Predicting Cervical Cancer Based on Behavioral Risk Factors

Rakeshkumar Mahto[1], Kanika Sood[2]
Department of Electrical and Computer Engineering, California State University, Fullerton, California 92831, USA[1]
Department of Computer Science, California State University, Fullerton, California 92831, USA[2]

***Abstract*—Machine learning (ML) based predictive models are increasingly used in various fields due to their ability to find patterns and interpret complex relationships between variables in an extensive dataset. However, getting a comprehensive dataset is challenging in the field of medicine for rare or emerging infections. Therefore, developing a robust methodology and selecting ML classifiers that can still make compelling predictions even with smaller and imbalanced datasets is essential to defend against emerging threat or infections. This paper uses behavioral risk factors to predict cervical cancer risk. To create a robust technique, we intentionally selected a smaller imbalanced dataset and applied Adaptive Synthetic (ADASYN) sampling and hyperparameter tuning to enhance the predictive performance. In this work, hyperparameter tuning, evaluated through 3-fold cross-validation, is employed to optimize the performance of the Random Forest, XGBoost, and Voting Classifier models. The results demonstrated high classification performance, with all models achieving an accuracy of 97.12%. Confusion matrix analysis further revealed the models' robustness in identifying cervical cancer cases with minimal misclassification. A comparison with previous work confirmed the superiority of our approach, showcasing improved accuracy and precision. This study demonstrates the potential of ML models for early screening and risk assessment, even when working with limited datasets.**

***Keywords*—Cervical cancer; random forest; voting classifier; Adaptive Synthetic Sampling (ADASYN); predictive model**

I. Introduction

According to WHO, cervical cancer is the fourth most prevalent cancer in women across the globe [1]. The same report highlights that 94% of fatalities happening due to cervical cancer are in low and middle-income countries [1]. However, according to [2], globally, only 36% of women aged 30-49 have been screened for cervical cancer. This is due to a host of social determinants, including socioeconomic status, access to care, and behavioral risks, which make detecting and preventing these types of cancer earlier extremely challenging. Cervical cancer is primarily caused by prolonged infection with high-risk human papillomavirus (HPV) types, and it can be prevented through early-stage screening and vaccination. Various well-established screening methods exist, including Pap smears [3], [4], [5] and HPV DNA [6], [7], [8] testing. Unfortunately, due to the high cost of diagnosis, lack of infrastructure, and awareness in underdeveloped countries, these techniques failed to make an impact. With these limitations in mind, researchers and healthcare professionals are now examining the implementation of predictive models to help identify high-risk individuals for developing cervical cancer. This area of interest is particularly relevant when it comes to using behavioral and social risk factors known to be associated with cervical cancer occurrence (e.g., sexual activity, diet, and personal/intimate hygiene). Integrating these factors in prognostic models efficiently improves current screening techniques that lead to early detection and provide personalized care.

Several research studies have identified various factors that contribute to the development of cervical cancer. Kadir et al. in [9], found that sexual behaviors such as early sexual activity and multiple sexual partners, poor diet, inadequate personal hygiene, and lack of social support are key risk factors that result in cervical cancers. Additionally, a woman's behavioral and mental state can affect their ability to participate in regular screening, follow preventive measures that include HPV vaccination, and adhere to the treatment plan. Thus, having a predictive model equipped with such knowledge will enhance its ability to identify women at higher risk. These kinds of models can also contribute to better allocation of resources by prioritizing those individuals who might otherwise not seek care.

In recent times, machine learning (ML) has increasingly been integrated into the medical diagnostic arena due to its ability to analyze large amounts of data, identify trends, and make accurate predictions. ML models have been utilized to diagnose cervical cancer using behavioral, demographic, and clinical data [10]. In this research work [10], multiple ML models that include Random Forest, AdaBoost, Gradient Boost, MultiLayer Perceptron (MLP), eXtreme Gradient Boosting (XGB), Decision Tree, Logistic Regression, SVM, and Gaussian Naive Bayes (GNB) was trained on a dataset consisting of an individual's demographic data, medical history, sexual behavior, and reproductive health to predict early prediction of cervical cancer. Among all of them, XGB Classifier showed superior accuracy of 98% and ROC AUC of 99%. Similarly, in [11], various ML classifiers such as K-Nearest Neighbors (KNN), Linear Support Vector Machine (SVM), and Naive Bayes were trained using a dataset consisting of patients' demography, biopsy, and medical history. KNN outperformed in all metrics compared to the other two classifiers with an impressive accuracy of 97.59%.

Besides applying ML classifiers and getting trained on individuals' demography, biopsy, and medical history, Pap sear images, HPV DNA test results, and biopsy samples are also utilized for diagnosing Cervical cancer. For example, in the research work presented by Sholik et al., they applied a process that combines advanced methods from the neural network, convolutional neural network (CNN), and vision transformers to capture both detailed and overall patterns in images [12]. The dataset utilized for this work were Herlev [13], Mendeley LBC [14], and SIPaKMeD [15], which consists

of Pap smear images. They achieved an impressive accuracy of 100% with SVM, K-NN, MLP, and Logistic Regression (LR). Similarly, advanced deep learning models such as ResNet and GoogLeNet were employed to classify cervical cancer using Pap smear images [16]. Using fine-tuned ResNet18, a test accuracy of 98.51% was obtained. On the other hand, HPV viral load with bacterial vaginosis status was utilized to train an LR model for early diagnosis of cervical cancer [17]. The model achieved an impressive accuracy with AUC values ranging from 0.915 to 0.9614. However, the effectiveness and accuracy of these ML classifiers depend on the size and distribution of the dataset.

Class imbalance is one of the biggest problems while building predictive models for cervical cancer. The count of patients who are suffering from Cervical cancer is significantly less in real-world datasets as compared to the ones that do not suffer from this disease, which leads to an imbalanced dataset. This class imbalance of "cervical cancer" to "no cervical cancer" can significantly influence the performance of ML models, especially in terms of not being able to classify the minority group. In ML, specifically classification problems, the models tend to favor the majority class in case of imbalanced data, which results in higher overall accuracy but low sensitivity (i.e. the inability of the model to identify positive cases). In healthcare applications, this can be particularly problematic where the model fails to correctly identify individuals at risk of a fatal ailment like cervical cancer, which could lead to missed early intervention and prevention.

