IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 June 2023, accepted 9 July 2023, date of publication 18 July 2023, date of current version 25 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3296456

RESEARCH ARTICLE

Design and Development of an Efficient Risk Prediction Model for Cervical Cancer

RITHVIK HARIPRASAD, NAVAMANI T M, TEJAS RAVINDRA ROTE, AND ISHITA CHAUHAN
School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology (VIT), Vellore, Tamil Nadu 632 014, India
Corresponding author: Navamani T M (navamani.tm@vit.ac.in)

ABSTRACT Cervical cancer is a major public health concern, especially in low- and middle-income countries. Lifestyle choices to some extent have an effect on causing cervical cancer. Most cervical cancers are caused by the sexually transmitted infection caused by the Human Papillomavirus (HPV). However, only persistent HPV infections lead to progression to pre-cancer and cancer. The persistence of this infection is influenced by many factors namely, age, sexually transmitted infections, number of sexual partners, age at first sexual intercourse, number of deliveries, tobacco consumption, etc. Risk-based prediction algorithms help to stratify women with a high risk to develop cervical cancer and screen them on a priority basis. In this study, a model has been developed to predict the risk of cervical cancer based on one's lifestyle choices. Important features have been delineated using the Extreme Gradient Boosting (XGBoost) Classifier. After oversampling, the data is fed into the model for training and testing. The Gradient Boost model was chosen to arrive at an accurracy of 98.9%. This model can be effective to associate risk factors with cervical cancer prediction which can help the in the effective prevention and management of cervical cancer.

INDEX TERMS Machine learning, digital health, cervical cancer, human papillomavirus, risk factors, predictive modeling.

I. INTRODUCTION

Cervical cancer originates in the cervical cells. The cervix is the slender, lower extremity of the uterus which joins the uterus to the vagina. This cancer is caused by the abnormal growth of cells in the cervix and is typically slow growing, which means that it may take years for symptoms to appear. Cervical cancer symptoms can include aberrant vaginal bleeding, such as bleeding between periods, after intercourse, or after menopause [1]. Women may also experience pelvic pain, pain during sex, or an unusual vaginal discharge.

Cervical cancer is a leading cause of mortality among women worldwide, especially in low-resource settings. Every year in India 123,907 new cases of cervical cancer are diagnosed and 77,348 cases succumbed to this disease accounting for nearly one-fourth of the global mortality due to cervical cancer [2]. Cervical cancer is not only preventable but also is amenable to elimination. Hence World Health Organization (WHO) delivered a call for the elimination of cervical cancer as a public health problem by the year 2030 through extensive

Human Papillomavirus (HPV) vaccination, screening, early diagnosis, and treatment of cervical precancer and cancer [3]. Despite the availability of screening methods and preventive measures, the incidence of cervical cancer remains high, emphasizing the need for a more accurate risk prediction model.

In recent years, risk-based prediction algorithms have emerged as a promising approach to improve the accuracy of cervical cancer screening and diagnosis [4]. These algorithms use various parameters, typically involve collecting clinical and demographic information from women including their age, number of sexual partners, age at the first sexual intercourse, HPV infection status, immune deficiency status like Human Immunodeficiency Virus (HIV) infection, and history of tobacco consumption to estimate an individual's risk of developing cervical cancer [5].

Machine learning techniques have been applied to develop and refine these algorithms, with the goal of achieving

The associate editor coordinating the review of this manuscript and approving it for publication was Jad Nasreddine.

higher sensitivity and specificity in predicting the risk of developing cervical cancer. As more data become available, including cytology results and epigenomic data, risk-based prediction algorithms are expected to become even more accurate, enabling earlier detection and better treatment outcomes for women at high risk of cervical cancer. Algorithms like neural networks, Decision Tree classifiers (DT), and logistic regression are often employed to develop and refine these algorithms [6]. The resulting algorithms can help clinicians to identify high-risk patients who may require further diagnostic testing, such as colposcopy or biopsy, or more frequent screening intervals. Early detection through screening is essential for improving survival rates, and several screening methods, such as Pap tests and HPV testing have been developed. However, these methods may not be accessible to all women, notably in low-income and middle-income nations. Furthermore, women face numerous obstacles in accessing these services [7].

This motivated a need for developing a cervical cancer risk prediction model that provides a percentage risk based on lifestyle factors that is driven by the need to improve early detection, personalize risk assessment, optimize resource allocation, and empower individuals to make informed decisions.

It is true that previous research on cervical cancer risk prediction has primarily concentrated on the creation of more accurate models using machine learning techniques [8], [9]. However, the selection of the objective variable in these models is also crucial for enhancing their clinical credibility. Several earlier studies [10], [11] have utilized pre-existing target variables, such as the Hinselmann model, which incorporates variables such as age, race, sexual behavior, and smoking status. However, these studies have taken a holistic perspective and have not given due credibility to the important factors.

Our proposed methodology takes a novel approach by using 'Dx:Cancer' and 'Dx:CIN' as the target variable and taking into account known cervical cancer risk factors. This strategy may provide a more comprehensive understanding of cervical cancer risk and facilitate earlier disease detection. Incorporating cancer diagnosis as the objective variable may result in a cervical cancer risk prediction model that is more clinically accurate. This could potentially enhance screening and prevention efforts and aid in reducing the overall incidence of cervical cancer.

The remaining sections of the paper are organized as follows. Section II provides a comprehensive review of the existing literature on cervical cancer risk prediction models employing different Machine Learning Techniques. In Section III, we describe the dataset used for training and testing the proposed model, as well as the evaluation metrics used to assess the efficacy of the model. Section IV discusses the results of feature extraction, model training, and model evaluation for the proposed model. Section V presents the results acquired after running the proposed model on the dataset successfully. In Section VI, we summarize the main findings of the study and provide concluding remarks regarding the proposed cervical cancer risk prediction model.

