Proactive cervical cancer risk assessment using data-driven analytics

Sreelatha[1,2], **Vrinda Shivashetty**[3]
[1]Department of Information Science and Engineering, Sai Vidya Institute of Technology, Visvesvaraya Technological University, Belagavi, India
[2]Department of Computer Science and Engineering, Presidency University, Bangalore, India
[3]Department of Information Science and Engineering, Sai Vidya Institute of Technology, Bangalore, India

Article Info

Article history:

Received Oct 22, 2023
Revised Feb 17, 2024
Accepted Mar 21, 2024

Keywords:

Cervical cancer
Lifestyle data
Machine learning
Multifactorial clinical data
Risk prediction

ABSTRACT

This study introduces a sophisticated predictive model integrating clinical and lifestyle data addressing the critical public health challenge of cervical cancer, particularly in regions lacking routine screenings. Leveraging data-driven analytics, the proposed model undergoes comprehensive preprocessing, including exploratory data analysis, missing value imputation, and feature extraction. Feature selection is carried out using the XGBoost classifier to ensure model efficacy. Data normalization and class balance via oversampling techniques are applied, with model validation conducted through stratified cross-validation. The optimized feature vector is then employed to train a LightGBM model. Utilizing a retrospective dataset of 858 patients from the Hospital Universitario de Caracas, Venezuela, comprising demographic, lifestyle, and medical history data, the LightGBM model achieves an impressive accuracy of 98%, outperforming similar existing approaches. The study outcome demonstrates the effectiveness of the proposed data modelling framework and feature selection, along with the choice of LightGBM as a suitable classifier. The proposed predictive framework can efficiently aid healthcare professionals in prioritizing high-risk patients for further evaluation and intervention.

Corresponding Author:

Sreelatha
Department of Computer Science and Engineering, Presidency University
Bangalore, India
Email: sreelatha.pk@presidencyuniversity.in

1. INTRODUCTION

Cervical cancer is a severe disease that affects millions of women worldwide. It is characterized by the abnormal growth of cells in the cervix, the opening to the uterus [1]. The leading cause of cervical cancer is persistent infection with certain types of human papillomavirus (HPV), a prevalent virus that is spread through sexual contact [2]. Most people who are infected with HPV clear the virus on their own, but some people develop a persistent infection, which can lead to the development of precancerous cells, which can eventually turn into cancer [3]. Medical advances have led to several ways to prevent and treat early-stage cervical cancer. Unfortunately, the reality is harsh; many women are diagnosed with cervical cancer when the disease has reached an advanced stage, making treatment more complex, expensive, and less likely to be successful [4]. Despite available preventive mechanisms, early detection and management of cervical cancer remains a significant health problem. Late diagnosis of cervical cancer is very common as many women are diagnosed with advanced cervical cancer, mainly due to a lack of regular screening, economic barriers to healthcare and the issue of socio-cultural [5], [6]. The World Health Organization (WHO) recommends that women aged 30-49 undergo cervical cancer screening at least once every five years, yet adherence to these guidelines is insufficient, particularly in low-resource settings and regions with limited awareness about cervical cancer and its implications [7]. There are many barriers to cervical cancer screening. One barrier is that women may not know about the risk factors and symptoms of cervical cancer, so they do not realize how important it is to get screened regularly [8]. Another barrier is that women may not be comfortable talking about their reproductive health, especially with male doctors. In conservative societies, women may be hesitant to talk about their reproductive health because of cultural norms or personal inhibitions [9]. This can lead to delayed or missed diagnoses. Another barrier is the availability and affordability of screening facilities. In many places, primarily rural areas, there is no infrastructure for regular, affordable cervical cancer screening [10]. This makes it difficult for women to access these essential healthcare services. Therefore, there is an urgent need for an effective solution that can facilitate early detection of cervical cancer. The emergence of artificial intelligence (AI) technology in healthcare is revolutionizing the diagnostic process [11]. Using clinical data and patient histories, AI-based methods can identify patterns and correlations that may not be much explored using traditional analytical methods [12]. However, when the data comes from clinical settings, it is difficult to build accurate predictive models that can be clinically acceptable because of the inherent heterogeneity and complexity of this form of data [13], [14].

In recent state-of-the-art works, many researchers have offered risk prediction models to various clinical contexts. In cervical cancer, the existing schemes adopt various attributes, involve collecting clinical data and demographic information from women and applying different machine learning approaches to model risk factors. Ijaz *et al.* [15] developed a prediction model to predict cervical cancer by evaluating risk factors. This model uses noise-resistant density clustering and anomaly isolation trees to handle outliers in the data and synthetic minority over-sampling technique (SMOTE) technique to balance the dataset. The model is based on a random forest classifier for classification tasks. Lilhore *et al.* [16] addressed the limitations of the standard Pap smear examination for cervical cancer, which can produce many false-positive results due to human error. The authors introduced an efficient feature selection and prediction model using Boruta analysis and the support vector machine (SVM) method. However, this study did not discuss the model's scalability to adapt to changing cancer screening guidelines and practices. Putri *et al.* [17] considered a case study of cervical cancer in Indonesia, where it ranks as the second deadliest form of cancer after breast cancer. The research presented a methodology combining region-aware segmentation, texture feature and artificial neural network (ANN) to localize cervical cancer areas and categories CT images into normal and abnormal. Curia [18] introduces an ensemble method for cervical cancer forecasting focusing on minimizing errors and false positives. It also integrates explainability and interpretability features to make the model's results and decision-making processes more precise and more understandable in the context of clinical decisions. Youneszade *et al.* [19] reviewed recent AI-based approaches to improve computer-aided diagnostic (CAD) systems for cervical cancer screening images. They identified areas where further research is needed and outlined future directions. Dhivya *et al.* [20] evaluated the performance of different supervised classifiers for automatic cervical tumour classification. They developed an optimized meta-learning algorithm to tune the hyperparameters of the classifiers and select the best model for the given dataset. Zhou *et al.* [21] developed a framework incorporating various statistical indicators to analyze risk factors more efficiently based on thorough correlation analysis and close relationships among indicators derived from existing literature. Priya and Karthikeyan [22] address the limitations of existing models, such as low accuracy and high processing time in cervical cancer detection. Their method uses SMOTE to address class imbalance issues, the artificial bee colony (ABC) optimization algorithm to select the most essential features, and the long short-term memory (LSTM) learning model to classify cervical cancer based on the selected features. Şentürk and Uzun [23] applied a median filter to remove noise from Pap smear images and used transfer learning for early cervical cancer diagnosis. Mudawi and Alazeb [24] compared the performance of different classical supervised classifiers for early cervical cancer prediction. They found that random forest and decision tree classifiers achieved the highest classification scores. Tanimu *et al.* [25] integrated recursive feature elimination and L1 regularization methods to identify the most pertinent attributes for the decision tree classifier for precise cervical cancer predictions.