Besides class imbalance, the other issue in developing a robust ML model is the challenges in collecting comprehensive datasets, especially in resource-limited settings or the occurrence of rare and emerging infectious diseases. In this paper, we intentionally chose a smaller and imbalanced dataset [18] to demonstrate that advanced ML techniques can achieve high accuracy and precision. The success of the proposed model in achieving this goal will showcase the robustness and effectiveness of ML in predicting cervical cancer risk, even with a small and imbalanced dataset.

Additionally, in this paper, we analyze the impact of ML in the development of an efficient prediction model for cervical cancer using behavioral risk factors. In this work, we use Adaptive Synthetic Sampling (ADASYN) [19] to address the problem of class imbalance and smaller datasets to achieve better performance for distinguishing women who are at high risk of developing cervical cancer. The present study also intends to determine the potential of various behavioral risk factors significantly associated with cervical cancer by using feature importance analysis. This effort will lead to a greater understanding of the primary behavioral and social factors that may influence risk for cervical cancer, which ultimately results in further improvement in cervical cancer screening and prevention strategies. Additionally, the ML model developed in this work can be replicated and applied to rare diseases or those cases where getting a comprehensive dataset is challenging.

The remainder of this paper is organized as follows: Section II describes relevant work in cervical cancer prediction by applying ML techniques to behavior risk factors. This is followed by Section III, presenting the methodology utilized in this paper. This includes analyzing and describing the dataset used for training the various ML models, including Random Forest, XGBoost, and Voting Classifier. Results and discussion are presented in Sections IV and V, respectively, where the model's performance will be evaluated through various metrics like accuracy, precision, recall, and F1-score. Finally, Section VI concludes the paper where the implications of these findings are presented.

II. Relevant Work

ML models show great promise in diagnosing cervical cancer by leveraging behavioral risk factors. In cancer prediction, and especially for cervical cancer, behavioral risk factors play a more significant role compared to other datasets like clinical or genetic data. Behavioral information can be collected non-intrusively through surveys, interviews, or questionnaires, which makes it easier to collect, especially in resource-limited contexts where clinical tests such as Pap smears or genetic screening are scarce or too expensive for widespread use. Including behavioral information as input features for ML models can provide a more holistic view of all likely risk factors that may result in earlier detection and intervention. Various studies have been conducted to improve early detection using ML, which is significant for the treatment and increasing survival rates. For example, using the UCI machine learning repository with 32 features [20], Mehmood et al. achieved an accuracy of 93.6% with the Random Forest technique [21]. On the same dataset [20], Suman et al. attained an accuracy of 96.38% using BayeNet algorithm [22]. In another work on the same dataset [20], SMOTE and Random Forest were applied, yielding an accuracy of 85% [23]. Sun et al. developed a unique stacking-integrated machine learning (SIML) using the same dataset from UCI [20]. The SIML model combined multiple algorithms that included TreeBag, XGBoost, and MonMLP to achieve an AUC of 0.877, sensitivity of 81.8%, and specificity of 81.9%. Though various studies were conducted utilizing ML models to predict cervical cancer risk as presented in [20], [21], [22], [23], most have focused on demographic or clinical data, where behavioral factors were ignored entirely.

In another research work, Akter et al. applied the Decision Tree, Random Forest, and XGBoost on a different dataset at UCI machine learning repository dataset [18] with 19 attributes regarding behavior risk that can lead to cervical cancer [24]. Their ML model was able to achieve an accuracy of 93.33%. On the same dataset, various ML classifiers were applied that included Gaussian Naive Bayes (GNB), k Nearest Neighbor, and Decision Tree. Among all of them, GNB demonstrated a superior performance of 94% [25]. While the accuracy of ML models in predicting cervical cancer was impressive, most of these studies employ large datasets and balanced datasets, which do not reflect the real-world challenges of data scarcity and class imbalance, which are more prevalent in healthcare settings.

In most developing countries where the majority of the sufferers of cervical cancers reside, getting large and granular datasets is challenging due to inadequate health infrastructure, limitations of resources, and financial constraints. Additionally, cultural and logistical barriers become inhibiting factors for participant to share their information, making it challenging to create an extensive dataset. Given these constraints, it is vital to develop ML models that can perform well with a smaller