II. RELATED WORKS

Juneja et al. [12] have surveyed the Indian demographic and found the factors which are likely to increase the risk of cervical cancer. They have found that the risk is greater in women with multiple sexual partners, unhygienic menstrual practices, early marriages and other unhealthy dietary and lifestyle choices. It majorly states that HPV has a high correlation with cervical cancer as HPV leaves the cervix prone to infections. Ratul et al. [13] performed a performance analysis of a given dataset using various machine learning models to analyze the models' capability of prediction. They achieved an accuracy of 93.33% in MLP and later the same accuracy in Decision Tree Classification (DTC), Random Forest Classifier (RFC), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). This accuracy was achieved with hyperparameter tuning. Ijaz et al. [14] used a chi-square test to conduct the feature extraction on the given dataset to remove unwanted columns and preserve the top 10 columns that displayed the maximum feature scores. The highest feature score was of the number of cigarettes smoked per year and the least among the top 10 were diagnosed with HPV and diagnosis. Then they used Density-based spatial clustering of applications with noise (DBSCAN) and iForest for the detection of outliers. These outliers were then excluded. Synthetic Minority Over-sampling Technique (SMOTE) and SMOTETomek have been used to balance the dataset. SMOTE and SMOTETomek are two popular techniques used in machine learning to address the problem of class imbalance. The researchers then did a performance evaluation for the four types of cervical cancer tests namely Hinselmann, Schiller, Cytology, and Biopsy. They chose the random forest classifier as the model algorithm as it showed optimal values in accuracy, sensitivity, and specificity. An app was made for getting input from users based on which the model would detect whether the person has cervical cancer or not. In a recent study, Alquran et al. [8] employed Principal Component Analysis (PCA), along with Canonical Correlation Analysis (CCA), to minimize dimensionality and identify the most prominent features that can be used to classify Pap smear images into five classes.

Prusty et al. [9] gave an alternative study showing that a Stratified k-folds cross-validation framework over the commonly used ML algorithm of choice would also help bring up the accuracy. Lihore et al. [15] gave another fresh approach to this problem was taken through the Boruta analysis that identifies subsets of features from the dataset that are relevant to the classification activity at hand. Yang et al. [16] presented a study on the development of a cervical cancer risk prediction model using machine learning techniques. They collected data from 280 patients with cervical cancer and 350 healthy individuals and analyzed the data to identify risk factors associated with cervical cancer. They used four machine learning algorithms, namely Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), and

Random Forest (RF), to develop and evaluate the performance of the prediction model. According to the results, age, number of pregnancies, smoking, and contraceptive use were significant risk factors for cervical cancer. The prediction model developed using the RF algorithm achieved the highest accuracy of 0.904 and Area Under the Receiver Operating Characteristic Curve(ROC AUC) of 0.966, indicating its high performance in predicting cervical cancer risk. The study highlights the potential of using machine learning techniques to develop accurate and reliable prediction models for cervical cancer risk assessment, which could aid in the early detection and prevention of the disease.

Rothberg et al. [17] described the development of a perpersonalized screening model for cervical cancer. They analyzed electronic health record data from over 58,000 women aged 30- 64 who had undergone cervical cancer screening in Northeast Ohio between 2007 and 2014. They used logistic regression to develop a model that could predict a woman's risk of developing cervical cancer within the next 5 years. The model included factors such as age, race/ethnicity, smoking status, and history of cervical dysplasia. They also found that the personalized screening model had higher sensitivity than current guidelines for cervical cancer screening. Overall, their study suggests that a personalized approach to cervical cancer screening based on individual risk factors may be more effective than the current one-size-fits-all approach.

Curia [18] presented a study on developing a cervical cancer risk prediction model using robust ensemble and explainable black box methods. They used data from 858 women who underwent cervical cancer screening in the region of Tuscany, Italy, between 2015 and 2018. Their study employed several machine learning algorithms, including Random Forest, Gradient Boosting, Support Vector Machine, and Artificial Neural Networks, to develop the prediction model. In addition, the study utilized two techniques, namely LIME and SHAP, for generating explanations of the predictions made by black box models. The results showed that the ensemble model developed using Random Forest, Gradient Boosting, and Artificial Neural Networks obtained an accuracy of 0.903 and AUC of 0.934 in predicting cervical cancer risk. The study also demonstrated the effectiveness of the explainable black box methods in generating interpretable explanations of the model's predictions. The study suggests that combining robust ensemble models with explainable black box methods can help improve the accuracy and interpretability of cervical cancer risk prediction models. The findings may have important implications for improving cervical cancer screening and prevention programs.

Li et al. [19] proposed a machine learning-based approach to develop an accurate prognosis prediction model for lung adenocarcinoma. They collected data from a large cohort of patients and used various machine-learning techniques to identify key features associated with prognoses, such as tumor size, histologic subtype, and lymph node status. These features were used to develop a prognosis prediction model that demonstrated high accuracy in predicting overall survival and disease-free survival. This approach highlights the potential of machine learning for improving the accuracy of cancer prognosis prediction and aiding personalized treatment planning.

Kourou et al. [20] explore the potential of machine learning algorithms, such as artificial neural networks, support vector machines, decision trees, and random forests, in cancer prognosis and prediction, and offer examples of their application in various cancer types. The importance of feature selection and extraction in machine learning-based cancer diagnosis is also discussed, along with the integration of machine learning with other technologies, like genomic data analysis and imaging analysis, to improve accuracy. The authors conclude by providing insights into the challenges and opportunities in this area. Huang et al. [21] developed a deep learning algorithm to predict Lung cancer risk following low-dose CT screening. Their approach integrated multiple clinical and imaging features and was trained on a large dataset of low-dose CT scans. The algorithm demonstrated high accuracy in predicting the risk of lung cancer and reduced the number of unnecessary follow-up screenings.

This approach shows potential for improving the efficiency and performance of programs used for lung cancer screening.

According to the works discussed, we learn that cervical cancer is influenced by multiple sexual partners, unhygienic menstrual practices, early marriages, unhealthy dietary and lifestyle choices, and a strong correlation with HPV. Various researchers have used multiple machine learning models to accurately predict cervical cancer. To balance the dataset and eliminate outliers, they have also employed techniques such as feature extraction, density-based spatial clustering, and synthetic minority oversampling. The studies have demonstrated that machine learning algorithms such as Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Gradient Boosting, and Artificial Neural Networks can be used to create accurate and trustworthy prediction models for cervical cancer risk assessment. Personalized screening models based on an individual's unique risk factors may be more effective than the conventional one-size-fits-all method.

III. DATASET DESCRIPTION

The dataset was obtained from the repository at UCI. The information was obtained from Universidad Central de Venezuela (UCV), a Venezuelan public university located in Caracas. The dataset contains data of 36 attributes on 858 female patients. Information on the attributes of the dataset is given in Table 1.

The four attributes namely Hinselmann, Schiller, Cytology and Biopsy are the tests for cervical cancer in the patient. Some patients did not wish to disclose some information due to privacy concerns and therefore a question mark (?) is present at that location in the dataset.

