



# Interpretable artificial intelligence (AI) for cervical cancer risk analysis leveraging stacking ensemble and expert knowledge

Priyanka Roy<sup>1,2,\*</sup>, Mahmudul Hasan<sup>1,\*</sup>, Md Rashedul Islam<sup>1</sup>  
and Md Palash Uddin<sup>1</sup> 

## Abstract

**Objectives:** This study develops a machine learning (ML)-based cervical cancer prediction system emphasizing explainability. A hybrid feature selection method is proposed to enhance predictive accuracy and stability, alongside evaluation of multiple classification algorithms. The integration of explainable artificial intelligence (XAI) techniques ensures transparency and interpretability in model decisions.

**Methods:** A hybrid feature selection approach combining correlation-based selection and recursive feature elimination is introduced. An ensemble model integrating random forest, extreme gradient boosting, and logistic regression is compared against eight classical ML algorithms. Generative artificial intelligence methods, such as variational autoencoders and generative teaching networks, were evaluated but showed suboptimal performance. The research integrates global and local XAI techniques, including individual feature contributions and tree-based explanations, to interpret model decisions. The effects of feature selection and data balancing on classification performance are examined to stabilize precision, recall, and F1 scores. Classical ML models without preprocessing achieve 95–96% accuracy but exhibit instability.

**Results:** The proposed feature selection and data balancing strategies significantly enhance classification stability, creating a robust predictive model. The ensemble model achieves 98% accuracy with an area under the curve of 99.50%, outperforming other models. Domain experts validate critical contributing features, confirming practical relevance. Incorporating domain knowledge with XAI techniques significantly increases transparency, making predictions interpretable and trustworthy for clinical use.

**Conclusion:** Hybrid feature selection combined with ensemble learning substantially improves cervical cancer prediction accuracy and reliability. The integration of XAI techniques ensures transparency, supporting interpretability and trustworthiness, demonstrating significant potential in clinical decision-making.

## Keywords

cervical cancer, feature selection, machine learning, explainable AI, domain knowledge

Received: 27 October 2024; accepted: 18 February 2025

## Introduction

Cancer is a complex and diverse group of diseases characterized by the abnormal and uncontrollable growth of cells, often leading to the formation of malignant tumors capable of infiltrating nearby tissues and metastasizing to other areas of the body.<sup>1</sup> The cervix is a vital, pear-shaped organ that connects the lower part of a woman's uterus to the vagina. Cervical cancer is one of the deadliest cancers occurring in the cells of the cervix, mostly caused by the human papillomavirus (HPV). According to the annual report

<sup>1</sup>Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh

<sup>2</sup>Computer Science and Engineering, Sylhet International University, Sylhet, Bangladesh

\*These authors equally contributed to this work.

### Corresponding author:

Md Palash Uddin, Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh.  
Email: palash\_cse@hstu.ac.bd



of the World Health Organization (WHO), it is the fourth-leading cancer among women around the globe.<sup>2</sup> The documented data for the year 2020 indicates an estimated 6,04,000 new cases and 3,42,000 deaths.<sup>3</sup> Moreover, it revealed a substantial increase in the number of new global cases, surpassing 700,000 within a mere 12 years from 2018 to 2030, representing a significant growth from the current figure of 570,000. An alarming 90% of these new cases and deaths were reported in low- and middle-income countries. This highlights the disproportionate impact of the disease in these regions, as the mortality rate for cervical cancer is increasing annually at a significant rate.<sup>4</sup> This is due to the lack of regular screening programs, limited healthcare access, and awareness about early-stage risk factors. In Bangladesh, cervical cancer is the second most common cancer in females. In the absence of a proper intervention, a cumulative total of 505,703 women in Bangladesh are projected to die of cervical cancer by the year 2070. Furthermore, this figure is expected to increase to 1,042,859 by the year 2120.<sup>5</sup>

The alarming statistics necessitate urgent attention and calls for immediate action to ensure early detection of cervical cancer. When detected in the early stage, this fatal disease is highly treatable. Regular screenings, pap smears, HPV tests, and cervical biopsies are effective in detecting abnormalities in the cervical cell wall. Although the expenses and discomfort associated with these surgical processes emphasize the necessity for more cost-effective and reliable approaches to the detection of cervical cancer. Identifying the risk factors for this cancerous condition also eliminates suffering and distress and improves the quality of life. Multiple research studies have demonstrated that early detection of cervical cancer greatly enhances the chances of a patient's swift recovery.<sup>6</sup> Therefore, by identifying the risk factors, healthcare providers can significantly improve the prognosis of individuals at risk of developing cervical cancer.