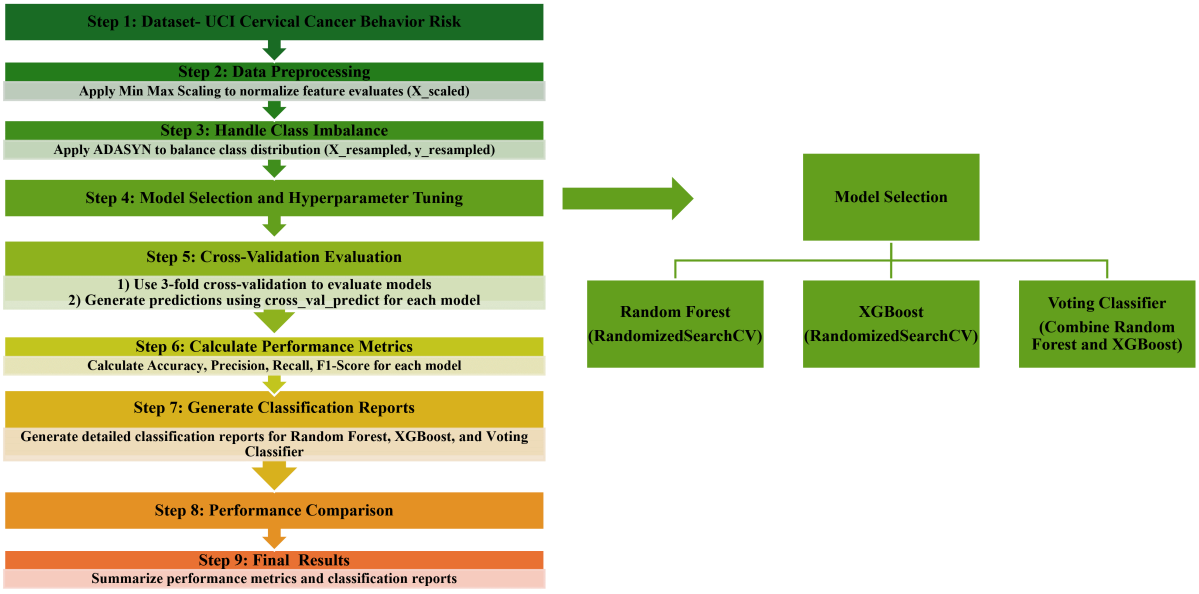


Fig. 1. Methodology flow diagram for developing an efficient prediction model for cervical cancer using behavioral risk factors.

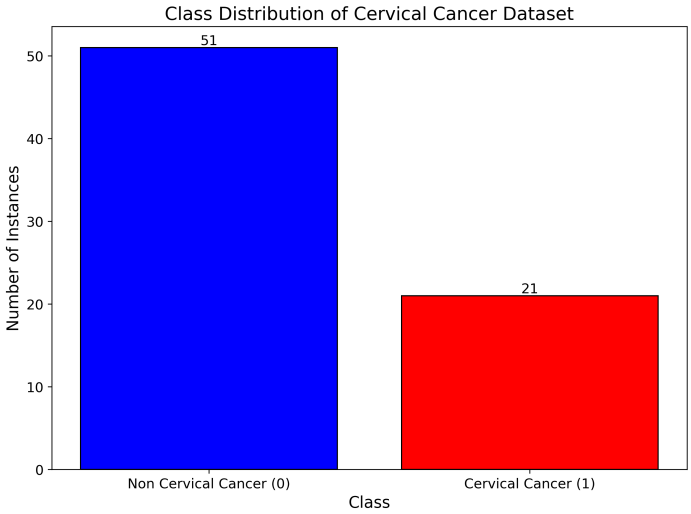


Fig. 2. Class distribution of the cervical cancer dataset.

number of data points and fewer attributes. Optimizing models to achieve high accuracy using few attributes will ensure that such models will remain useful when a comprehensive dataset is unavailable. These models are scalable, and their predictive performance can be further improved when trained on larger datasets.

III. Methodology

The methodology adopted for this work involves a comprehensive set of steps, including data preprocessing, handling class imbalance, hyperparameter tuning, model training, and evaluation using cross-validation techniques, as shown in Fig. 1.

A. Dataset Description

The dataset used in this study was obtained from the UCI Machine Learning Repository [18]. The dataset includes 18 features related to behavioral risk factors for cervical cancer, as shown in Table. I and a total of 72 datapoints. The target variable is ca_cervix, where a value of '0' represents non-cervical cancer, and '1' represents cervical cancer. As shown in Fig. 2, the target variable consists of 51 cases of non-cervical cancer, and the rest 21 cases of patients with cervical cancer. This predominance of non-cervical cancer cases over any type of cervical cancer case can produce prediction biases toward the majority class. Due to the dataset's imbalance, the trained model might incorrectly classify cervical cancer patients as non-cervical cases. Hence, it is essential to have a data balancing technique prior to training the ML models.

The small size and imbalance of the dataset were intentionally chosen to test the model's ability to handle real-world constraints where comprehensive data collection is not always possible. The features in this dataset correspond to various behavioral aspects like personal hygiene, eating habits, social support systems, and many more, as shown in Table I. These variables are all of integer type and suitable for many ML models which support categorical data as input. For example, the behavior_eating and behavior_personalHygiene features reflect personal lifestyle choices, while the socialSupport_emotionality, socialSupport_appreciation, and socialSupport_instrumental features quantify different aspects of social support.

A correlation plot is presented in Fig. 3 shows a correlation between various features in the dataset. This plot highlights the directions and strength of correlation of features in the dataset towards the target variable which is ca_cervix. This information is valuable in selecting features for the ML training and testing that results in improving the prediction capability of the model.