Hence, it can be realized that the existing literature consists of a wide range of solutions to develop a predictive model for cervical cancer. Since clinical data often suffers from data inconsistency and incompleteness. It has been identified that most existing works lack sophisticated feature engineering and are more specific to addressing either class imbalance problems or feature selection. Some studies considered addressing both but were subjected to biased predictive learning, hindering their applicability in real-world applications. This paper introduces a novel cervical cancer risk assessment framework, employing the power of data analytics to scrutinize various risk factors. The proposed framework then utilizes AI algorithms to infer from clinical data, predicting the probability of individuals with a high risk of developing cervical cancer. The next section presents the framework design and the implementation procedure adopted.

2. METHOD

This section presents the proposed system analytical design and a detailed discussion of the computing procedure adopted in the system implementation for cervical cancer risk prediction based on the input patient's demographic and habits-related information. Basically, the proposed research focuses on developing an early detection model for cervical cancer, which is of utmost importance, given the real-world challenges surrounding cancer diagnosis and its associated health implications. The main objective of the proposed methodology is to create an accurate and reliable predictive model that can be utilized to identify patients who are at risk of cervical cancer. This model will assist healthcare professionals in making informed decisions about which patients should undergo more comprehensive examinations. The schematic architecture of the proposed cervical cancer risk prediction framework is illustrated in Figure 1.

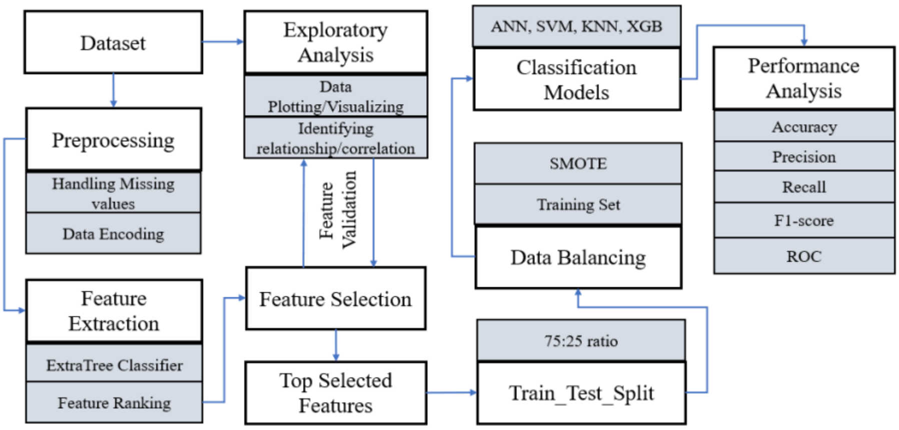


Figure 1. Illustrates the schematic architecture of the proposed framework, following a block-based workflow

The methodology adopted in the proposed system adopts a systematic implementation procedure. The first step carried out in the system development is exploratory data analysis (EDA) to gain insights about the dataset attributes and understand the complexity of the dataset. This phase involves visual analysis, checking missing values and performing descriptive analysis. Next, a preprocessing mechanism is adopted to provide suitable treatment to the dataset to address any outliers and inconsistencies. Further, statistical analysis is done to understand and identify potential relationships between risk attributes and outcome class labels. This process helps to analyze and select the predictive model's relevant attributes. The study then performs a ranking of the features under consideration using an extra-tree classifier test. Afterwards, the most dominating features are selected based on the ranking score. The top optimal features are then vectorized and subjected to the learning analytics-based predictive task. The study in this phase employed various computational intelligence, particularly a shallow machine learning classifier, which is trained using selected features and deployed to perform prediction of the presence of high-risk HPV strains or cervical cancer risk. The benchmarking of the proposed system is done concerning confusion matrix, accuracy, f1-score, and comparative analysis with different machine learning models and similar existing studies.

The proposed system retains high utility in the real-world healthcare application. It can be deployed in hospitals, clinics, and screening centres, where it can serve as a support system for the initial screening of cervical cancer and help in a better and more rapid decision-making process. For example, if the patient is identified or predicted as at risk of cervical cancer, the experts then prioritise patients for further examinations and intervention using an advanced screening process, ultimately leading to earlier detection of cancer and treatment, thereby better patient outcomes.

2.1. Dataset description

The cervical cancer dataset used in this study was obtained from the machine learning repository of the University of California, Irvine (UCI) [26]. This dataset is essential for predicting risk factors and diagnosing cervical cancer. It was carefully prepared using data collected from 858 patients at the Hospital Universitario de Caracas in Venezuela. The specific details of the dataset are outlined in Table 1.

Based on the statistics provided in Table 1, it can be analyzed that the dataset includes demographic information, lifestyle habits, and medical history. This dataset contains four crucial response variables for cervical cancer predictive modelling: Hinselmann, Schiller, Citology, and Biopsy. These features are essential for diagnosing cervical cancer in patients. Their accumulation provides a comprehensive representation of the variables that influence cervical cancer. The dataset contains missing values, as some patients declined to answer certain questions due to privacy concerns.