### *Applications of machine learning in ensuring digital healthcare*

Artificial intelligence and machine learning (AI–ML) has become a transformative tool in the healthcare sector, enabling more accurate diagnoses, personalized treatments, and efficient healthcare management.<sup>7</sup> By leveraging different sizes of medical data, dealing with inherent discrepancies, and handling challenges imposed by real-world inconsistent data, ML algorithms assist healthcare professionals in making informed decisions, reducing errors, and improving patient outcomes. One of the primary aspects of the application of ML in healthcare is designing optimized algorithms for the early diagnosis of diseases and highlighting the key influencing factors.<sup>8</sup>

Additionally, ML models, particularly deep learning algorithms, have demonstrated exceptional performance in diagnosing diseases such as cancer, cardiovascular conditions, and brain-related disorders. Image-based diagnostics

using advanced computer vision algorithms have significantly improved radiology, pathology, and dermatology by providing automated and highly accurate image analysis and segmentation techniques. Furthermore, beyond diagnostics and personalized treatment planning, AI–ML is extensively being utilized in interdisciplinary research in biomedical, health informatics, and drug discovery.<sup>9</sup> As ML technology continues to evolve, its integration into healthcare holds immense potential to enhance medical practice and improve overall public health.

### *AI-powered predictive and histopathological analysis for cervical cancer*

AI plays a critical role in cervical cancer detection and classification, offering innovative solutions to enhance the accuracy, efficiency, and accessibility of diagnostic procedures.<sup>10</sup> Recent advances in AI hold tremendous potential for improving the automated detection of cervical pre-cancer and cancer, leading to more accurate and objective outcomes. Moreover, AI has significant potential in recognizing molecular markers linked to cervical cancer, facilitating its diagnosis. ML, a subset of AI, is a powerful tool to identify underlying patterns in input data and discover key indicators of cervical cancer.<sup>11</sup> The demand for applying cutting-edge computational models such as AI and ML has greatly increased due to recent advancements in this field.

ML algorithms are highly effective in handling imbalanced datasets, a prime concern in the healthcare industry where instances are unevenly distributed across target classes. An imbalanced dataset can potentially bias toward the majority class, resulting in the deterioration of the prediction model's performance.<sup>12</sup> AI and ML algorithms have demonstrated remarkable potential in enhancing diagnostic precision and forecasting patient outcomes by addressing such real-world imbalanced data. Consequently, the healthcare sector actively incorporates these technologies into various areas of patient care and administrative procedures to identify high-risk individuals and develop personalized intervention strategies. This can ultimately lead to earlier detection and more effective treatment of cervical cancer.

Conversely, explainable AI (XAI) methods break down the complex decision-making processes of various sophisticated ML models to ensure transparency and interpretability.<sup>13,14</sup> Different XAI methods determine the key factors influencing the prediction model, revealing which features are more important for accurate diagnosis. This ensures the understandability and trustworthiness of the ML models. While most studies primarily concentrate on enhancing classification accuracy through various advanced techniques contributing to feature selection and model development, some research has shifted its focus to XAI for elucidating black-box models using diverse XAI tools.

However, these studies often overlook the integration of domain experts' opinions and their alignment with XAI

explainability. This gap can lead to mismatches in certain cases, necessitating efforts to minimize such discrepancies and gain the trust of domain experts in AI tools. The identified gaps in existing studies motivate us to propose a new direction in research, aiming to enhance transparency for both users and experts in this field.