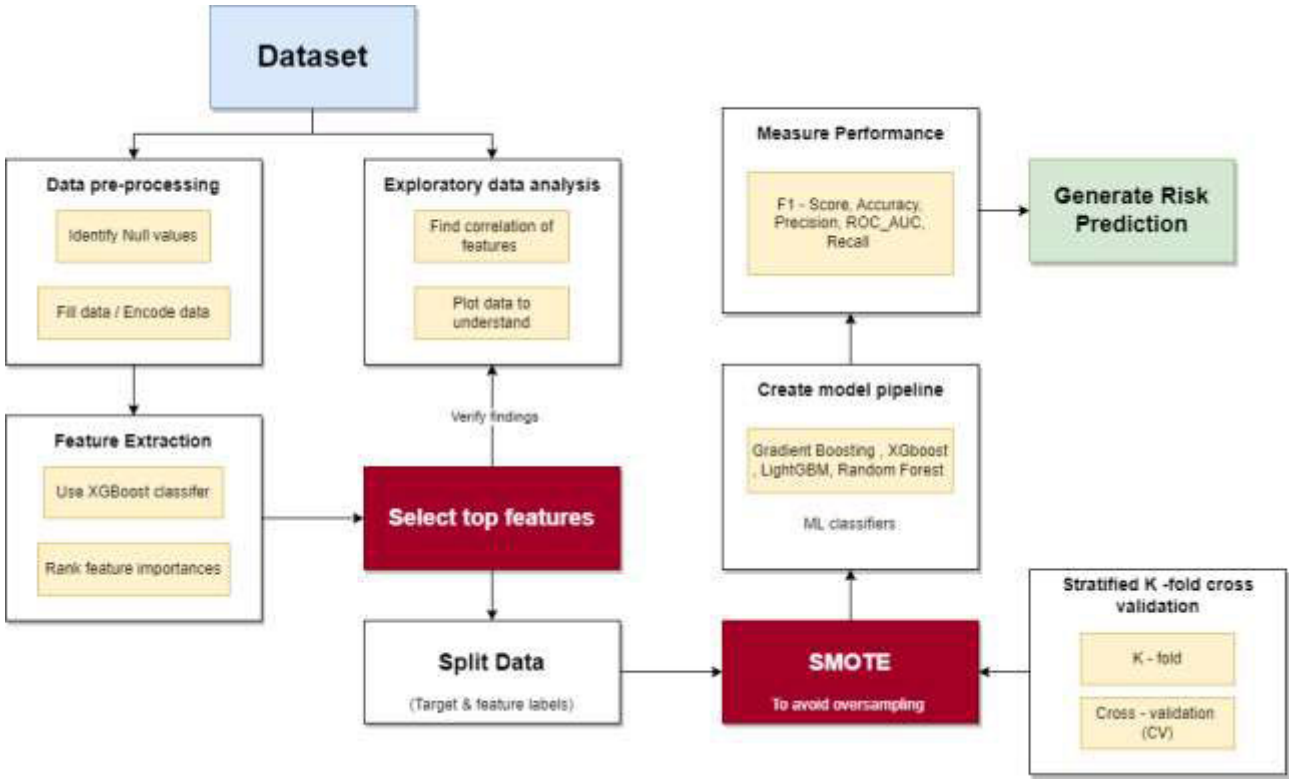


FIGURE 1. Overview of the proposed model.

IV. METHODOLOGY

The methodology involves several steps for processing and analyzing a dataset to construct a precise and trustworthy machine learning model as described in Figure 1. Initially pre-processing is performed, which includes converting the data to a numeric format, managing null values, and creating a new target variable. Then, to simplify the model, Feature Extraction is performed, which involves identifying the most important features and removing the less important ones. Model training involves constructing a pipeline of classification models, oversampling the data to account for class imbalance, and employing cross-validation to evaluate model performance. Model evaluation involves comparing the models using ROC AUC, precision, recall, and F1 score, among other metrics.

A. PRE-PROCESSING

We begin by importing the dataset and then converting the data to numeric data. Then, we check the dataset for null values. As the dataset contains several null values (?), it is necessary to fill them in order to increase model precision, eliminate data bias, and improve overall data quality. Classification of features and target columns to fill the empty values is done. Then, the feature columns were divided based on variable type, such as categorical or continuous variables. The median of the non-null data was used to replace null values in features containing continuous variable. Similarly, for categorical data features, the mode of non-null data was used to replace the null values.

A new column was named 'Cancer status' which sums the rows of the columns 'Dx:Cancer' and 'Dx:CIN'. This is because CIN is termed as precancer and therefore means the patient has a high risk of progressing into invasive cancer. We did this because the number of positives in the 'Dx:Cancer' alone was 18 which was quite less and the number of positives for 'Dx:CIN' was 9. The union of both these columns gave 27 instances of patients diagnosed with cancer. This in turn would give a better estimate of the risk factors. In the end, 'Cancer status' is made the target variable.

$$Cancer\_status = Dx : Cancer \cup Dx : CIN$$

Now that the dataset has undergone a thorough cleaning process, it is deemed ready for further processing. We have ensured the integrity and completeness of the data, enabling subsequent analyses and computations to be performed accurately and reliably. With the elimination of null values, the dataset is now in an optimal state for further exploration, feature engineering, modeling, and extracting valuable insights.

B. FEATURE EXTRACTION

This is the most vital part of the process as it helps in decreasing the dimensionality of the data which leads to an increase

TABLE 1. Attributes of the dataset.

S No.	Attribute	Type
1	Age	int
2	Number of sexual partners	int
3	First sexual intercourse (age)	int
4	Num of pregnancies	int
5	Smokes	bool
6	Smokes (years)	int
7	Smokes (packs/year)	int
8	Hormonal Contraceptives	bool
9	Hormonal Contraceptives (years)	int
10	IUD	bool
11	IUD (years)	int
12	STDs	bool
13	STDs (number)	int
14	STDs:condylomatosis	bool
15	STDs:cervical condylomatosis	bool
16	STDs:vaginal condylomatosis	bool
17	STDs:vulvo-perineal condylomatosis	bool
18	STDs:syphilis	bool
19	STDs:pelvic inflammatory disease	bool
20	STDs:genital herpes	bool
21	STDs:molluscum contagiosum	bool
22	STDs:AIDS	bool
23	STDs:HIV	bool
24	STDs:Hepatitis B	bool
25	STDs:HPV	bool
26	STDs: Number of diagnosis	int
27	STDs: Time since first diagnosis	int
28	STDs: Time since last diagnosis	int
29	Dx:Cancer	bool
30	Dx:CIN	bool
31	Dx:HPV	bool
32	Dx	bool
33	Hinselmann	bool
34	Schiller	bool
35	Cytology	bool
36	Biopsy	bool

in model performance as only the relevant data is used and the model can better capture the underlying patterns. We split the data into training and testing using a 75-25% split. Meaning that 75% of the data is used for training and the rest 25% for testing. After the data partitioning, normalization of the data is done which involves scaling the numerical features of a dataset to a standard range. This is done to ensure fair comparison and interpretation of the features.