Table 1. Summary of the cervical cancer risk dataset

SI. No	Attributes	Data Type	Description
0	Age	int32	Age of a woman
1	Number of sexual partners	int32	Total number of sexual partners
2	First sexual intercourse	int32	Age of a woman when she had her first sexual intercourse.
3	Num of pregnancies	int32	Total number of times the woman got pregnant.
4	Smokes	bool	Whether the women smokes or not.
5	Smokes (years)	int32	Number of years for which the woman is smoking.
6	Smokes (packs/year)	int32	Total number of packets of cigarettes per year the woman smokes.
7	Hormonal Contraceptives	bool	Whether the women use hormonal contraceptives or not.
8	Hormonal Contraceptives (years)	int32	Total years for which contraceptive method was used by women.
9	IUD	bool	The intrauterine contraceptive device was used or not.
10	IUD (years)	int32	For how many years the IUD was used.
11	STDs	bool	The presence of sexually transmitted diseases (STD).
12	STDs (number)	int32	Total number of STD present with the patient.
13	STDs: condylomatosis	bool	The presence of condylomatosis with the patient.
14	STDs: cervical condylomatosis	bool	The presence of cervical condylomatosis.
15	STDs: vaginal condylomatosis	bool	The presence of vaginal condylomatosis.
16	STDs: vulvo-perineal condylomatosis	bool	The presence of vulvo- perineal condylomatosis.
17	STDs: syphilis	bool	The presence of syphilis.
18	STDs: pelvic inflammatory disease	bool	The presence of pelvic inflammatory disease.
19	STDs: genital herpes	bool	The presence of genital herpes.
20	STDs: molluscum contagiosum	bool	The presence of molluscum contagiosum.
21	STDs: AIDS	bool	The presence of AIDS in the patient.
22	STDs: HIV	bool	The presence of HIV in the patient.
23	STDs: Hepatitis B	bool	The presence of hepatitis B in the patients.
24	STDs: HPV	bool	The presence of HPV in the patients.
25	STDs: Number of diagnoses	int32	The total number of times the STDs have been diagnosed.
26	STDs: Time since first diagnosis	int32	The total number of years since the first diagnose.
27	STDs: Time since last diagnosis	int32	The total number of years elapsed since the last diagnose.
28	Dx: Cancer	bool	The presence of cancer after the diagnose.
29	Dx: CIN	bool	The presence of cervical intraepithelial neoplasia.
30	Dx: HPV	bool	The presence of human papilloma viruses (HPV).
31	Dx	bool	The presence of CIN or HPV.
32	Hinselmann	bool	A colposcopy test to examine a magnified view of the cervix
33	Schiller	bool	Iodine test to diagnose cervical cancer.
34	Citology	bool	A PaP smears test, helps detect abnormal cells in the cervix
35	Biopsy	bool	A surgical procedure where a small tissue is removed from the cervix

2.2. Preprocessing

Preprocessing is an important operation in data-driven analytics as it alone contributes 70% to performance improvement. The initial preprocessing step involves importing the dataset and conducting EDA to understand its characteristics and inherent complexities of the dataset. During EDA, missing data instances represented as '?' were identified. To standardize the handling of these missing values, '?' instances were replaced with 'not a number' (NAN). Imputing missing values is critical for ensuring data integrity and accuracy, thereby improving model precision and mitigating potential biases. The dataset contains numerical and categorical attributes, so the imputation approach varied accordingly. For numerical attributes, missing values were substituted with the median value of non-null entries in the respective feature, and for categorical attributes, missing values were replaced with the mode, i.e., the most frequently occurring non-null value within the feature.

A new column titled "cancer risk" was also introduced to augment the dataset's predictive capacity. This column combines the entries from the "Dx: Cancer" and "Dx: CIN" columns. It is important to note that "Dx: CIN" is indicative of a precancerous condition, indicating that patients diagnosed with CIN have an increased risk of developing invasive cancer. The motivation for adding this column was the relatively low incidence rates in the individual columns: "Dx: Cancer" had 18 positive instances, and "Dx: CIN" had 9. By merging the data from these two columns, a total of 27 instances indicating cancer diagnosis were obtained, which is more representative of the underlying risk factors.

2.3. Feature extraction and selection

Dimensionality reduction is crucial in data preprocessing as it eliminates irrelevant features and only retains important and highly correlated features. This process simplifies the dataset by removing unnecessary features and enables the model to identify and utilize key patterns in the data more effectively. By reducing the number of features, the model becomes more efficient and less prone to overfitting, where it learns to memories the training data rather than generalize to new, unseen data. It also reduces the computational complexity in data processing and during model training. The initial step in the proposed feature engineering process involves splitting the dataset into training and testing subsets using a 70-30% split ratio. Following this partitioning, normalization is performed to scale the numerical attributes of the dataset to a standardized range, typically [0,1]. This normalization ensures that no single feature disproportionately influences the model due to its scale.

$$\text{normalized}(x) = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

The study then performs feature selection using the ExtraTree classifier, a tree-based ensemble learning technique. Unlike regular decision trees that choose the best split among a set of available features, ExtraTree selects features and split points at random, which helps reduce the variance. The decision-making process of the tree is regulated by entropy value, which helps in determining the purity of a split, numerically given as (2).

$$Entropy(S) = p_+ + log_2(p_+) - p_- log_2(p_-) \tag{2}$$

Where in (2) S is the set of samples, p_+ is the proportion of positive samples and p_- is the proportion of negative samples. The ExtraTree classifier aggregates multiple such trees and computes an average to determine the importance of each feature. The significance of a feature is directly related to the frequency with which it appears in the trees and the depth at which it appears. Once the feature importance is derived, these values are subjected to standardization operation to ensure that the collective importance across all features sums up to one. Mathematically, given a set of feature importance values F, the normalized importance F_{standr} for feature *i* as in (3).

$$F_{standr_i} = \frac{F_i}{\sum_{j=1}^{n} F_j} \tag{3}$$

Where, *n* is the total number of features. Following this procedure, the study selects the most significant features and excludes the least significant attributes. This process not only refines the model but also enhances its performance and diminishes computational overhead, offering a more efficient learning environment. Figure 2 illustrates the top selected features, following cross-correlation analysis in Figure 3.