The primary focus of this investigation is to establish a connection between traditional healthcare methods and advanced AI through XAI. The goal is to address the outlined challenges by leveraging the extensive capacities of ML. To achieve this, a cutting-edge ML system integrated with XAI and domain knowledge is proposed. The ML system aims to make accurate predictions and explain its decisions through XAI techniques. The system improves its interpretability and trustworthiness in critical decision-making processes by incorporating domain knowledge. Furthermore, integrating traditional healthcare methods with advanced AI and XAI can revolutionize patient care and treatment by providing valuable insights and explanations for medical practitioners. This innovative approach can enhance diagnostic accuracy, optimize treatment plans, and ensure that medical decisions are transparent and understandable to healthcare professionals and patients. Ultimately, this research aims to bridge the gap between healthcare and AI, paving the way for more informed and efficient healthcare delivery.

## Contributions

The main contributions of this study are summarized as follows:

- We propose and design an ML-based cervical cancer risk prediction system, integrating domain knowledge to enhance black-box model explainability for secondary and primary data.
- We propose a novel hybrid feature selection process combining filter and wrapper methods and introduce a comprehensive preprocessing pipeline that addresses missing values, detects and removes outliers, and handles imbalanced data, resulting in improved outcomes for the ML models to ensure optimal performance.
- We design a stacking ensemble model to predict cervical cancer with enhanced accuracy and stability, as measured by the precision, recall, and F1 score for individual classes.
- We integrate two generative AI models in the prediction system to compare the prediction capability of the proposed ML model with generative AI approaches for tabular data.
- We employ an exploratory data analysis (EDA) system and XAI methods, providing global and local explainability to uncover the underlying patterns within the dataset.

- We validate the efficacy of XAI explainability by conducting surveys with domain experts, establishing a robust relationship at different levels of understanding.

The paper is structured as follows: Section “Related studies” presents the related literature, Section “Methodology” outlines the methodology, Section “Results” discusses the results, and Section “Conclusion and future work” concludes the findings while pinpointing some future research directions.

## Related studies

Cervical cancer detection and prediction have witnessed significant advancements by integrating ML techniques. ML models analyze vast amounts of patient data, assisting in early detection and prognosis. Consequently, numerous researchers have focused on using diverse images and numerical datasets to predict and detect the onset of cervical cancer in its initial phases. Wu and Zhou<sup>15</sup> applied a support vector machine (SVM) classifier to the “Cervical Cancer (Risk Factors)” dataset.<sup>15</sup> The dataset comprises 858 subjects, each with 36 unique attributes. The authors combined SVM with feature selection techniques such as recursive feature elimination (RFE) and principal component analysis (PCA) to identify the top 10 risk factors contributing most to the model’s prediction outcomes. The experimental results showed the superiority of traditional SVM over SVM-RFE and SVM-RF.

Subsequently, Hariprasad et al.<sup>16</sup> utilized the extreme gradient boosting (XGB) model on the same dataset to develop an innovative approach for predicting cervical cancer risk.<sup>16</sup> They extracted the most prominent features and concluded that the random forest (RF) was the best performer, achieving 98.7% accuracy after applying the oversampling technique. Ensemble models have also demonstrated encouraging results in the early prediction of cervical cancer. One study utilized several ensemble methods, including bagging, boosting, and stacking.<sup>17</sup> The results indicate that using a combination of RF, SVM, ExtraTreeClassifier, XGB, and bagging in a stacked ensemble produces the highest classification accuracy of 94.4%. Another recent study, in 2024, proposed such an ensemble model with extensive preprocessing techniques to identify the presence of cervical cancer.<sup>18</sup> This study aimed to enhance the predictive power of the ML model by introducing an RF-based ensemble model. They incorporated two different datasets, namely the “Cervical Cancer (Risk Factors)” dataset from the UCI ML repository and the Behavioral risk dataset (sobar-72) to achieve accuracies of 98.06% and 95.45%, respectively.<sup>19,20</sup> Ilyas and Ahmad<sup>21</sup> implemented a voting ensemble classification approach that achieved 94% accuracy, surpassing other models such as decision tree (DT), SVM, KNN, RF, Naïve Bayes, etc.<sup>21</sup> They explored both the aforementioned behavioral risk dataset and the earlier UCI dataset. Researchers conducted a meta-analysis of available

literature concerning ML-driven cervical cancer diagnosis, particularly emphasizing ensemble learning methodologies. Kumawat<sup>22</sup> applied six traditional ML algorithms, including an artificial neural network, a Bayesian network, an SVM, a random tree (RT), a logistic tree, and an XGB tree.<sup>22</sup> Although the authors applied three different feature selection algorithms, namely relief rank, wrapper method, and least absolute shrinkage and selection operator regression, the maximum accuracy they could achieve is only 94.94% utilizing all 36 features. This significantly increased the overall model training time and, therefore, model overhead.