XGBoost Classifier is used to find the feature importance of the attributes [22], [23]. XGBoost Classifier is an ensemble algorithm which uses tree-based methods that optimizes model performance via gradient boosting. It computes feature significance based on the number of times a feature is employed to divide the data across all the model's trees. The importance of each feature is then normalized, so that the sum of all feature importance equals one. By this, we have identified all relevant features and we can remove the least important features from the dataset to simplify the model, improve performance and reduce the computational complexity of the model.

C. MODEL TRAINING

For training the model, the training set is used which contains 75% of the data. To improve the data quality and overall performance of the model, the most important features were chosen, and the remaining features were removed from the dataset after determining the importance of each feature. We use a pipeline of models to identify the most suitable model to train with a high degree of precision to provide accurate prediction and risk percentage. In the proposed model, we have attempted to address data/class imbalance, also known as the oversampling problem. Class imbalance occurs when one class has significantly fewer samples than the other class in a binary classification problem. In such circumstances, a model's performance may be subpar since it will be partial towards the dominant class. To balance the class distribution, oversampling entails the creation of synthetic minority samples. Using techniques such as Random Oversampling, SMOTE, or ADASYN, these synthetic samples can be generated [24]. Since the dataset contains only 858 samples and more than 15 features, the accuracy achieved by the model may be subpar. To resolve this problem, cross-validation is used to evaluate the efficacy of a model on a standalone dataset. It entails partitioning the available data into numerous subsets (called folds), training the model on a few of the folds, and evaluating it on each remaining fold. This is done because it can reduce model overfitting on training data, tune the model using hyper parameters, and provide a more accurate estimate of model performance. In this proposed method we used Stratified Cross validation that is frequently used to evaluate the performance of a model on a given dataset. The primary purpose of this technique is to guarantee that data sets are divided into training and testing in such a way that the distribution of classes in the original dataset is preserved in each fold of the cross-validation process. Then, we cre- ated a pipeline of common machine learning classification models such as SVM, Random Forest classifier, Naive Bayes classifier, K-nearest neighbor classifier, LightGBM (Gradient boosting model), and Adaboost (ensemble learning model). These models were then tuned using hyperparameters for making the model more sensitive and adaptive to the unique characteristics of the data, leading to improved performance.

D. MODEL EVALUATION

We then perform cross-validation where we use the test set that contains 25% of the data to validate each machine learning model, compare the results, and then select the optimal model for making risk predictions. We compare models using machine learning metrics such as ROC, AUC score, F1 score, Precision, and recall. ROC Curve (AUC) is a metric for measuring the model's overall performance across all thresholds. It provides a single number representing the model's capacity to differentiate between positive and negative classes.

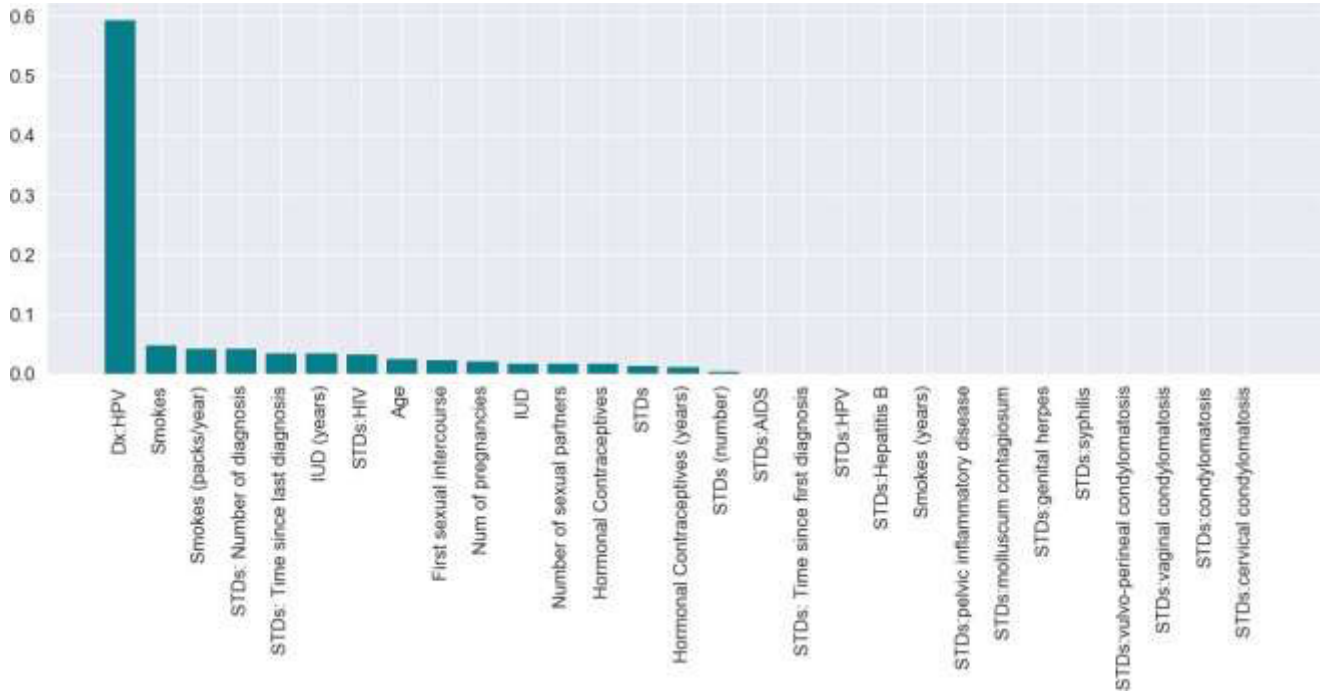


FIGURE 2. Extracted feature importance using XGBoost classifier.

A model with a ROC AUC score of 0.5 is equivalent to random chance while a score of 1 indicates a perfect model.

Precision is a metric that evaluates the proportion of correct predictions that are generated by a model. It describes the model's capacity to avoid false positives. A high precision score indicates that the model predicts few false positives.

$$Precision = \frac{TruePositive}{TruePositve + FalsePositive} \tag{1}$$

Recall is a metric which evaluates the proportion or correct predictions that are favorable relative to all actual instances of positive data. It describes the model's ability to identify positive instances. A high recall score indicates that the model correctly identifies the majority of positive data instances.

$$Recall = \frac{TruePositive}{TruePositve + FalseNegatives} \tag{2}$$

The F1 Score is the harmonic mean of precision and recall. It offers a singular number that represents the model's overall precision and recall performance. A model with a high F1 score performed admirably in both precision and recall.

$$F1score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3}$$

In conclusion, the ROC AUC, precision, recall, and F1 Score are essential metrics for evaluating the performance of a machine learning model. These metrics provide insight into the capability of the model to distinguish between positive and negative classes, avoid false positives, and identify positive instances in the data.