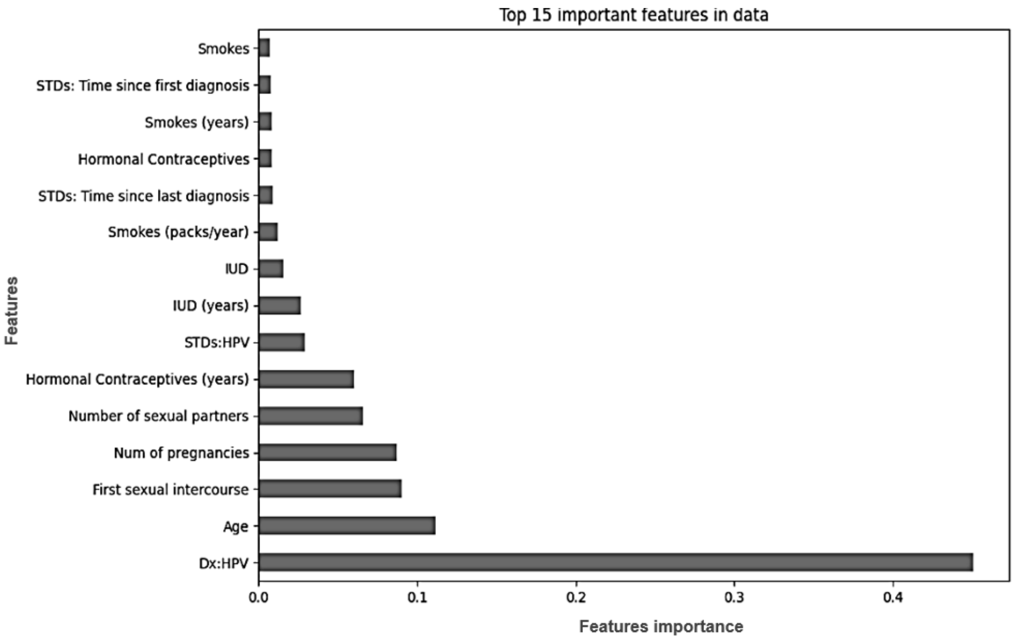
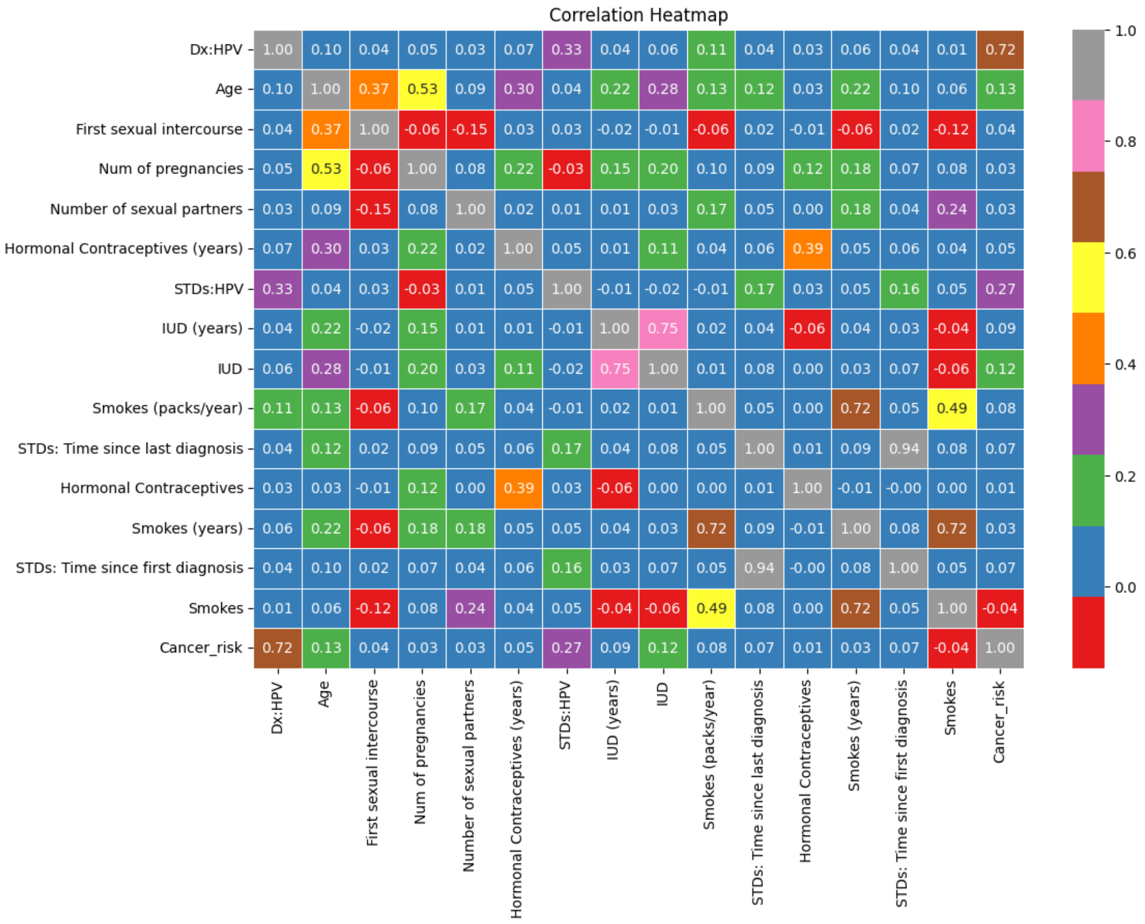


Figure 2. Illustrates the schematic architecture of the proposed framework, following a block-based workflow



	Dx:HPV	Age	First sexual intercourse	Num of pregnancies	Number of sexual partners	Hormonal Contraceptives (years)	STDs:HPV	IUD (years)	IUD	Smokes (packs/year)	STDs: Time since last diagnosis	Hormonal Contraceptives	Smokes (years)	STDs: Time since first diagnosis	Smokes	Cancer_risk
Dx:HPV	1.00	0.10	0.04	0.05	0.03	0.07	0.33	0.04	0.06	0.11	0.04	0.03	0.06	0.04	0.01	0.72
Age	0.10	1.00	0.37	0.53	0.09	0.30	0.04	0.22	0.28	0.13	0.12	0.03	0.22	0.10	0.06	0.13
First sexual intercourse	0.04	0.37	1.00	-0.06	-0.15	0.03	0.03	-0.02	-0.01	-0.06	0.02	-0.01	-0.06	0.02	-0.12	0.04
Num of pregnancies	0.05	0.53	-0.06	1.00	0.08	0.22	-0.03	0.15	0.20	0.10	0.09	0.12	0.18	0.07	0.08	0.03
Number of sexual partners	0.03	0.09	-0.15	0.08	1.00	0.02	0.01	0.01	0.03	0.17	0.05	0.00	0.18	0.04	0.24	0.03
Hormonal Contraceptives (years)	0.07	0.30	0.03	0.22	0.02	1.00	0.05	0.01	0.11	0.04	0.06	0.39	0.05	0.06	0.04	0.05
STDs:HPV	0.33	0.04	0.03	-0.03	0.01	0.05	1.00	-0.01	-0.02	-0.01	0.17	0.03	0.05	0.16	0.05	0.27
IUD (years)	0.04	0.22	-0.02	0.15	0.01	0.01	-0.01	1.00	0.75	0.02	0.04	-0.06	0.04	0.03	-0.04	0.09
IUD	0.06	0.28	-0.01	0.20	0.03	0.11	-0.02	0.75	1.00	0.01	0.08	0.00	0.03	0.07	-0.06	0.12
Smokes (packs/year)	0.11	0.13	-0.06	0.10	0.17	0.04	-0.01	0.02	0.01	1.00	0.05	0.00	0.72	0.05	0.49	0.08
STDs: Time since last diagnosis	0.04	0.12	0.02	0.09	0.05	0.06	0.17	0.04	0.08	0.05	1.00	0.01	0.09	0.94	0.08	0.07
Hormonal Contraceptives	0.03	0.03	-0.01	0.12	0.00	0.39	0.03	-0.06	0.00	0.00	0.01	1.00	-0.01	-0.00	0.00	0.01
Smokes (years)	0.06	0.22	-0.06	0.18	0.18	0.05	0.05	0.04	0.03	0.72	0.09	-0.01	1.00	0.08	0.72	0.03
STDs: Time since first diagnosis	0.04	0.10	0.02	0.07	0.04	0.06	0.16	0.03	0.07	0.05	0.94	-0.00	0.08	1.00	0.05	0.07
Smokes	0.01	0.06	-0.12	0.08	0.24	0.04	0.05	-0.04	-0.06	0.49	0.08	0.00	0.72	0.05	1.00	-0.04
Cancer_risk	0.72	0.13	0.04	0.03	0.03	0.05	0.27	0.09	0.12	0.08	0.07	0.01	0.03	0.07	-0.04	1.00