Segmentation of cervical cell images has proven to be effective in diagnosing cervical cancer, allowing the tracking of changes in cell walls and lesions. In a study by Song et al.,<sup>23</sup> a dynamic multitemplate deformation model was designed to segment each cell from pap-smear images.<sup>23</sup> One major challenge in segmentation tasks is the issue of cell overlapping. Lee and Kim<sup>24</sup> proposed a superpixel partitioning model combined with the triangle threshold for automatically segmenting multiple overlapping cervical cells, demonstrating competitive results compared to other approaches.<sup>24</sup> Another study integrated a data-mining approach with traditional ML algorithms.<sup>25</sup> The authors proposed a two-level cascade integration system combining CR4.5 and logistic regression (LR), achieving an abnormal cell recognition rate of 95.642%, surpassing individual classifiers. Devi et al.<sup>26</sup> implemented a graph-cut-based segmentation model for efficient cervical cancer detection. Applying their method to the Herlev dataset improved overall accuracy by 5.24%.<sup>27</sup>

In 2021, a statistical analysis utilized instance-based K-nearest neighbor (KNN), K-Star, split-point and attribute reduced classifier (SPAARC), RT, and RF algorithms.<sup>28</sup> The dataset was collected from both Kaggle and UCI ML repositories, and RF performed best, achieving 98.33% accuracy. In a recent study in 2022, hyperparameter tuning was conducted for early risk prediction of cervical cancer using grid search cross-validation.<sup>29</sup> The performance of various models, including KNN, DT, SVM, RF, and multilayer perceptron (MLP), was thoroughly analyzed, resulting in an accuracy of 93.33%. Kumari et al.<sup>30</sup> applied PCA to overcome the potential problems imposed by the high-dimensional data.<sup>30</sup> After reducing the size of the dataset, they fed the input data into a deep neural network (DNN) classifier. The experimental investigation secured a maximum accuracy of 87.4%.

The predominant focus of research in this domain revolves around improving classifier performance by integrating various preprocessing techniques. Glučina et al.<sup>31</sup> explored different class balancing techniques to tackle the challenge imposed by employing an imbalanced dataset.<sup>31</sup> They achieved the highest area under the curve (AUC) score of 0.95 utilizing MLP and KNN combined with random oversampling, synthetic minority over-sampling technique (SMOTE), and SMOTE Tomek (SMOTETomek). Some studies also explore the elucidation of black-box models using XAI tools, shedding light on diverse aspects of cervical cancer. A study by AlMohimeed et al.<sup>32</sup> introduced a stacked

ensemble model for predicting early stage cervical cancer.<sup>32</sup> The authors employed different wrapper and embedded approaches for feature selection, optimized using Bayesian optimization, and finally applied Shapley additive explanations (SHAP) for global explainability, achieving 99.44% accuracy. To ensure model transparency at both local and global scales, another group of researchers incorporated local interpretable model-agnostic explanations (LIME) alongside SHAP.<sup>33</sup> Although the proposed ensemble model achieved 94.5% accuracy, the authors tried to explain clinical observations and interpret features. However, there remains an opportunity to enhance classifier performance further. It is essential in health informatics to craft models that demonstrate superior precision and recall. Consequently, an immediate need exists to align global explainability with local explainability. Equally vital is forging a link between the perspectives of domain experts and the results of XAI tools. This approach fosters trust from domain experts and maximizes the benefits of such studies.