V. RESULTS AND DISCUSSION

This section discusses the results of the feature extraction, oversampling, and the evaluation of the models.

The feature extraction process showed significant results. The important features are ranked from most relevant to least relevant as shown in Figure 2. The figure depicts that one feature stands out as tremendously more significant than the others. The 'HPV' feature has the highest feature importance, which corresponds with the real world, i.e. people with HPV are highly at risk of getting cervical cancer. Other than HPV, other factors namely smoking, age, number of sexual partners, number of pregnancies, and so on are also features that affect the output but they are not as significant as HPV. This result is clinically accurate and can also be confirmed with the feature correlation heat map as shown in Figure 3.

After removing the unimportant features and oversampling using SMOTE, the dataset is used to train a pipeline of various classification models. The models used are Gradient Boosting, XGBoost, Naive Bayes, Ada Boost, Decision Tree, LightGBM, Random Forest, SVM using Radial Basis Function(RBF) kernel, SVM Linear, SVM Polynomial, Multi-Layer Perceptron (MLP), K-Nearest Neighbors(KNN) and Logistic Regression. The results of the models are calculated with and without feature extraction, K-folds

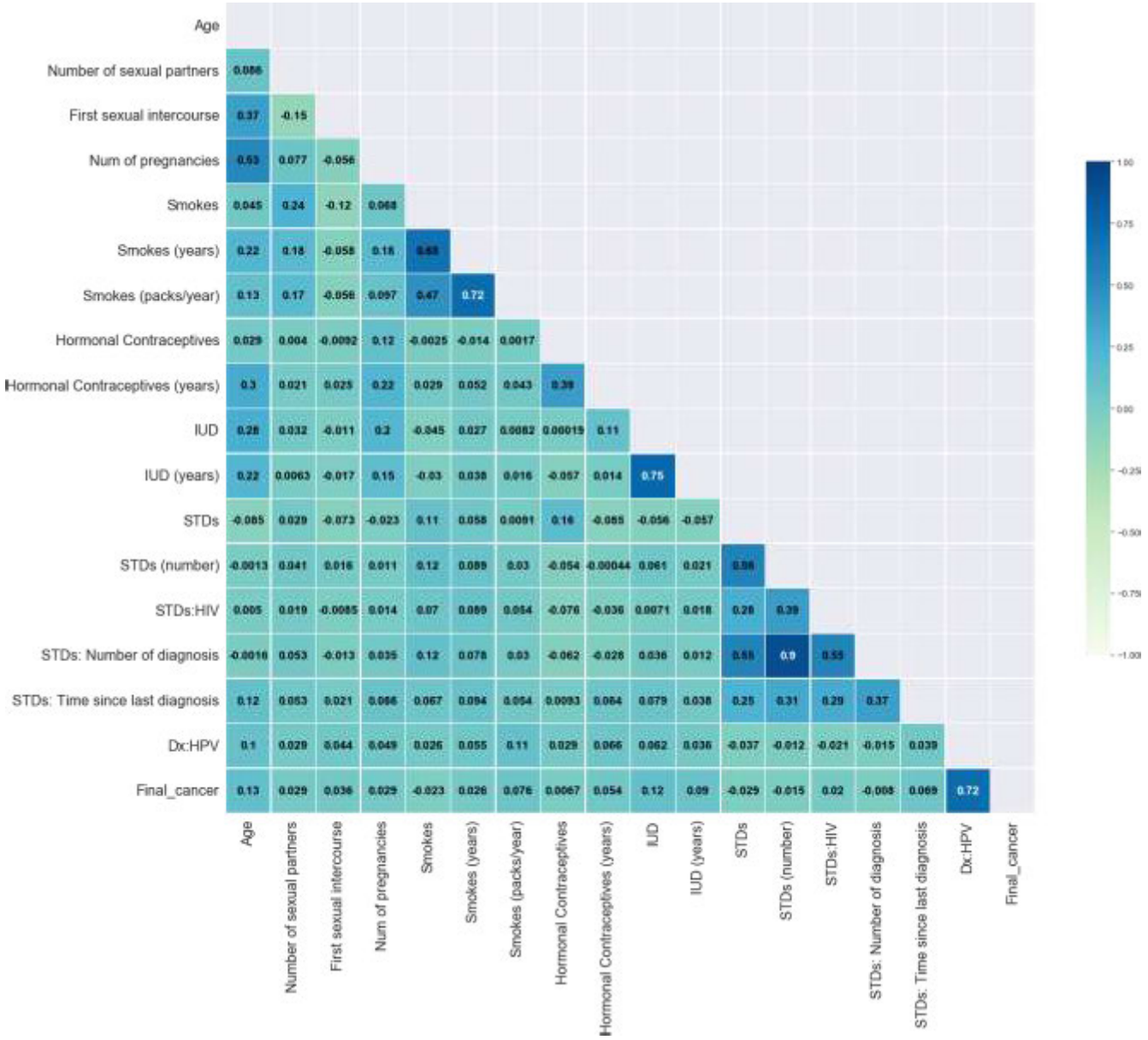


FIGURE 3. Extracted feature importance using correlation Heatmap.

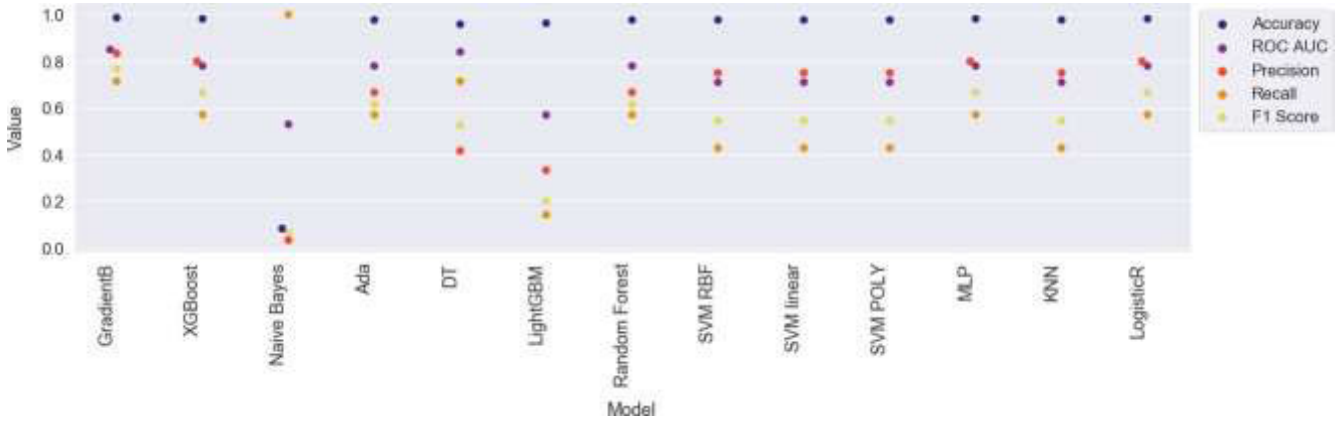
TABLE 2. Without using feature extraction, K-folds, and oversampling.