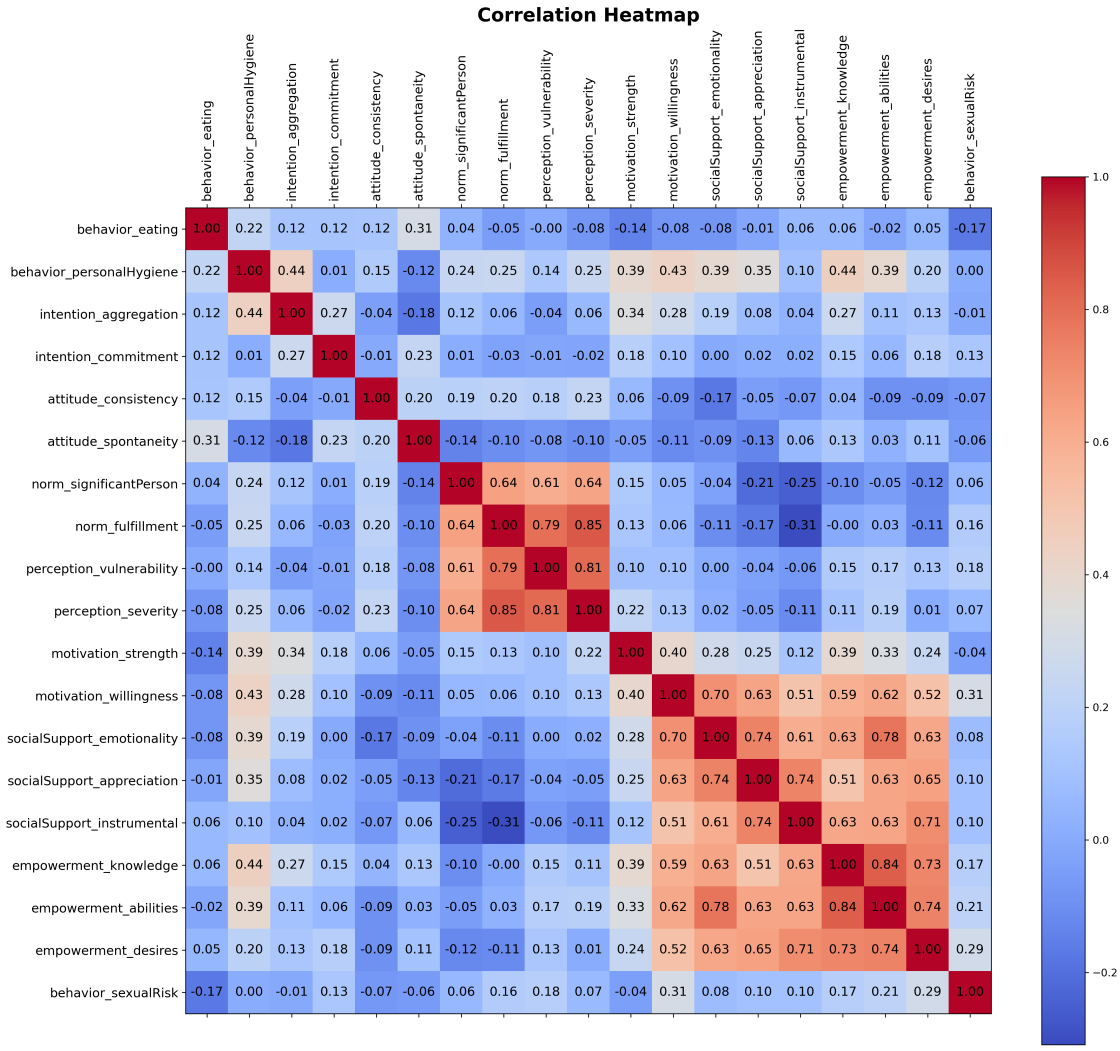


Fig. 3. Correlation heatmap of behavioral risk factors for cervical cancer.

B. Data Preprocessing

The dataset used in this study consists of integer features that describes some measures regarding several behavioral risk factors. Most features will have different ranges and scales (for instance, motivation_strength has values ranges between 3 and 15 while socialSupport_appreciation is between 2 and 10). Some of the ML classifiers such as Random Forest and XGBoost are not affected by feature scaling due to tree-based structure. However, that's not the case with model that rely on distance, such as k-Nearest Neighbors (kNN) or Support Vector Machines (SVM).

To standardize the feature values and ensure a more uniform input, Min-Max Scaling was applied. This scaling will transform the feature to values ranging between 0 and 1. The formula for Min-Max Scaling is:

$$X_{\text{scaled}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{1}$$

Where, X_{scaled} is the normalized value, X_{max} and X_{min} represent the maximum and minimum values of the feature, respectively.

C. Handling Class Imbalance

Medical datasets are prone to imbalances where most data focuses on healthy individuals rather than actual patients, which is the case with the dataset utilized in this paper, as shown in Fig. 2. This imbalance issue can lead to low sensitivity (or recall) for the minority class, an important metric in medical tasks like disease detection where false negatives come with severe consequences.

To address this, prior to scaling steps, ADASYN is applied on the dataset. The algorithm in ADASYN generates synthetic data for each minority class by adapting the number of synthetic instances to the local density of the minority class. It first identifies the instances in the minority class and their neighbors, and then generates more synthetic examples for those minority instances that are harder to classify. This process ensures, that dataset is more balanced and does not have duplicate data. Hence, the use of the ADASYN technique

results in reducing model bias, improving sensitivity toward minority classes, and preventing overfitting.

D. Model Selection and Hyperparameter Tuning

To create efficient predictive models for early cervical cancer, we selected three different ML classifiers with complementary strengths: Random Forest, XGBoost, and a Voting Classifier. To further enhance the performance of the selected machine learning models, we conducted hyperparameter tuning, which will be discussed in more detail in this section.

1) Model selection:

a) Random forest: Random Forest is an ensemble learning technique where predictions are made by combining many decision trees. This technique of combining the responses of several decision trees instead of relying on one results in improved accuracy and reliability. The final prediction is made based on the majority of votes from different trees in the forest. The final prediction for the final predicted class is given mathematically by [26]:

$$\hat{y} = \text{argmax}\left(\sum_{i=1}^{N} I\left(T_i(x) = j\right)\right), \quad j \in \{0, 1\} \tag{2}$$

Where $\hat{y}$ is the final predicted class, N is the number of trees in the Random Forest. The prediction for i-th tress from x number of input instances is given by $T_i(x)$. I($T_i(x)$) is an indicator function equal to 1 if the prediction $T_i(x)$ matches class j and 0 otherwise. In Eq. (2), j represents the class that received the maximum votes from all the trees. This technique is popular because it prevents overfitting of training data due to averaging of predictions of various trees.

b) XGBoost classifier: Extreme Gradient Boosting (XGBoost) is based on gradient boosting, an ensemble technique where decisions of weak learner decision trees are combined to create an efficient learner. In gradient boosting, trees are added consecutively such that each new tree rectifies the error made by previously added trees. The overall prediction in XGBoost, $\hat{y}$ is given by [27]:

$$\hat{y} = \sum_{m=1}^{M} f_m(x) \tag{3}$$

TABLE I. GROUP BY THEME TABLE FOR CERVICAL CANCER DATASET

Theme	Category	Attributes
Psychological	**Behavior**	behavior_personalHygiene, behavior_eating, behavior_sexualRisk
	Intention	intention_commitment, intention_aggregation
	Attitude	attitude_spontaneity, attitude_consistency
Social	**Norm**	norm_fulfillment, norm_significantPerson
	Social Support	socialSupport_emotionality, socialSupport_appreciation, socialSupport_instrumental
Perceptual & Motivational	**Perception**	perception_severity, perception_vulnerability
	Motivation	motivation_willingness, motivation_strength
Empowerment	**Empowerment**	empowerment_knowledge, empowerment_abilities, empowerment_desires

Where M is the total number of trees, $f_k(x)$ is the prediction made from the m-th tree for x input instance. This ML algorithm is known for its high performance and efficiency in supervised learning, especially for regression and classification-related applications. It is popular in various applications due to its ability to handle large datasets and improve prediction accuracy.

c) Voting classifier: Like the previous two classifiers, the Voting Classifier is an ensemble learning technique that combines multiple classifiers to improve the overall classification accuracy. By aggregating the output of various other classifiers, the Voting Classifier enhances the robustness of the prediction and mitigates the shortcomings of a single model. Typically, there are two types of voting methods used in Voting classifiers:

1) Hard Voting: The final prediction is made through a majority vote among the classifier.
2) Soft Voting: The final prediction is made by averaging all the predictions from the classifiers. This led to a balanced and nuanced decision.