## Methodology

### Approach overview

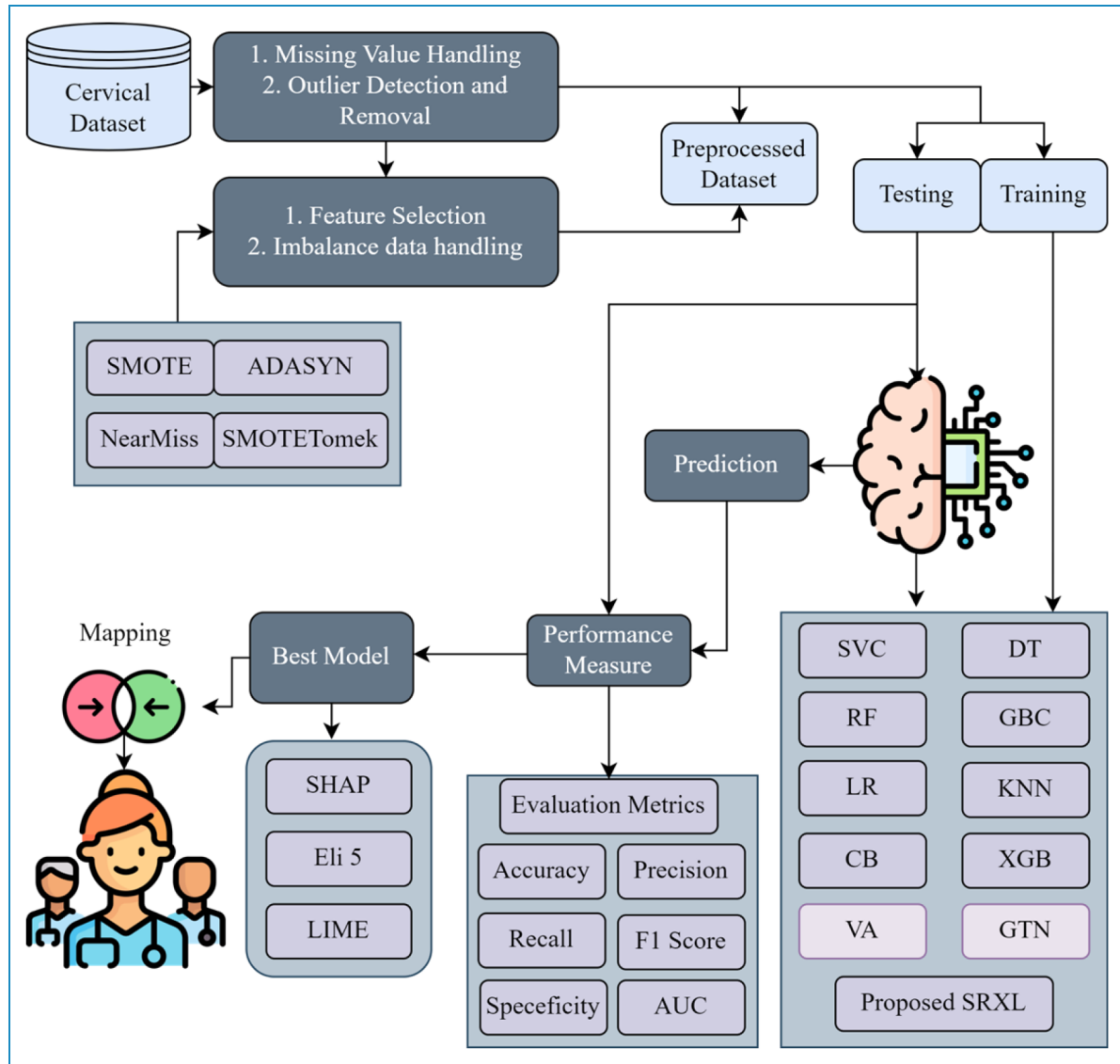
In this study, we employ baseline ML models and introduce a stacking ensemble model for the prediction of cervical cancer. The dataset used is sourced from the UCI ML repository, and Figure 1 outlines our two-phase data preprocessing and balancing technique to ensure data consistency and quality.

To make the dataset suitable for ML training, we perform basic preprocessing steps such as handling missing values, detecting and removing outliers, etc. We apply over-sampling and under-sampling methods to address the challenge of imbalanced classes and investigate the impact of different sampling techniques. The top features are selected using our proposed hybrid feature selection mechanism, which combines filter and wrapper methods. The final preprocessed data is split into training and testing sets and fed into the ML classifiers. Subsequently, we employ XAI tools such as SHAP, ELI5, and LIME to align the results with domain experts and determine the significance of individual features, establishing their impact on the classifiers.