Models	Accuracy	ROC AUC	Precision	Recall	F1 Score
GradientB	0.986	0.996	0.988	0.984	0.986
XGBoost	0.986	0.996	0.985	0.986	0.985
Naive Bayes	0.724	0.827	0.835	0.575	0.675
Ada	0.978	0.992	0.977	0.978	0.977
DT	0.953	0.955	0.947	0.960	0.954
LightGBM	0.961	0.987	0.954	0.970	0.962
Random Forest	0.987	0.997	0.989	0.984	0.987
SVM RBF	0.923	0.986	0.891	0.967	0.927
SVM linear	0.763	0.852	0.818	0.697	0.747
SVM POLY	0.749	0.843	0.822	0.686	0.731
MLP	0.906	0.976	0.900	0.918	0.907
KNN	0.891	0.956	0.831	0.989	0.902
LogisticR	0.743	0.836	0.794	0.680	0.726

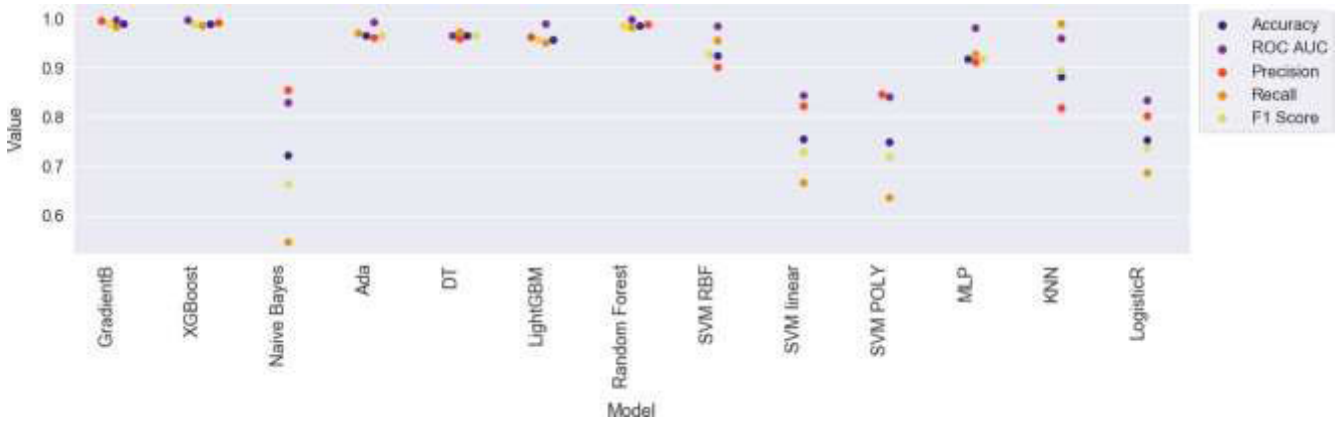
TABLE 3. Without using feature extraction, K-folds, and oversampling.

Models	Accuracy	ROC AUC	Precision	Recall	F1 Score
GradientB	0.986	0.85	0.833	0.714	0.769
XGBoost	0.981	0.78	0.8	0.571	0.667
Naive Bayes	0.084	0.53	0.034	1	0.066
Ada	0.977	0.78	0.667	0.571	0.615
DT	0.958	0.84	0.417	0.714	0.526
LightGBM	0.963	0.57	0.333	0.143	0.2
Random Forest	0.977	0.78	0.667	0.571	0.615
SVM RBF	0.977	0.71	0.75	0.429	0.545
SVM linear	0.977	0.71	0.75	0.429	0.545
SVM POLY	0.977	0.71	0.75	0.429	0.545
MLP	0.981	0.78	0.8	0.571	0.667
KNN	0.977	0.71	0.75	0.429	0.545
LogisticR	0.981	0.78	0.8	0.571	0.667

cross-validation, and oversampling. The models' evaluation results are depicted in Table 2 and Table 3 respectively and the results are visualized in Figure 4. The models do not demonstrate a significant increase in accuracy, but AUC, precision, recall, and F1 scores have all increased significantly.



(a) Without feature extraction, K-folds, and oversampling



(b) With feature extraction, K-folds, and oversampling

FIGURE 4. Comparison of models.

During previous training without the use of any techniques, the accuracy of linear models such as SVM classifiers was extremely high, while other metrics were low. We can infer that this was because the dataset was imbalanced. According to Table 3, despite having a recall of 1.00, the Naive Bayes model had a precision of 0.034 and an accuracy of 0.084. After feature extraction and oversampling, Naive Bayes accuracy was 0.732 and precision was 0.837, but recall decreased to 0.587. Similarly, the accuracy of gradient boosting did not increase significantly, but AUC, precision, recall, and precision did, which is advantageous for predicting the percentage of cervical cancer risk.

A. RISK PREDICTION

In general, the patient's risk percentage can be calculated as follows. The probability score or risk percentage is computed using the gradient boost model because it has been determined to be the most effective model in the pipeline.

$$Probabilityscore = \frac{M+1}{N+K} \tag{4}$$

Here,

M = No. of times the instance belongs to that class in the training set.

N = Total No. of instances in the training set.

K = No. of classes.

B. SAMPLE TEST CASES

An 18-year-old woman who has had five sexual partners, smokes an average of 37 packs of cigarettes per year, has been diagnosed with STDs three times, but does not have HPV has a risk of only 19%. In contrast, if the same woman had been diagnosed with HPV, her risk of developing cervical cancer would increase to 91.1.

Similarly, if a 45-year-old woman had only one sexual partner and her first sexual encounter occurred at age 30. With these feature values alone, the risk is 0.02%; if the same woman had been diagnosed with HPV, the risk would increase to 80%.