The final prediction $\hat{y}$ Voting Classifier is given by [28]:

$$\hat{y} = \text{argmax}\left(\sum_{i=1}^{N} I\left(y_i = j\right)\right), \quad j \in \{0, 1\} \tag{4}$$

Where $\hat{y}$ is the final prediction, N is the number of classifiers, y_i is the prediction of the i-th classifier, and $I(y_i = j)$ shows that the prediction belongs to one of the classes in the target. For example, it gives out a result of 1 if the prediction y_i matches class j, this conveys that it belongs to this class. Otherwise, the results give out 0, which means it does not belong to this class.

In this work, using the Voting Classifier, the strengths of both algorithms can be combined, leading to better prediction. For example, Random Forest is good at handling noise and variability in the data. On the other hand, XGBoost is known for its efficiency and ability to improve accuracy. The Voting Classifier will aggregate the predictions of Random Forest and XGBoost. Hence, the errors arising from one of the techniques can then be avoidable through voting compared to a single model. Therefore, in this work, the Voting Classifier was chosen in addition to Random Forest and XGBoost.

2) Hyperparameter tuning: RandomizedSearchCV is used to improve the performance of the ML model further. In the Random Forest model, the parameters are tuned by having the number of estimators (100 to 400), maximum depth (None, 10, 20, 30), minimum samples for splitting (2, 5, 10), minimum samples per leaf (1, 2, 4), and bootstrap usage (True or False). Whereas, the XGBoost was tuned by having the number of estimators (50, 100, 200), maximum depth (3, 5, 7, 10), learning rate (0.01 to 0.3), subsampling ratio (0.6, 0.8, 1.0), and column sampling by tree (0.6, 0.8, 1.0).

E. Cross-Validation

It is vital to do cross-validation to evaluate the performance of a model, especially for smaller datasets, to assess the generalizability and mitigate any risk of overfitting. Typically,

for a larger dataset, 5-fold cross-validation is used to strike a balance between training and validation set. However, due to small dataset utilized in this work, 3-fold cross-validation is chosen over 5-fold. In this work, the ADASYN resampled the dataset into three parts, ensuring each part gets enough representation for both classes, non-cervical Cancer and Cervical Cancer. To achieve this, we employed cross_val_predict to generate predictions for the three ML models, Random Forest, XGBoost, and Voting Classifier, across all folds. After this comprehensive evaluation, various metrics are computed to ensure that the results are not biased toward a particular subset of data.

F. Performance Metrics

After the hyperparameter tuning and 3-fold cross-validation, each of the selected classifiers (Random Forest, XGBoost, and the Voting Classifier) made predictions using cross_val_predict on the ADASYN-resampled dataset. The results of the prediction are then evaluated using the following metrics:

a) Accuracy: It measures the total correct prediction with respect to the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

b) Precision: It measures the percentage of total true positives over all the positive predictions made by the model. Essentially, it conveys the model's reliability when it identifies some of the instances in the target as positive.

$$\text{Precision} = \frac{TP}{TP + TN} \tag{6}$$

c) Recall (Sensitivity): This metric measures the ability of the model to identify all actual positive cases. Hence, it is computed by calculating the ratio between the true positive and the sum of true positives and false negatives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

d) F1-Score: It is the measure where precision and recall are combined into a single metric to provide balanced view of a model's performance hence it is called a harmonic mean between prediction and recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

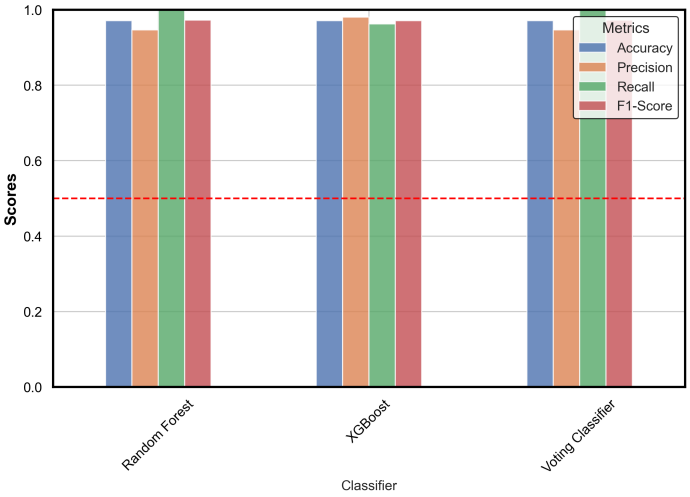


Fig. 4. Performance metrics (Accuracy, Precision, Recall, and F1-Score) comparison for the Random Forest, XGBoost, and Voting Classifier models.