The nature of this study is retrospective and analytical, utilizing publicly available cervical cancer datasets. Additionally, for domain knowledge-driven explainability, this method collects data from the domain experts. This study was conducted from January 2024 to April 2024 in Bangladesh. Most of the data are collected virtually through a Google form by some questionnaire and we meet some of the respondents face to face and collect the information.

### Description of the dataset

The dataset employed in this study is the “Cervical Cancer (Risk Factors)” dataset, sourced from the UCI ML



**Figure 1.** Overview of the proposed methodology, including data processing, explainable artificial intelligence (AI) tools, and model evaluations.

repository.<sup>19</sup> Collected from the “Hospital Universitario de Caracas” in Caracas, Venezuela, this dataset encompasses detailed medical records of 858 patients. It comprises 32 attributes and includes four target classes: Hinselmann, cytology, Schiller, and biopsy. Biopsy involves microscopic examination to determine the presence of abnormal cells. In this study, the use of biopsy as the target variable indicates the presence of cervical cancer based on biopsy samples. The target variable comprises two classes: positive, indicating the presence of cervical cancer, and negative, indicating its absence. Table 1 concisely describes the dataset.

## Data preprocessing pipeline

This research involves a two-phase data processing approach. Initially, we undertake fundamental preprocessing tasks,

addressing issues such as handling missing values and detecting outliers. In the subsequent phase, we introduce a novel feature selection process to identify the essential features within the dataset. To evaluate the impact of various data balancing techniques, we employ SMOTE<sup>34</sup> and adaptive synthetic (ADASYN)<sup>35</sup> as oversampling methods, NearMiss as the undersampling method,<sup>36</sup> and the combined over and undersampling technique SMOTETomek.<sup>37</sup> A systematic evaluation of each algorithm is conducted to assess its effectiveness.

**Handling missing values.** Handling missing values in datasets is crucial to prevent biased results and maintain the precision of ML algorithms. In this study, we substitute the mean for continuous variables and the median for discrete values to address missing values.



**Table 1.** Description of the detailed attributes of the dataset.

SL.	Attribute name	Description	Type
1	Age	Indicates the age of the patient (woman)	Int
2	Number of sexual partners	Expresses the total count of people with whom the woman is engaged in sexual activity together	Int
3	First sexual intercourse	Denotes the age at which the woman had her first sexual intercourse	Int
4	Number of pregnancies	Denotes the total number of childbirths by the woman	Int
5	Smokes	Either the woman smokes (one) or not (zero)	Bool
6	Smokes (years)	Denotes the total smoking period (in years)	Int
7	Smokes (packs/year)	Depicts the annual count of cigarette packets consumed	Int
8	Hormonal contraceptives	Indicates whether the woman takes hormonal contraceptives or not	Bool
9	Hormonal contraceptives (years)	It indicates the duration for which the contraceptive method has been used	Int
10	IUD	Indicates whether the intrauterine contraceptive device was used (one) or not (zero)	Bool
11	IUD (years)	Denotes the total usage period of IUD (in years)	Int
12	STDs	Indicates the presence of STDs—either yes (one) or no (zero)	Bool
13	STDs (number)	This numerical feature shows the overall count of STDs detected in the patient	Int
14	STDs: Condylomatosis	Presence of condylomatosis with the patient—expressed in ones (yes) and zeros (no)	Bool
15	STDs: cervical condylomatosis	Presence of cervical condylomatosis—expressed in ones (yes) and zeros (no)	Bool
16	STDs: vaginal condylomatosis	Presence of vaginal condylomatosis—expressed in ones (yes) and zeros (no)	Bool
17	STDs: vulva-perineal condylomatosis	Presence of vulvo-perineal condylomatosis—expressed in ones (yes) and zeros (no)	Bool
18	STDs: syphilis	Presence of syphilis— expressed in ones (yes) and zeros (no)	Bool
19	STDs: pelvic inflammatory disease	Presence of pelvic inflammatory diseases—expressed in ones (yes) and zeros (no)	Bool
20	STDs: genital herpes	Presence of genital herpes—expressed in ones (yes) and zeros (no)	Bool
21	STDs: molluscum contagiosum	Presence of molluscum contagiosum—expressed in ones (yes) and zeros (no)	Bool
22	STDs: AIDS	Presence of AIDS—expressed in ones (yes) and zeros (no)	Bool
23	STDs: HIV	Presence of HIV infection—expressed in ones (yes) and zeros (no)	Bool

(continued)

Table 1. Continued.