Figure 3. Illustrates the Pearson correlation plot for the top 15 selected features

Figure 3 presents a Pearson correlation plot for the top 15 selected features. A Pearson correlation is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. The visual analysis reveals the relationships between variables, with coefficients ranging from -1 (negative relationship) to 1 (positive relationship). It can be observed that the feature Dx: HPV is strongly positively correlated with Cancer_risk (0.718944), IUD duration and presence also correlate positively (0.749288), Smoking years and packs/year exhibit a similar trend (0.724320), recent STD diagnoses correlate with the time since the first diagnosis (0.935614), and age has a moderate relationship with pregnancies (0.525892).

2.4. Model training

In the training phase, the dataset is split into testing and training subsets, where 80% of the dataset is considered for training and validation and 20% is considered for testing purposes. The study implements different supervised classifiers to identify the one that exhibited the highest precision and reliability in predictions and risk evaluation. A significant challenge addressed during this model selection process is the class imbalance issue, where one class dominates the other, potentially biasing the model's predictions. To mitigate this, oversampling via the SMOTE is employed. Additionally, different supervised classifiers, such as SVM, random forests, naive Bayes, K-nearest neighbors (KNNs), LightGBM, and AdaBoost, are utilized, as each offering distinct learning patterns with their own advantages. Hyperparameter optimization used the grid search technique to refine and calibrate the models. This process involved systematically exploring a range of hyperparameter values to find the combination that resulted in the best model performance. For each model, hyperparameters, such as C (regularization parameter) and kernel type for SVM, number of trees and maximum depth for random forests, and learning rate and boosting rounds for LightGBM, were tuned to achieve optimal performance. Hyperparameters such as alpha for naive Bayes and K for KNNs were also optimized.

3. RESULTS AND DISCUSSION

The proposed system was designed and developed using Python scripting in the Anaconda environment. This section presents the performance of the proposed system with different classification models. The effectiveness of the proposed computational framework is also shown through a comparative analysis with similar existing research work. The performance of the proposed system is measured in terms of accuracy, precision, recall, F1-score, and the receiver operating characteristic (ROC) curve.

$$\text{Accuracy} = (\text{True positives} + \text{True negatives}) / (\text{Total number of predictions}) \tag{4}$$

Accuracy is the proportion of all correct predictions, i.e. the overall correctness of the model. However, its dependance on correct predictions for all classes can be misleading in case when data set is severely imbalanced (i.e., there are more negative cases than positive cases). In this case, the model only needs to predict all cases as negative to achieve high accuracy.

$$Precision = True\ positives / True\ positives + False\ positives \tag{5}$$

Precision is basically a positive predictive value that measures the proportion of positive predictions that are actually correct. Precision is a good measure of a model's ability to identify positive cases. A high accuracy score means the model does not produce many false positives.

$$Recall = True\ positives / (True\ positives + False\ negatives) \tag{6}$$

Recall is the true positive rate, which measures the proportion of actual positives that are correctly predicted. Performance metrics are a good measure of a model's ability to capture all positive cases. A high recall indicates that the model is not missing many true positive results. The F1 score refers to the harmonic mean of precision and recall metrics, providing a balance between the two when the class distribution is uneven.

$$F1\ score = 2 \times (Precision \times Recall) / (Precision + Recall) \tag{7}$$

The evaluation of the proposed prediction model also considers the ROC curve, which shows the performance of the classification model at all possible thresholds. It is created by plotting the true positive rate (TPR) versus the false positive rate (FPR) at different thresholds. TPR is defined as the proportion of actual positive examples that are correctly predicted, and FPR is the proportion of actual negative examples that are incorrectly predicted. The higher the area under the curve (AUC), the better the model performance.

3.1. Numerical outcome

The outcome statistics presented in Table 2 show the performance of the different popular classification models across several critical metrics, namely accuracy, precision, recall, f1-score, and ROC score. Based on the critical analysis of the accuracy score, it has been identified that the random forest classifier achieved the highest accuracy score at 98.06%. Followed closely by both the SVM and LightGBM classification models, each achieving an accuracy of 97.67%.