The preceding examples demonstrate that the risk percentage is heavily dependent on whether or not a person has HPV. The model still accounts for additional variables such as

TABLE 4. Comparison with the existing works.

Title	Model	Accuracy (%)	Precision (%)	Recall (%)
Ishrak et al. [13]	Bayes Net Gaussian Naive Bayes Random Forest Classifier	96.38 86.6 93.33	90.91 81.82 90.91	100 100 95.45
Curia F. [18]	Ensemble	95	100	67
Sujay et al. [25]	Decision Tree SVM Decision Stump	93.33 96.15 96.15	89 89 75	96.4 96.1 96.1
Lu et al. [26]	Ensemble MLP Logistic Regression	83.16 82.93 82.78	51.73 48.03 45.85	28.35 28.61 21.42
Akter et al. [27]	XGBoost Decision Tree Random Forest	93.33 93.33 93.33	0 80 100	100 100 75
Present work	**Gradient** **XGBoost** **Random Forest**	**98.6** **98.6** **98.7**	**99.8** **99.5** **98.9**	**98.4** **99.6** **98.4**

age, number of sexual partners, smoking, etc. Therefore, this model places the clinical observations in the right perspective and the model correlates well with the pathogenicity of the disease.

C. COMPARATIVE ANALYSIS

In Table 4, all three models utilized in the proposed work-Gradient Boosting, XGBoost, and Random Forest - have performed exceptionally well. The accuracy obtained by the Gradient Boosting and XGBoost models is 98.6%, while that of the Random Forest model is 98.7%.

In terms of accuracy, precision, and recall, it is evident that the proposed work models have outperformed the other models discussed in the table.

Ishrak et al. [13] proposed three models namely, Bayes net, Gaussian Naive Bayes, and Random Forest Classifier. Our models achieved a higher performance in terms of accuracy and precision. The ensemble model proposed by Curia F. [18] achieved an accuracy of 95%, a precision of 100%, signifying a high accuracy in correctly classifying positive instances. However, the model's recall was 67%, implying that it missed some positive instances during classification.

Suman and Hooda [25] achieved 96.38 percent accuracy with Bayes Net and SVM models, which is lower than the pro- posed work models. With their ensemble model, Lu et al. [26] achieved an accuracy of 83.16 percent, which is significantly lower than the proposed work models.

The Random Forest model proposed by Akter et al. [27] has a precision of 100 but a recall of 75. Likewise, their XGBoost and Decision Tree models have a recall of 100 but significantly low precision.

While comparing the different works to our own, several common aspects emerge. Some studies [25], [26] have also considered the overfitting problem and have used K-folds cross-validation. Similar to the study [13], we utilized hyperparameter optimization techniques to optimize our machine learning models in our study.

There are also several distinct aspects in our study compared to the other works. In the work [18] the authors have utilized ensemble learning and interpretability methods like LIME and SHAP whereas we employed a simpler approach using a correlation heatmap to explore the relationships in the data. Some research works [25], [26], [27] have not used over-sampling techniques to counter overfitting which can lead to biased and unreliable predictions. One of the studies [26] also introduced a gene assist module using a genomic sequencing dataset to enhance the robustness of the predictions, which differentiates their work from ours.

Our models produce better results across all metrics as we have accounted for all problems such as overfitting which can be caused by insufficient and imbalanced training data. We have also chosen only the most relevant features, resulting in accurate and efficient models. Our predictive model goes beyond providing a simple binary response by considering a vast array of risk factors. By incorporating these variables, our model provides a more comprehensive risk assessment, expressed as a percentage. This estimation based on percentages provides a more thorough understanding of the level of risk, enabling individuals to make more informed decisions and change their lifestyle based on circumstances.

Despite the promising results of our study, there are several limitations that must be considered. One of the main limitations is the small size of our dataset, which may have limited the generalizability of our findings. The small sample size may have also impacted the statistical

effectiveness of our study and hampered our capacity to detect significant implications or associations between variables.

VI. CONCLUSION

The proposed model provides results that can help in recognising women who are at a higher risk of developing cervical cancer enabling healthcare providers to offer early screening and detection such as Pap smear and HPV testing. This can help to catch the disease in its early stages; when it is most treatable.

Women found to be at greater risk can be counseled for lifestyle modifications such as quitting smoking, following safe sex practices, etc. This model can help health-care providers to tailor screening and prevention strategies to individual patients based on their risk factors. This can decrease the number of redundant examinations and improve the efficacy of screening programs.

While this study has provided valuable insights into the factors that influence cervical cancer, there is still scope for future work in this area. We aim to replicate our findings in larger, more diverse datasets to increase the generalizability of our results. A dataset is considered substantial when it consists of a minimum of 1000 or more data points, depending on the specific topic under consideration. Acquiring a good dataset will ensure better model performance as well as better insights into the factors at play.

REFERENCES

[1] National Health Portal. (2021) *Cervical Cancer*. Accessed: Apr. 8, 2023. [Online]. Available: https://www.nhp.gov.in/disease/cancer/cervical-cancer

[2] J. Ferlay, M. Ervik, F. Lam, M. Colombet, L. Mery, M. Piñeros, A. Znaor, Soerjomataram, and F. Bray, (2020). *Global Cancer Observatory: Cancer Today*. Lyon, France: International Agency for Research on Cancer. Accessed: Mar. 27, 2023. [Online]. Available: https://gco.iarc.fr/today

[3] WHO Director-General's statement. *Who Director-General's Statement on the Call to Eliminate Cervical Cancer as a Public Health Problem*. Accessed: Mar. 29, 2023. [Online]. Available: https://www.who.int/initiatives/cervical-cancer-elimination-initiative

[4] N. Al Mudawi and A. Alazeb, "A model for predicting cervical cancer using machine learning algorithms," *Sensors*, vol. 22, no. 11, p. 4132, May 2022.

[5] N. Kashyap, N. Krishnan, S. Kaur, and S. Ghai, "Risk factors of cervical cancer: A case-control study," *Asia–Pacific J. Oncol. Nursing*, vol. 6, no. 3, pp. 308–314, Jul. 2019.

[6] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Inform.*, vol. 35, nos. 5–6, pp. 352–359, Oct. 2002.

[7] J. P. Dsouza, S. Van den Broucke, S. Pattanshetty, and W. Dhoore, "Exploring the barriers to cervical cancer screening through the lens of implementers and beneficiaries of the national screening program: A multi-contextual study," *Asian Pacific J. Cancer Prevention (APJCP)*, vol. 21, no. 8, pp. 2209–2215, Aug. 2020.