SL	Attribute name	Description	Type
24	STDs: Hepatitis B	Presence of Hepatitis B virus infection in the patient—expressed in ones (yes) and zeros (no)	Bool
25	STDs: HPV	Presence of HPV in the patient—expressed in ones (yes) and zeros (no)	Bool
26	STDs: Number of diagnoses	Indicates the total number of times the STDs have been diagnosed	Int
27	STDs: Time since first diagnosis	Indicates the total number of years since the first diagnosis	Int
28	STDs: Time since last diagnosis	Specifies the length of time since the last diagnosis	Int
29	Dx:Cancer	Indicates the presence of cancer after the diagnose	Bool
30	Dx:CIN	Indicates the presence of CIN after the diagnosis	Bool
31	Dx:HPV	Indicates the presence of HPV after the diagnosis	Bool
32	Dx	It indicates the presence of cancer, CIN, or HPV	Bool
33	Hinselmann	Hinselmann or colposcopy is an intensive medical test that uses magnification to accurately identify abnormalities in cervical cells, vagina, and vulva	Bool
34	Schiller	Schiller iodine test is a medical test in which iodine solution is applied to the cervix to diagnose cervical cancer	Bool
35	Cytology	It is the PaP smears test which helps to detect abnormal cells in the cervix	Bool
36	Biopsy (TARGET)	A surgical procedure to examine a small amount of tissue removed from the cervix to determine whether a carcinogenic cell is present (one) or not (zero)	Bool

AIDS: acquired immune deficiency syndrome; CIN: cervical intraepithelial neoplasia; HIV: human immunodeficiency virus; HPV: human papillomavirus; IUD: intrauterine device; STD: sexually transmitted disease.

**Outlier detection.** Detecting outliers is crucial in dataset analysis, as these anomalies can significantly impact the effectiveness of ML algorithms.<sup>38</sup> One statistical method for outlier detection is the interquartile range (IQR). This technique calculates the IQR, representing the range between the first quartile (Q1) and the third quartile (Q3). Emphasizing the central tendency of the data, IQR outlier detection provides a measure of variability that is less susceptible to the influence of extreme values than some other statistical approaches.

**Handling imbalance data.** When dealing with imbalanced datasets, where classification categories are not evenly distributed, the performance of a model may be compromised. Resampling techniques, such as oversampling and under-sampling, are frequently employed to overcome this

challenge. In this research, three methods—SMOTE, NearMiss, and the SMOTETomek ensemble method—address imbalances in the data.<sup>37</sup>

SMOTE acts as a data augmentation method, creating synthetic data samples for the minority class through interpolation among existing data samples.<sup>34</sup> ADASYN is a data balancing technique designed to generate synthetic examples for the minority class, particularly focusing on skewed class distributions.<sup>35</sup> NearMiss, on the other hand, as an undersampling technique, identifies samples from the majority class that are most similar to those from the minority class and retains only a subset of those samples.<sup>36</sup> The SMOTETomek approach combines under- and oversampling strategies. SMOTE is employed to oversample the minority class, while Tomek Links—a method for finding pairs of closely related samples from different classes—is

used to delete samples from the majority classes.<sup>37</sup> This research aims to improve the classification model's performance and accuracy by addressing the challenges posed by imbalanced datasets through various resampling strategies.

**Feature selection techniques.** This is the most significant phase in the study since it contributes to decreasing the dimensionality of the data, which improves model performance because only relevant data is utilized, and the model can more accurately represent the underlying patterns. To identify the optimal features, we devise a streamlined hybrid feature selection technique that integrates correlation-based feature selection (a filter method) with subsequent RFE (a wrapper method).<sup>39,40</sup> Algorithm 1 represents the proposed hybrid feature selection method.

### Inclusion and exclusion criteria

We list the inclusion and exclusion criteria of this study below.