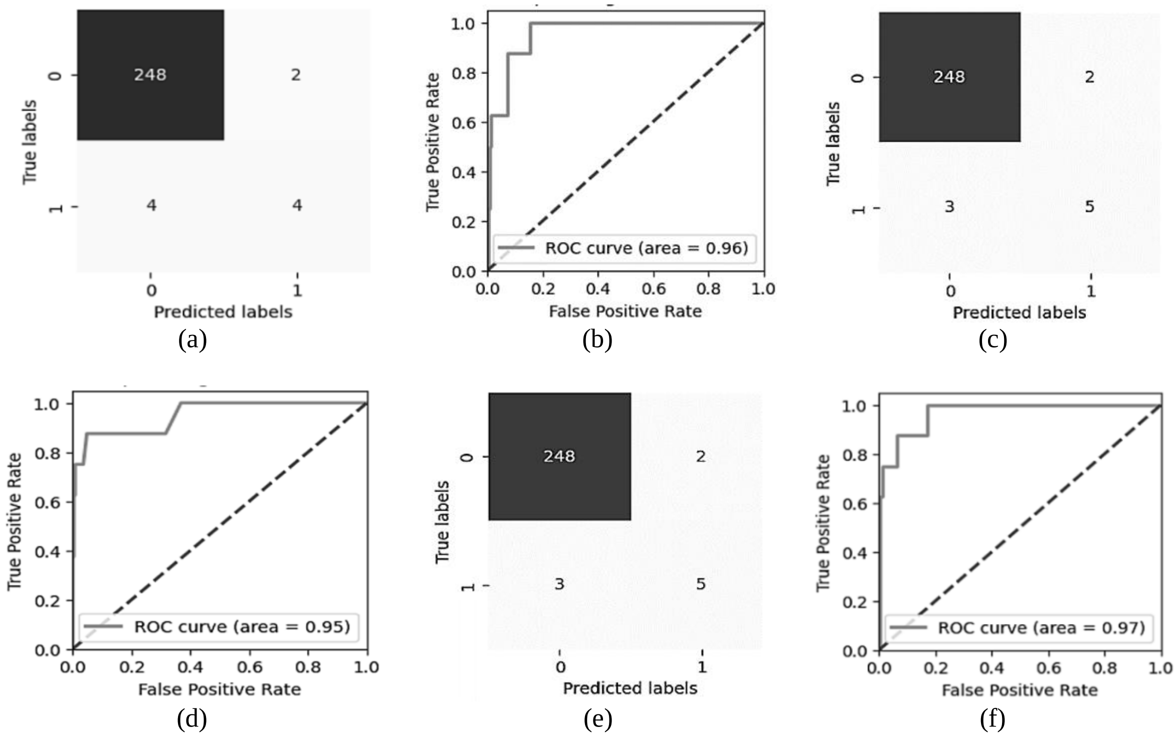
Table 2. Analysis of the numerical outcome statistics for classification model

	Accuracy	Precision	Recall	F1-Score	ROC Score
SVM	0.976744	0.973173	0.976744	0.972961	0.96
Random Forest	0.980620	0.979559	0.980620	0.979994	0.95
Naive Bayes	0.968992	0.972774	0.968992	0.970653	0.85
KNN	0.957364	0.973420	0.957364	0.963550	0.96
LightGBM	0.976744	0.976744	0.976744	0.976744	0.97
AdaBoost	0.965116	0.967207	0.965116	0.966107	0.80

It is to be noted that the higher accuracy indicates the overall correctness of the classification models and its suitability for a given dataset. However, when the dataset suffers from class imbalance problem, performance analysis considering other metrics such as precision and recall becomes a critical issue. The precision captures the model's capability to correctly classify positive instances out of those predicted as positive. Based on the numerical outcome, it can be seen that the random forest model again outperformed with a precision of 97.95%, though the difference with other top-performing models like KNN (97.34%) and SVM (97.31%) is relatively minimal. Here high precision score implies a lower false-positive rate, ensuring that the predictions made are reliable. In terms of recall rate, which reflects the ability of the classification models in identifying all possible positive instances, random forest classifier maintains its superiority with a score of 98.06%. However, both SVM and LightGBM achieved similar performance as well, each presenting a recall of 97.67%. The F1-Score, representing the harmonic mean of precision and recall, is significant when dealing with datasets with imbalanced class distributions. Here, random forest maintains its dominance with an F1 score of 97.99%. Furthermore, LightGBM exhibits unique characteristics with the same F1 score, precision and recall of 97.67%, indicating balanced performance in terms of false positives and false negatives. Finally, the ROC score represents the model's ability to distinguish classes for all thresholds; the classifier LightGBM achieves an ROC score of 97%, outperforming all other classifiers.

3.2. Visual outcome

Figure 4 provides a detailed comparison of the performance metrics for three of the top-performing classification models: SVM, random forest, and LightGBM. For each model, the performance is visualised through its confusion matrix, which illustrates the accuracy of predictions in terms of true positives and true negatives, and the ROC curve, which evaluates the trade-off between sensitivity and specificity. Analysis of Figures 4(a) and 4(b) reveals a large number of true negatives (248) and a very small number of false positives (2). In terms of positive categories, it correctly identified 4 instances, but also had 4 false negatives. This indicates a slight decrease in the sensitivity of the SVM model. The SVM model achieved an ROC score of 96% and successfully distinguished between positive and negative classes at different thresholds. Figures 4(c) and 4(d) analysis shows that the random forest model identified 248 true negatives and only 2 false positives. For the positive class, it accurately predicted 5 instances and only 3 false negatives. This configuration slightly improves sensitivity compared to SVM. Furthermore, the ROC score of the random forest model is 95%, which is almost comparable to SVM. Meanwhile, a careful analysis of Figures 4(e) and 4(f) shows that LightGBM also has 248 true negatives and 2 false positives, similar to the random forest classifier. It has the same accurate predictions, i.e. 5 positive class predictions and 3 false negative class predictions. LightGBM stands out with a ROC score of 97%, slightly higher than SVM and random forest. This excellent ROC score indicates that LightGBM has a better ability to distinguish classes at different thresholds. A high ROC score indicates a model's strength in maintaining a high true positive rate. According to the result statistics, LightGBM models outperform other classification models on all metrics. It achieves the highest accuracy, precision, recall, F1 score, and ROC score.



(a) (b) (c)

(d) (e) (f)

Figure 4. Illustration of visual outcome of top classifiers: (a) SVM confusion matrix, (b) SVM ROC, (c) random forest confusion matrix, (d) random forest ROC, (e) LightGBM confusion matrix, and (f) LightGBM ROC

3.3. Use case scenario

Based on the comprehensive analysis of the results, the LightGBM model emerged as the best model due to its consistency, accuracy, precision, recall and high ROC score. Therefore, for patients at risk of cervical cancer, the probability score from this model can be used to calculate the risk percentage, expressed as in (8).

$$\mathcal{P}(\mathcal{Y}=1)=\frac{1}{1+e^{-\sum \text{raw score}}} \tag{8}$$

Where $\mathcal{P}(\mathcal{Y}=1)$ is the probability that the instance belongs to class 1 (e.g., "at risk"), and the $\sum$ raw score is the sum of the outputs of all the trees for that instance. The (8) is the logistic function, which maps the raw score between 0 and 1, yielding a probability value. For example, if $\mathcal{P}(\mathcal{Y}=1)$ is 0.8, it means that there is an 80% probability, according to the model, that the patient is at risk. In a practical setting, consider a 26-year-old woman who has had 8 sexual partners, been pregnant twice, smokes an average of 25 packs of cigarettes per year, and has had 3 STD diagnoses. Without an HPV diagnosis, the model puts her risk of cervical cancer at 18%. Conversely, an HPV diagnosis increases her risk to an estimated 90 percent. Another example is a 40-year-old woman with 3 sexual partners who started having sex when she was 26 years old. Based on these attributes alone, her risk is only 0.09%; however, an HPV diagnosis increases this risk to 99%. These examples highlight the significant impact of HPV status on cervical cancer risk. Factors that increase the risk of HPV include having multiple sexual partners, having sex too early, concurrent sexually transmitted infections (such as chlamydia, gonorrhea, syphilis, HIV/AIDS), and a compromised immune system. Additionally, another reason is that pre-existing medical conditions, harmful lifestyle habits, or smoking may lead to a weakened immune system, which can also lead to squamous cell cervical cancer.