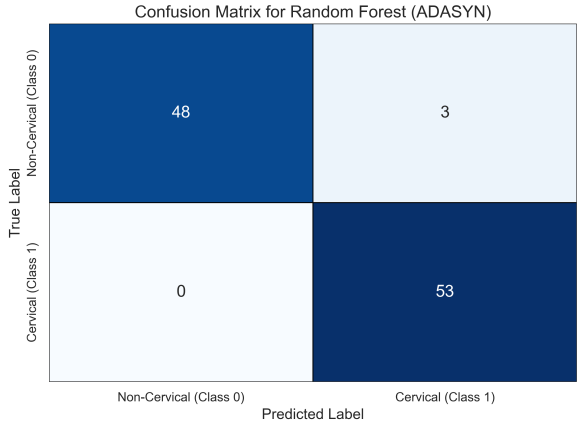
IV. Results

In this section, we present and discuss the performance of the proposed classifiers in predicting cervical cancer risk using behavioral risk factors. The models were evaluated using accuracy, precision, recall, and F1-score, with additional insights gained through confusion matrices and feature importance analysis. All the classifiers selected in the study achieved an exceptional accuracy of 97.12%.

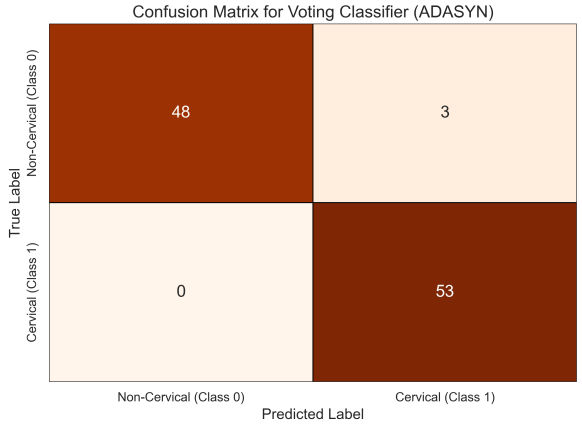
A. Performance Metrics

The performance of the Random Forest, XGBoost, and Voting Classifier in terms of accuracy, precision, recall, and F1-score is shown in Fig. 4. Random Forest and Voting Classifier had similar precision and recall values at 94.64% and 100%, respectively. XGBoost, on the other hand, had slightly higher precision (98.08%) but a little lower recall (96.23%). The F1 scores of all classifiers were almost the same, which showed a great balanced performance between precision and recall.

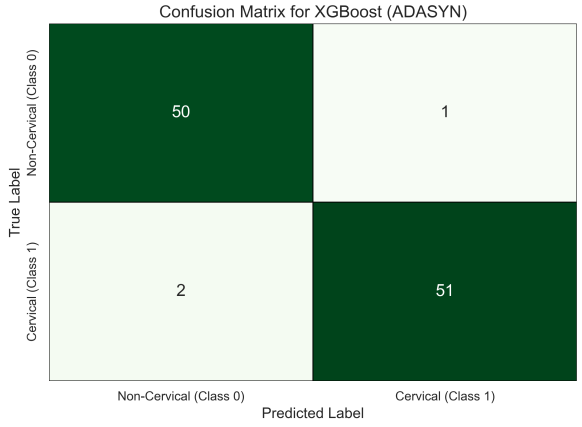
These results show that all three classifiers are very effective in predicting cervical cancer risk. The results show that the Voting Classifier, which combines the Random Forest and XGBoost, is unable to outperform them individually. Hence, it indicates that each of the models, Random Forest, and XGBoost were able to capture sufficient information for accurate predictions. Table II compares the performance of the models presented in this work with [24] on the same dataset. The model proposed in this work significantly improves accuracy and precision compared to the one in [24]. Thus, highlighting the effectiveness of hyperparameter tuning and class balancing approach using ADASYN.



(a)



(b)



(c)

Fig. 5. Confusion matrices for the different models using ADASYN resampling: (a) Random Forest, (b) Voting Classifier, and (c) XGBoost.

B. Confusion Matrix Analysis

Fig. 5a, Fig. 5b, and Fig. 5c shows the confusion matrix for Random Forest, Voting Classifier, and XGBoost, respectively. The confusion matrix shown in Fig. 5a, and Fig. 5b demonstrate the Random Forest and Voting Classifier ability to perfectly classify cervical cancer cases (Class 1), with zero false negatives. This demonstrating their ability to identify all positive instances. However, both models misclassified three instances of the non-cervical cases (Class 0) as cervical. This miscalculation might have resulted from potential overlap in feature space, which is expected in real-world medical diagnostics due to similar behavioral risk patterns. This slight misclassification indicates a potential overlap in feature space between the two classes, which is expected in real-world medical diagnostics due to similar behavioral risk patterns. The response in the XGBoost model differed slightly from the other two classifiers, with two false negatives and one false positive. Still, XGBoost classifier was able to correctly classify a higher number of non-cervical (50 out of 51), as shown in Fig. 5c.

TABLE II. COMPARISON OF PROPOSED MODELS WITH [24]

Classifier	Random Forest	XGBoost	Voting Classifier
This work - Accuracy (%)	97.12	97.12	97.12
This work - Precision (%)	94.64	98.08	94.64
This work - Recall (%)	100.00	96.23	100.00
This work - F1-Score (%)	97.25	97.14	97.25
[24] - Accuracy (%)	93.33	93.33	-
[24] - Precision (%)	92	93	-
[24] - Recall (%)	100	100	-
[24] - F1-Score (%)	96	97	-

C. Feature Importance Analysis

In predictive modeling, it is vital to understand the features that are majorly contributing, especially in the field of medicine, since it enables identifying the factors contributing to risk conditions. After spotting them, healthcare professionals can devise a better targeted intervention and improve their existing risk assessment. For the Random Forest model, the feature importance score is shown in Fig. 6. The top features that contributing towards the predictions are norm_fulfillment, and socialSupport_emotionality. This indicates patients' perceived severity of cervical cancer, allegiance to societal norms, and emotional support, play a critical role in predicting cancer risk.

Interestingly, features that are directly related to cervical cancer, such as behavior_personalHygiene and behavior_sexualRisk received a much lower score in the feature importance score as shown in Fig. 6. Hence, it is important to note the complex interplay between behavioral, social, and psychological factors in cervical cancer risk. Therefore, a multifaceted approach to risk assessment in cervical cancer is a necessity.

