. A total of 13 columns contained missing values, with some features missing over 10% of their entries. Categorical encoding was then applied so that binary features could be turned into 0s and 1s, and features with more than two categories were one-hot encoded to ensure the model could interpret them properly. For standardization and scaling, StandardScaler was used to help improve the models’ performance and convergence.

To handle the class imbalance, Adaptive Synthetic Sampling (ADASYN) was applied to generate synthetic minority‑class instances, helping the model generalize to cancer‑positive cases. Hyperparameter tuning was also performed to enhance each model's performance. For Random Forest, GridSearchCV was used to adjust n_estimators, max_depth, min_samples_split, min_samples_leaf, bootstrap, and class_weight. Hyperparameter tuning with GridSearchCV was used for XGBoost as well, with parameters including a learning rate of 0.05, a max_depth of 3, 200 estimators, and a subsample setting of 1. With ADASYN, sampling_strategy was set to 0.7 to generate a sufficient minority class for improving recall. After applying these methods, the training data expanded to 1,284 samples (with 33 features), and the test set contained 172 samples, making it possible to thoroughly evaluate each model's performance. Data was split into 80% training and 20% testing.

Algorithm Choice

Recent studies have shown how effective machine learning can be for predicting cervical cancer. Mahto and Sood (2024) applied ADASYN to a small, highly imbalanced cervical cancer dataset using Random Forest, XGBoost, and a voting classifier to achieve high accuracy across minority classes. The authors used different hyperparameters and 3-fold cross-validation to reach 97.13% across all three models, demonstrating how ADASYN can significantly improve predictions for minority classes.

Akter et al. (2021) reported results similar to those of Mahto and Sood (2022) but emphasized that Random Forest was the strongest predictor for clinical diagnosis, achieving 93.33%. However, for early detection, both XGBoost and Random Forest can be valuable in identifying early stages of cervical cancer.

A similar group of researchers, Tanimu et al. (2022), used comparable models—including Adaptive Boosting and Gradient Boosting—but they also introduced a hybrid sampling method combining SMOTE for oversampling and Tomek for undersampling. Additionally, they employed Recursive Feature Elimination (RFE) with Least Absolute Shrinkage and Selection Operator (LASSO) for feature selection. Their results showed 98.72% accuracy and 100% sensitivity, indicating that decision trees can effectively handle selective features while addressing class imbalance.

While these studies show strong performance, most of them focus on either sampling techniques or boosting methods alone, without combining both strategies. This project builds on that by applying ADASYN with XGBoost to improve recall, while also evaluating the trade-off in precision. Unlike previous work that emphasized either oversampling (like SMOTE) or ensemble classifiers, this approach focuses on the integration of sampling and boosting to better detect minority cancer cases.

In addition, none of the reviewed studies explored how simple models like Random Forest perform when directly compared to ADASYN-enhanced models. By doing so, this project helps evaluate whether model complexity or sampling strategy contributes more to performance in an imbalanced cervical cancer dataset.

Mehmood et al (2021) mentions the importance of using hyperparameters to help improve diagnosis and early detection for cervical cancer. The author uses shallow neural networks and is trained by a scaled gradient backpropagation method. This technique helps the experiment with feature selection using Random forest. This was able to reduce noise by selecting only the most important features. This reduces the complexity and gives a faster and easier model to interpret. The study was able to achieve a high true positive rate to not miss any patents that are positive for early detection. The shallow neural network is great for a clinical setting because it does not require retraining and is great in a limited resource environment. Another benefit it provided was an accuracy above 90% and with minimal computational cost.

These findings highlight the importance of balancing recall and precision to minimize both missed diagnoses and false positives. Drawing on these studies provides insight into building a model that helps detect cervical cancer at early stages and diagnosis. The selected algorithms for this project are Random Forest, XGBoost, and ADASYN + XGBoost. Random Forest offers interpretability and handles diverse feature types without overfitting, while XGBoost helps distinguish which ensemble approach might be more effective. Lastly, applying ADASYN with XGBoost addresses the minority-class imbalance to improve detection of positive cases.

Analysis Result

Model	Accuracy	Precision	Recall	F1-score
XGBoost (Base Model)	94.77%	58%	64%	61%
XGBoost (Best Parameters)	94.18%	53%	73%	62%
ADASYN +XGBoost	94.77%	56%	82%	67%
Random Forest	94.77%	60%	55%	57%

The primary focus is for the models to analyze the minority class (class 1) that represents cancer. A base model was used to provide a standard level of performance before adding improvements, as it is important to see how the model performs on its own. XGBoost's base model achieved an accuracy of 94.77%, with a precision of 58% and a recall of 64%. These results are consistent with the study by Tanimu et al. (2022), which emphasizes the effectiveness of XGBoost by itself.

An optimized version of XGBoost was then created using the best tuning parameters, resulting in an improved recall of 73% and an F1-score of 62%. This reflects the findings of Mahto and Sood (2024), demonstrating the importance of hyperparameter tuning in increasing sensitivity for detecting cervical cancer. When ADASYN was applied to XGBoost, the model achieved the highest recall of 82% while raising precision to 56%, leading to the highest F1-score and thus the best harmony between precision and recall. These results closely align with Mahto and Sood (2024), showing how ADASYN can enhance sensitivity in an imbalanced dataset.