3.4. Comparative analysis

Figure 5 provides a comparative analysis to demonstrate the effectiveness of the proposed cervical cancer risk prediction model with existing similar research works. The risk prediction model Gaussian naive Bayes used in the study [27] has an accuracy of 81%, precision is 86% and achieved the recall rate with 100%. The ensemble model utilized in [18] exhibited superior accuracy at 95%, boasting a perfect precision score. However, its recall is comparatively lower at 67%. Suman and Hooda [28] adopted decision tree, yielding accuracy of 93%, while the precision and recall were 89% and 96%. Lu *et al.* [29] employed logistic regression and attained an accuracy of 82%, but its precision and recall scores, at 45% and 21%, respectively, are quite lower than other models.

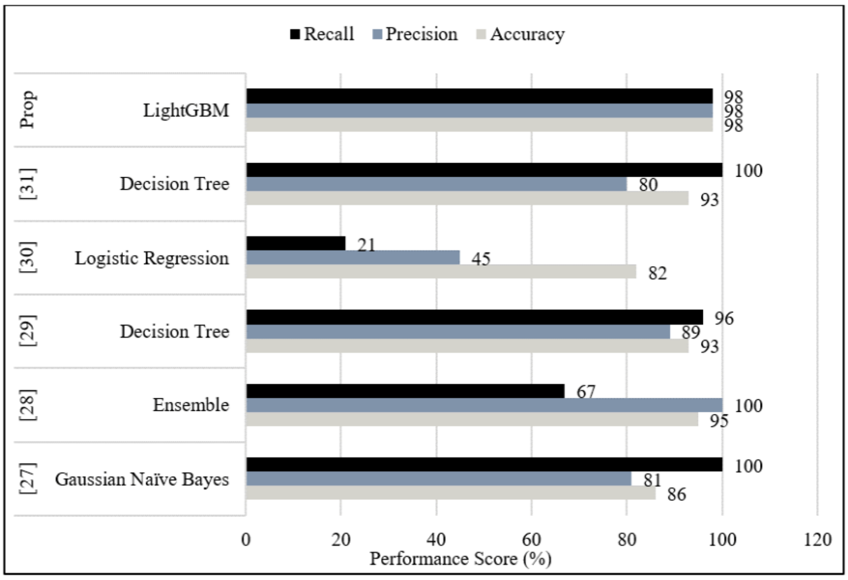


Figure 5. Shows comparative analysis in terms of accuracy, precision, and recall rate

Akter *et al.* [30] using decision tree, showed the same accuracy of 93%, but varied particularly in recall, with a rise to 100%. The proposed risk prediction framework (Prop) harnesses the potential of

LightGBM that outperforms the models mentioned above with accuracy, precision, and recall, all reaching 98%. This highlights the efficacy of the LightGBM model and shows its balanced approach and consistent performance across all metrics, suggesting its suitability as a reliable predicitve model in the propsoed framework for cervical cancer risk prediction.

4. CONCLUSION

This study has presented a robust data-driven approach designed to proactively identify cervical cancer risk, enabling healthcare professionals to make informed decisions about patient care. By applying computational intelligence algorithms to clinical and patient lifestyle data, this research work attempted to improve early detection of cervical cancer and increase diagnostic accuracy and efficiency. This study identified key features and correlations for building an effective cancer prediction model using the LightGBM classifier through efficient analysis and preprocessing of a cervical cancer dataset. The effectiveness of the proposed prediction model is fully verified with an accuracy of up to 98%. By facilitating early and accurate detection of cervical cancer, the proposed risk prediction model enables healthcare professionals to prioritize and fast-track examination and treatment of high-risk patients. This proactive approach to risk analysis and patient care has the potential to significantly improve clinical outcomes and increase the likelihood of successful treatment and recovery. Furthermore, integrating the proposed predictive model into daily healthcare practice is expected to improve patient care standards and optimize the allocation of medical resources. In future, the study extends the scope of the proposed predictive model with sophisticated deep learning model to diagnose cervical cancer using pep smear images.

REFERENCES

[1] M. Arbyn *et al.*, "Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis," *The Lancet Global Health*, vol. 8, no. 2, pp. e191–e203, 2020, doi: 10.1016/S2214-109X(19)30482-6.

[2] K. S. Okunade, "Human papillomavirus and cervical cancer," *Journal of Obstetrics and Gynaecology*, vol. 40, no. 5, pp. 602–608, 2020, doi: 10.1080/01443615.2019.1634030.

[3] J. Huber, A. Mueller, M. Sailer, and P. A. Regidor, "Human papillomavirus persistence or clearance after infection in reproductive age. What is the status? Review of the literature and new data of a vaginal gel containing silicate dioxide, citric acid, and selenite," *Women's Health*, vol. 17, 2021, doi: 10.1177/17455065211020702.

[4] C. A. Burmeister *et al.*, "Cervical cancer therapies: Current challenges and future perspectives," *Tumour Virus Research*, vol. 13, 2022, doi: 10.1016/j.tvr.2022.200238.

[5] R. Mehrotra and K. Yadav, "Cervical Cancer: formulation and implementation of Govt of India guidelines for screening and management," *Indian Journal of Gynecologic Oncology*, vol. 20, no. 1, 2022, doi: 10.1007/s40944-021-00602-z.

[6] L. Allahqoli *et al.*, "Delayed cervical cancer diagnosis: a systematic review," *European Review for Medical and Pharmacological Sciences*, vol. 26, no. 22, pp. 8467–8480, 2022, doi: 10.26355/eurrev_202211_30382.

[7] S. S. Shastri *et al.*, "Secondary prevention of cervical cancer: ASCO resource–stratified guideline update," *JCO Global Oncology*, no. 8, 2022, doi: 10.1200/go.22.00217.

[8] M. Akinlotan, J. N. Bolin, J. Helduser, C. Ojinnaka, A. Lichorad, and D. McClellan, "Cervical cancer screening barriers and risk factor knowledge among uninsured women," *Journal of Community Health*, vol. 42, no. 4, pp. 770–778, 2017, doi: 10.1007/s10900-017-0316-9.