D. Comparison with Previous Work

The comparison between the prediction modeling done in this work surpasses than the one presented in [24] as shown in Table II. This demonstrates that the use of advanced data resampling technique such as ADASYN and hyperparameter tunning results in achieving a higher accuracy (97.12% vs 93.3%) and improved precision. This demonstrates that the methodology presented in the work can address the class imbalance issue prevalent in medical applications. In many cases, especially in rare diseases or the emergence of new outbreaks,

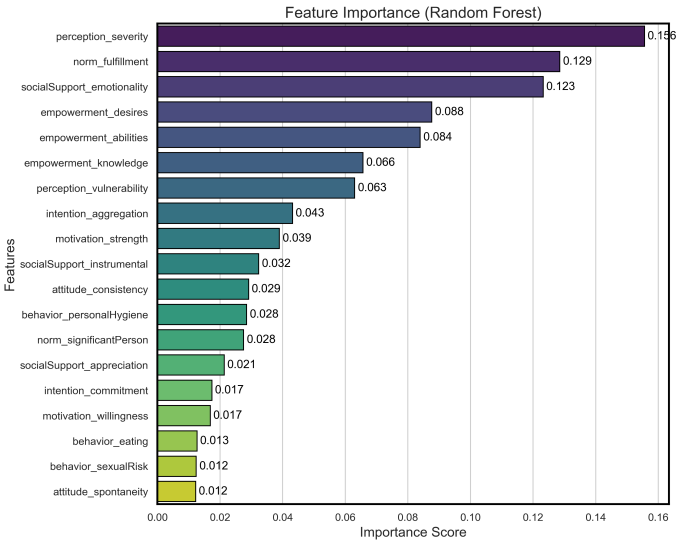


Fig. 6. Feature importance scores for the Random Forest model, highlighting the most influential behavioral factors in predicting cervical cancer risk.

getting a comprehensive and dataset is highly challenging. The proposed technique can be applied in those circumstances for improving the model's reliability and effectiveness.

V. Discussion

A. Comparative Results on Multiple Datasets

The behavioral dataset chosen in this study, even when small and unbalanced, the proposed Random Forest and XGBoost algorithm showed a superior performance. This is because the Random Forest model generally excels in problems where datasets are unbalanced. Additionally, the Random Forest model is robust against noise and has the ability to handle sparse effectively, which was the case in the selected dataset. Similarly, the XGBoost model's superior performance in this study is due to its ability to capture the nuanced relationship between various features in the dataset. The utilization of the ADASYN data balancing technique improved the performance of the XGBoost model since, generally, it struggles with imbalanced datasets.

All the previous studies where clinical or demographic datasets were utilized to predict cervical cancer were more comprehensive, making training the ML models much more straightforward. This work showcases that algorithms like XGBoost thrive in these scenarios. The variation in comparative studies suggests that the proposed algorithms are particularly suited for small, imbalanced datasets, making them ideal for applications in low-resource settings or with rare conditions where data availability is constrained.

B. Suitability of Proposed Algorithms

Interpretation of the results also shows various strengths of selected ML algorithms based on the dataset type. Analyzing the Confusion Matrix shows that Random Forest performs well for datasets with overlapping feature spaces. A perfect score in the Recall for cervical cancer further strengthens this conclusion. On the other hand, the XGBoost model has superior precision, which shows its capability to reduce false positives. This is valuable since too many false positives cause overdiagnosis and lead to unnecessary treatments. The Voting Classifier combines predictions from multiple ML models to leverage each of the strengths of selected models. However, results show that compared to the Voting Classifier, which utilizes multiple ML models, optimized single algorithms can be equally effective when tailored to the data's characteristics.

C. Implications for Healthcare

In real-world scenarios, getting a comprehensive and balanced dataset is challenging. Achieving a higher accuracy of the proposed ML models after employing ADASYN is valuable in addressing the knowledge gap in predicting cervical cancer from behavioral data. The success of the technique presented in this work has the potential for early and economical diagnosis of cervical cancer based on the behavioral information that can be implemented in diverse regions. Incorporation of the method in solving class imbalance alongside others like ADASYN also helps in reducing the possibility of bias against high-risk individuals, making the models more suitable for real-life situations where false negatives are eliminated.

D. Strengths, Limitations and Future Directions

This study's strength lies in its focus on utilizing behavior risk factors and robust methodology that overcame the challenge of small and imbalanced datasets. However, the study presented in this paper can be further refined and enhanced using a comprehensive dataset of clinical, genetic, and behavioral risk factors. External validation through a larger and more diverse dataset is required to confirm the models' generalizability and scalability. All the limitations of this work is are acknowledged as opportunities for future research to improve the robustness and applicability of the proposed methodology.

VI. Conclusion

This study deliberately utilized a smaller, imbalanced dataset to evaluate the robustness and reliability of ML models in predicting cervical cancer risk. Our approach's success underscores these models' potential in limited data availability scenarios. Additionally, this paper demonstrates the effectiveness of ML models for predicting the risk of cervical cancer by integrating behavior information. Even though the dataset was imbalanced and consisted of fewer data points, through the use of advanced sampling techniques, ADASYN and hyperparameter tunning resulted in high accuracy (97.12%), Precision (94.64%), Recall (100.0%), and F1-score (97.25%) for Random Forest. The confusion matrix analysis validated our model's reliability. Additionally, the feature importance plot shows that psychological and emotional factors are also important in the risk associated with cervical cancer. Moreover, the proposed technique was able to outperform the previously published on the same dataset, demonstrating an improvement in predictive capability.

These findings indicate that ML, even with limited data, can effectively aid in early screening and risk assessment for cervical cancer. Future research should explore integrating

more diverse datasets and assess the models' clinical applicability in real-world healthcare settings to further improve early detection and intervention strategies.

REFERENCES

[1] M. Kyrgiou, A. Athanasiou, M. Arbyn, S. F. Lax, M. R. Raspollini, P. Nieminen, X. Carcopino, J. Bornstein, M. Gultekin, and E. Paraskevaidis, "Terminology for cone dimensions after local conservative treatment for cervical intraepithelial neoplasia and early invasive cervical cancer: 2022 consensus recommendations from esgo, efc, ifcpc, and esp," *The Lancet Oncology*, vol. 23, no. 8, pp. e385–e392, 2022.