[8] H. Alquran, M. Alsalatie, W. A. Mustafa, R. A. Abdi, and A. R. Ismail, "Cervical Net: A novel cervical cancer classification using feature fusion," *Bioengineering*, vol. 9, no. 10, p. 578, Oct. 2022.

[9] S. Prusty, S. Patnaik, and S. K. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Frontiers Nanotechnol.*, vol. 4, Aug. 2022, Art. no. 972421.

[10] K. Adem, S. Kiliçarslan, and O. Cömert, "Classification and diagnosis of cervical cancer with stacked autoencoder and softmax classification," *Expert Syst. Appl.*, vol. 115, pp. 557–564, Jan. 2019.

[11] M. M. Ali, K. Ahmed, F. M. Bui, B. K. Paul, S. M. Ibrahim, J. M. W. Quinn, and M. A. Moni, "Machine learning-based statistical analysis for early stage detection of cervical cancer," *Comput. Biol. Med.*, vol. 139, Dec. 2021, Art. no. 104985.

[12] A. Juneja, A. Sehgal, A. Mitra, and A. Pandey, "A survey on risk factors associated with cervical cancer," *Indian J. Cancer*, vol. 40, no. 1, pp. 15–22, 2003.

[13] I. J. Ratul, A. Al-Monsur, B. Tabassum, A. M. Ar-Rafi, M. M. Nishat, and F. Faisal, "Early risk prediction of cervical cancer: A machine learning approach," in *Proc. 19th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON). IEEE*, May 2022, pp. 1–4.

[14] M. F. Ijaz, M. Attique, and Y. Son, "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods," *Sensors*, vol. 20, no. 10, p. 2809, May 2020.

[15] U. K. Lilhore, M. Poongodi, A. Kaur, S. Simaiya, A. D. Algarni, H. Elmannai, V. Vijayakumar, G. B. Tunze, and M. Hamdi, "Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques," *Comput. Math. Methods Med.*, vol. 2022, May 2022, Art. no. 4688327.

[16] W. Yang, X. Gou, T. Xu, X. Yi, and M. Jiang, "Cervical cancer risk prediction model and analysis of risk factors based on machine learning," in *Proc. 11th Int. Conf. Bioinf. Biomed. Technol.*, May 2019, pp. 50–54.

[17] M. B. Rothberg, B. Hu, L. Lipold, S. Schramm, X. W. Jin, A. Sikon, and G. B. Taksler, "A risk prediction model to allow personalized screening for cervical cancer," *Cancer Causes Control*, vol. 29, no. 3, pp. 297–304, Mar. 2018.

[18] F. Curia, "Cervical cancer risk prediction with robust ensemble and explainable black boxes method," *Health Technol.*, vol. 11, no. 4, pp. 875–885, Jul. 2021.

[19] Y. Li, D. Ge, J. Gu, F. Xu, Q. Zhu, and C. Lu, "A large cohort study identifying a novel prognosis prediction model for lung adenocarcinoma through machine learning strategies," *BMC Cancer*, vol. 19, no. 1, pp. 1–14, Dec. 2019.

[20] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Jan. 2015.

[21] P. Huang, C. T. Lin, Y. Li, M. C. Tammemagi, M. V. Brock, S. Atkar-Khattra, Y. Xu, P. Hu, J. R. Mayo, H. Schmidt, M. Gingras, S. Pasian, L. Stewart, S. Tsai, J. M. Seely, D. Manos, P. Burrowes, R. Bhatia, M.-S. Tsao, and S. Lam, "Prediction of lung cancer risk at follow-up screening with low-dose CT: A training and validation study of a deep learning method," *Lancet Digit. Health*, vol. 1, no. 7, pp. e353–e362, Nov. 2019.

[22] T. Maguire, L. Manuel, R. Smedinga, and M. Biehl, "A review of feature selection and ranking methods," in *Proc. 19th SC@ RUG*, 2022, pp. 15–20.

[23] X. Shi, Y. D. Wong, M. Z.-F. Li, C. Palanisamy, and C. Chai, "A feature learning approach based on XGBoost for driving assessment and risk prediction," *Accident Anal. Prevention*, vol. 129, pp. 170–179, Aug. 2019.

[24] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya, and M. Ismail, "SMOTE for handling imbalanced data problem: A review," in *Proc. 6th Int. Conf. Informat. Comput. (ICIC)*, Nov. 2021, pp. 1–8.

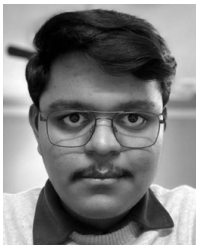
[25] S. K. Suman and N. Hooda, "Predicting risk of cervical cancer: A case study of machine learning," *J. Statist. Manage. Syst.*, vol. 22, no. 4, pp. 689–696, May 2019.

[26] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, "Machine learning for assisting cervical cancer diagnosis: An ensemble approach," *Future Gener. Comput. Syst.*, vol. 106, pp. 199–205, May 2020.

[27] L. Akter, Ferdib-Al-Islam, M. M. Islam, M. S. Al-Rakhami, and M. R. Haque, "Prediction of cervical cancer from behavior risk using machine learning techniques," *Social Netw. Comput. Sci.*, vol. 2, no. 3, pp. 1–10, May 2021.

RITHVIK HARIPRASAD was born in Bengaluru, India, in 2002. He received the primary and secondary education from the Delhi Public School, Vasant Kunj, New Delhi, India. He is currently pursuing the B.Tech. degree in computer science and engineering with the Vellore Institute of Technology, Tamil Nadu, India. His research interests include medical imaging, machine learning, and oncology.

NAVAMANI T M received the M.E. (C.S.E.) and Ph.D. degrees from Anna University, Chennai. She is currently a Senior Associate Professor with the School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Vellore, India. She has 25 years of teaching experience. She has published more than 40 papers in international and national journals, conferences, and book chapters. Her research interests include data sciences, human–computer interaction, wireless ad hoc networks, network security, and mobile computing. She is a Life Member of the Indian Society for Technical Education (ISTE).



TEJAS RAVINDRA ROTE was born in Maharashtra, India, in 2002. He is currently pursuing the degree in computer science and engineering with the Vellore Institute of Technology (VIT), Vellore, India. His research interests include machine learning and creating prediction models for solving problems.

ISHITA CHAUHAN was born in New Delhi, India, in 2002. She is currently pursuing the B.Tech. degree in computer science with the Vellore Institute of Technology (VIT), Vellore, India. Her research interests include natural language processing, image processing, machine learning, and genomics.