[9] S. Singh and S. Badaya, "Factors influencing uptake of cervical cancer screening among women in india: a hospital-based pilot study," *Journal of Community Medicine & Health Education*, 2012, doi: 10.4172/2161-0711.1000157.

[10] A. B. Bansal, A. P. Pakhare, N. Kapoor, R. Mehrotra, and A. M. Kokane, "Knowledge, attitude, and practices related to cervical cancer among adult women: A hospital-based cross-sectional study," *Journal of Natural Science, Biology and Medicine*, vol. 6, no. 2, pp. 324–328, 2015, doi: 10.4103/0976-9668.159993.

[11] T. U. Zaman *et al.*, "Artificial intelligence: the major role it played in the management of healthcare during COVID-19 pandemic," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, pp. 505–513, 2023, doi: 10.11591/ijai.v12.i2.pp505-513.

[12] J. Jeyshri and M. Kowsigan, "A comprehensive assessment of recent advances in cervical cancer detection for automated screening," *Image Processing and Intelligent Computing Systems*, pp. 171–184, 2022, doi: 10.1201/9781003267782-11.

[13] W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, "Data processing and text mining technologies on electronic medical records: A review," *Journal of Healthcare Engineering*, vol. 2018, 2018, doi: 10.1155/2018/4302425.

[14] A. Kalantari, A. Kamsin, S. Shamshirband, A. Gani, H. A. -Rokny, and A. T. Chronopoulos, "Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions," *Neurocomputing*, vol. 276, pp. 2–22, 2018, doi: 10.1016/j.neucom.2017.01.126.

[15] 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, 2020, doi: 10.3390/s20102809.

[16] U. K. Lilhore *et al.*, "Hybrid model for detection of cervical cancer using causal analysis and machine learning techniques," *Computational and Mathematical Methods in Medicine*, vol. 2022, 2022, doi: 10.1155/2022/4688327.

[17] E. R. Putri, A. Zarkasi, P. Prajitno, and D. S. Soejoko, "Artificial neural network for cervical abnormalities detection on computed tomography images," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 1, pp. 171–179, 2023, doi: 10.11591/ijai.v12.i1.pp171-179.

[18] F. Curia, "Cervical cancer risk prediction with robust ensemble and explainable black boxes method," *Health and Technology*, vol. 11, no. 4, pp. 875–885, 2021, doi: 10.1007/s12553-021-00554-6.

[19] N. Youneszade, M. Marjani, and C. P. Pei, "Deep learning in cervical cancer diagnosis: architecture, opportunities, and open research challenges," *IEEE Access*, vol. 11, pp. 6133–6149, 2023, doi: 10.1109/ACCESS.2023.3235833.

[20] P. Dhivya, M. Karthiga, A. Indirani, and T. Nagamani, “Cervical cancer prediction using optimized meta-learning,” *Lecture Notes in Networks and Systems*, vol. 565, pp. 393–401, 2023, doi: 10.1007/978-981-19-7455-7_30.

[21] T. Zhou, Y. Tang, L. Gong, H. Xie, M. Shan, and L. Wang, “MIC model for cervical cancer risk factors deep association analysis,” *Computational Data and Social Networks*, vol. 13116, pp. 147–155, 2021, doi: 10.1007/978-3-030-91434-9_14.

[22] S. Priya and N. K. Karthikeyan, “Deep learning classification to improve diagnosis of cervical cancer through swarm intelligence-based feature selection approach,” *Intelligent Systems, Technologies and Applications*, pp. 247–264, 2021, doi: 10.1007/978-981-16-0730-1_17.

[23] Z. K. Şentürk and S. Uzun, “An improved deep learning based cervical cancer detection using a median filter-based preprocessing,” *European Journal of Science and Technology*, 2022, doi: 10.31590/ejosat.1045538.

[24] N. A. Mudawi and A. Alazeb, “A model for predicting cervical cancer using machine learning algorithms,” *Sensors*, vol. 22, no. 11, 2022, doi: 10.3390/s22114132.

[25] J. J. Tanimu, M. Hamada, M. Hassan, H. A. Kakudi, and J. O. Abiodun, “A machine learning method for classification of cervical cancer,” *Electronics*, vol. 11, no. 3, 2022, doi: 10.3390/electronics11030463.

[26] K. Fernandes, J. Cardoso, and J. Fernandes, “Cervical cancer (risk factors),” *UCI Machine Learning Repository*, 2017. Accessed: Oct 4, 2023. [Online]. Available: http://archive.ics.uci.edu/ml.

[27] I. J. Ratul, A. A. -Monsur, B. Tabassum, A. M. Ar-Rafi, M. M. Nishat, and F. Faisal, “Early risk prediction of cervical cancer: A machine learning approach,” *19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2022*, 2022, doi: 10.1109/ECTI-CON54298.2022.9795429.

[28] 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, doi: 10.1080/09720510.2019.1611227.

[29] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud, “Machine learning for assisting cervical cancer diagnosis: An ensemble approach,” *Future Generation Computer Systems*, vol. 106, pp. 199–205, 2020, doi: 10.1016/j.future.2019.12.033.

[30] L. Akter, F. A.-Islam, M. M. Islam, M. S. A. -Rakhami, and M. R. Haque, “Prediction of cervical cancer from behavior risk using machine learning techniques,” *SN Computer Science*, vol. 2, no. 3, 2021, doi: 10.1007/s42979-021-00551-6.

BIOGRAPHIES OF AUTHORS

Sreelatha holds the Bachelors and Masters degree from Visvesvaraya Technological University. She is currently working as an assistant professor in Department of Computer Science at Presidency University, Bangalore. Her research area includes image processing and machine learning. She has published over 5 papers in international journals and conferences. She can be contacted at email: sreelatha.sajeev2@gmail.com or sreelatha.pk@presidencyuniversity.in

Dr. Vrinda Shivashetty received the Doctorate degree in Computer Science and Engineering from Gulbarga University. She is currently working as a Professor in Department of Information Science and Engineering at Sai Vidya Institute of Technology, Bangalore. She is recipient of many awards like “Emerging Leader in Engineering” and “Outstanding Women in Engineering” powered by Venus International Foundation in 2018 at Chennai, “Senior Woman Educator & Scholar award” powered by NFED Coimbatore in 2017, and also “Excellent Teacher” award 5 times for achieving 100% results in university exams. She can be contacted at email: hodise@saividya.ac.in