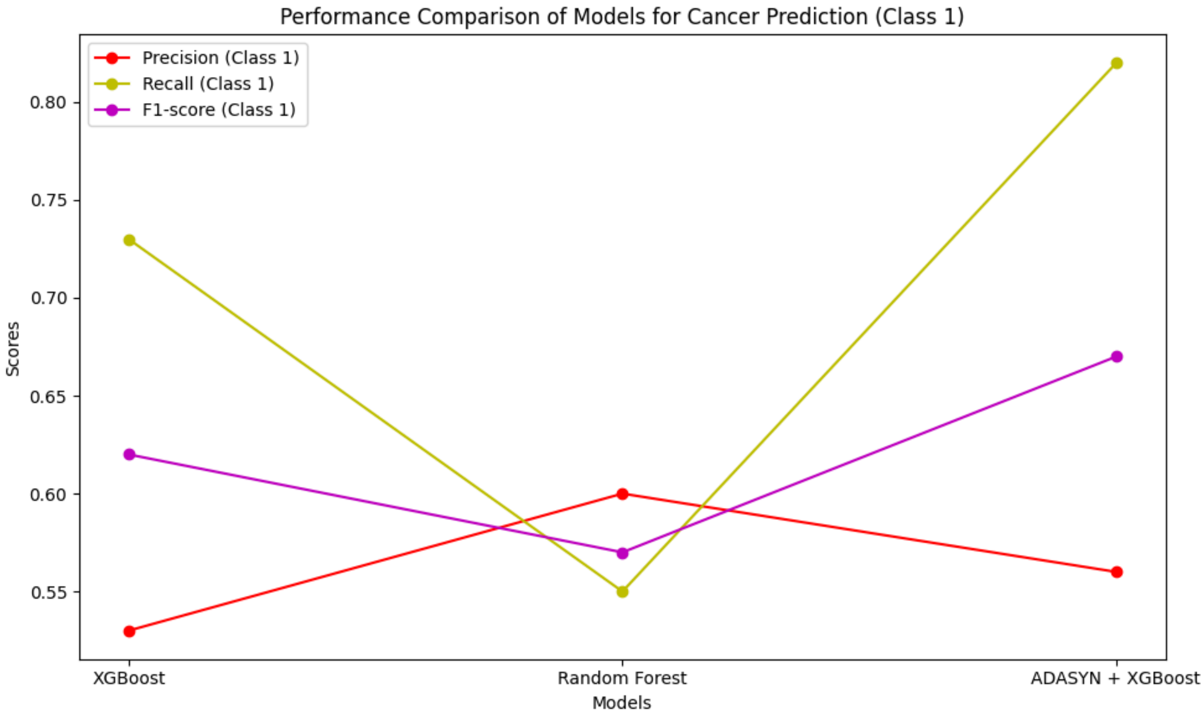
Random Forest achieved the highest precision at 60%, yet it posted the lowest recall at 55%, which lowered its F1-score to 57%. These findings support Akter et al. (2021), indicating how minimizing false positives could be beneficial for clinical diagnostics.

Comparison

ADASYN-enhanced XGBoost achieved the best recall, which is ideal for early detection of actual cancer cases. High recall is particularly valuable because it ensures positive cancer patients are identified. As shown in the figure, ADASYN has the highest recall rate, while Random Forest has the highest precision. Random Forest's ability to provide the highest precision is critical for diagnosing patients who are truly positive for cervical cancer, as it minimizes false positive results. The figure reflects these findings, illustrating how each model predicts the minority classes.

Although maximizing recall helps reduce missed cancer diagnosis, it is also essential to maintain a decent precision level so that fewer healthy individuals are flagged. In the future, further optimization of ADASYN and XGBoost to improve precision, or exploring hybrid approaches, could leverage both high precision and recall—balancing early detection with minimal false positives.

Figure 1: Comparing Models



The figure above analyzes the performance of XGBoost, Random Forest, and ADASYN with XGBoost. The figure visualizes the modes that had the best parameters. Although, both baseline XGBoost and Random Forest posted higher F1‑scores overall, suggesting they provide a more balanced trade‑off between precision and recall.

Conclusion

Integrating ADASYN with XGBoost was able to achieve the highest recall to be able to detect early stages of cervical cancer. Random forest was able to target positive cancer patients by minimising false positives. Based on (Mehmood et al., 2021) was able to show the importance of hyper parameters to improve diagnosis and early detection. To build on the resarch further, future researchers could focus on optimizing both ADASYN and XGBoost to investigate and improve both recall and precall. It could enhance model robustness and scalability in detecting cervical cancer.

Jetwyn Wilson
CSC 5800

Reference

Mahto, R., & Sood, K. (2024). Predicting Cervical Cancer Based on Behavioral Risk Factors. IJACSA) International Journal of Advanced Computer Science and Applications, 15(11). https://thesai.org/Downloads/Volume15No11/Paper_1-Predicting_Cervical_Cancer_Based_on_Behavioral_Risk_Factors.pdf

UCI Machine Learning Repository. (n.d.). Archive.ics.uci.edu. https://archive.ics.uci.edu/dataset/383/cervical+cancer+risk+factors

Tanimu, J. J., Hamada, M., Hassan, M., Kakudi, H., & Abiodun, J. O. (2022). A Machine Learning Method for Classification of Cervical Cancer. Electronics, 11(3), 463. https://doi.org/10.3390/electronics11030463

Akter, L., Ferdib-Al-Islam, Islam, Md. M., Al-Rakhami, M. S., & Haque, Md. R. (2021). Prediction of Cervical Cancer from Behavior Risk Using Machine Learning Techniques. SN Computer Science, 2(3). https://doi.org/10.1007/s42979-021-00551-6

Mehmood, M., Rizwan, M., Gregus ml, M., & Abbas, S. (2021). Machine Learning Assisted Cervical Cancer Detection. Frontiers in Public Health, 9, 788376. https://pubmed.ncbi.nlm.nih.gov/35004588/