[2] L. Bruni, B. Serrano, E. Roura, L. Alemany, M. Cowan, R. Herrero, M. Poljak, R. Murillo, N. Broutet, L. M. Riley, and S. de Sanjose, "Cervical cancer screening programmes and age-specific coverage estimates for 202 countries and territories worldwide: a review and synthetic analysis," *The Lancet Global Health*, vol. 10, no. 8, pp. e1115–e1127, 2022.

[3] W. William, A. Ware, A. H. Basaza-Ejiri, *et al.*, "A pap-smear analysis tool (pat) for detection of cervical cancer from pap-smear images," *BioMed Eng OnLine*, vol. 18, p. 16, 2019.

[4] S. Pankaj, S. Nazneen, S. Kumari, A. Kumari, A. Kumari, J. Kumari, V. Choudhary, and S. Kumar, "Comparison of conventional pap smear and liquid-based cytology: A study of cervical cancer screening at a tertiary care center in bihar," *Indian Journal of Cancer*, vol. 55, pp. 80–83, Jan–Mar 2018.

[5] M. Patel, A. Pandya, and J. Modi, "Cervical pap smear study and its utility in cancer screening, to specify the strategy for cervical cancer control," *National journal of community medicine*, vol. 2, no. 01, pp. 49–51, 2011.

[6] P. Mahmoodi, M. Fani, M. Rezayi, A. Avan, Z. Pasdar, E. Karimi, I. S. Amiri, and M. Ghayour-Mobarhan, "Early detection of cervical cancer based on high-risk hpv dna-based genosensors: A systematic review," *Biofactors*, vol. 45, pp. 101–117, Mar 2019.

[7] N. Bhatla and S. Singhal, "Primary hpv screening for cervical cancer," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 65, pp. 98–108, 2020.

[8] G. Ronco and P. G. Rossi, "Role of hpv dna testing in modern gynaecological practice," *Best Practice & Research Clinical Obstetrics & Gynaecology*, vol. 47, pp. 107–118, Feb 2018.

[9] A. Kadir, S. Shenoda, and J. Goldhagen, "Effects of armed conflict on child health and development: a systematic review," *PLOS ONE*, vol. 14, p. e0210071, Jan 2019.

[10] M. Hasan, J. Islam, M. A. Mamun, A. A. Mim, S. Sultana, and M. S. H. Sabuj, "Optimizing cervical cancer prediction, harnessing the power of machine learning for early diagnosis," in *2024 IEEE World AI IoT Congress (AIIoT)*, pp. 552–556, 2024.

[11] A. H. Elmi, A. Abdullahi, and M. Ali Bare, "A comparative analysis of cervical cancer diagnosis using machine learning techniques," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, p. 1010, May 1 2024.

[12] M. Sholik, C. Fatichah, and B. Amaliah, "Deep feature extraction of pap smear images based on convolutional neural network and vision transformer for cervical cancer classification," in *2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, (BALI, Indonesia), pp. 290–296, 2024.

[13] J. Jantzen, J. Norup, G. Dounias, and B. Bjerregaard, "Pap-smear benchmark data for pattern classification," in *Nature Inspired Smart Information Systems (NiSIS)*, vol. 30, pp. 1–9, 2005.

[14] E. Hussain, L. B. Mahanta, H. Borah, and C. R. Das, "Liquid based-cytology pap smear dataset for automated multi-class diagnosis of pre-cancerous and cervical cancer lesions," *Data in Brief*, vol. 30, p. 105589, Jun. 2020.

[15] M. E. Plissiti, P. Dimitrakopoulos, G. Sfikas, C. Nikou, O. Krikoni, and A. Charchanti, "Sipakmed: A new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 3144–3148, Oct. 2018.

[16] A. Goswami, N. G. Goswami, and N. Sampathila, "Deep learning-based classification of cervical cancer using pap smear images," in *2024 4th International Conference on Intelligent Technologies (CONIT)*, (Bangalore, India), pp. 1–6, 2024.

[17] B. Meng, G. Li, Z. Zeng, *et al.*, "Establishment of early diagnosis models for cervical precancerous lesions using large-scale cervical cancer screening datasets," *Virology Journal*, vol. 19, no. 177, 2022.

[18] UCI Machine Learning Repository, "Cervical cancer behavior risk." https://doi.org/10.24432/C5402W, 2019.

[19] H. He, Y. Bai, E. A. Garcia, and S. Li, "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, 2008.

[20] K. Fernandes, J. Cardoso, and J. Fernandes, "Cervical cancer (risk factors) [dataset]," 2017.

[21] M. Mehmood, M. Rizwan, M. L. Gregus, and S. Abbas, "Machine learning assisted cervical cancer detection," *Frontiers in Public Health*, vol. 9, p. 788376, 2021.

[22] S. K. Suman and N. Hooda, "Predicting risk of cervical cancer: A case study of machine learning," *Journal of Statistics and Management Systems*, vol. 22, no. 4, pp. 689–696, 2019.

[23] X. Deng, Y. Luo, and C. Wang, "Analysis of risk factors for cervical cancer based on machine learning methods," in *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pp. 631–635, 2018.

[24] L. Akter, Ferdib-Al-Islam, M. Islam, *et al.*, "Prediction of cervical cancer from behavior risk using machine learning techniques," *SN Computer Science*, vol. 2, p. 177, 2021.

[25] M. Çakır, A. Degirmenci, and O. Karal, "Exploring the behavioural factors of cervical cancer using anova and machine learning techniques," in *Science, Engineering Management and Information Technology* (A. Mirzazadeh, B. Erdebilli, E. Babaee Tirkolaee, G.-W. Weber, and A. K. Kar, eds.), vol. 1808 of *Communications in Computer and Information Science*, Springer, Cham, 2023.

[26] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[27] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), p. 785–794, Association for Computing Machinery, 2016.

[28] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: Chapman and Hall/CRC, 2012.