**Algorithm 1** Hybrid feature selection hfs

<b>function</b> FEATURE_SELECTION
<b>Set</b> corr_threshold $\alpha$
<b>Set</b> corr_features
<b>Find</b> corr_matrix
<b>for</b> i in corr_matrix.columns <b>do</b>
<b>for</b> j in i <b>do</b>
<b>if</b> (corr_matrix.iloc[i,j]) $\geq \alpha$ <b>then</b>
colname=correlation_matrix.columns[i]
correlated_features.add(colname)
<b>end if</b>
Select filtered_columns
<b>end for</b>
<b>end for</b>
Create object rf for Random Forest Classifier
Create selector for feature selection using rf
<b>return</b> selected_features
<b>end function</b>

### Inclusion criteria

- Patients with cervical cancer-related medical records are available in the dataset.
- Complete records with sufficient feature availability (i.e., essential clinical, demographic, and diagnostic variables required for model training).
- Subjects with labeled outcomes (e.g. diagnosed with or without cervical cancer) for supervised learning.
- We have collected data from the responses through face-to-face meetings and online survey questionnaires with the respondent's proper consent.

### Exclusion criteria

- Patients with missing or incomplete data for key variables such as age, medical history, or risk factors. We handle the missing values with proper methods.
- We have handled the imbalanced data and removed the inconsistent information from the dataset before analysis.

### ML classifiers

We use several state-of-the-art ML algorithms from secondary data to predict the risk of cervical cancer. We employ various probabilistic, distance-based, and tree-based ML models in this study to make predictions and evaluate their performance. The algorithms were chosen based on meticulous considerations to comprehensively explore the full spectrum of ML techniques and identify the optimal approach for predicting the presence of cervical cancer. The details of the models are as follows:

**Existing ML classifier.** Support vector classifier (SVC) is a commonly employed ML method utilized for both regression and classification tasks.<sup>41</sup> To classify data points, it employs hyperplanes in  $n$ -dimensional space, where support vectors denote the nearest points to the hyperplane.<sup>42</sup> Another widely used ML algorithm is the DT.<sup>43</sup> It works by recursively partitioning the dataset based on features to generate a tree-like decision structure where each internal node corresponds to a feature, each branch to a decision rule, and each leaf node to the final output or decision. LR is a statistical classification method.<sup>44</sup> It converts a linear combination into a probability value between 0 and 1 and predicts the probability of a binary result by using predictor variables and a logistic function. KNN is a classification technique that attempts to classify an unknown sample using the categorization of its neighbors based on various distance matrices.<sup>45</sup> However, the RF algorithm makes predictions by combining several decision trees.<sup>46</sup> To limit variance and avoid overfitting, it makes use of aggregation and bootstrapping.<sup>47</sup> Gradient boosting classifier (GBC) is



an ensemble method that optimizes a loss function using gradient descent.<sup>48</sup> The CatBoost classifier is another ML technique that is effective for the prediction of categorical features.<sup>49</sup> Binary decision trees are used as the foundation predictors in the gradient boosting technique CatBoost.<sup>50</sup> XGB is a scalable ensemble method that employs the gradient boosting concept and performs better because it uses a more regularized model formalization.<sup>48,51</sup>

**Generative AI approaches.** Generative models are ML algorithms designed to generate new data samples that resemble a given training dataset. These algorithms generate synthetic data samples without losing the essence of the real data by learning the underlying distribution of numerical data. Besides, the generative teaching networks (GTNs) represent an innovative approach to enhancing the efficiency and effectiveness of training ML models. This is achieved by generating optimal training data and/or training environments. These DNNs excel at creating synthetic data specifically designed to expedite the learning process of a newly generated neural network.<sup>52</sup>

**Proposed stacking SXRL ensemble classifier.** For better classification, we propose an ensemble classifier using a stacking procedure. Stacking is leveraging predictions from multiple models to construct a new model, which is subsequently directed to utilize these predictions on the test set. RF, XGB, and LR are in the level 0 estimator, and RF is the final estimator in this new proposed model. Stacking outperforms other ensembles by leveraging a meta-model to learn from multiple base model predictions.<sup>53–55</sup> Unlike voting or blending, it captures complex relationships and dependencies among classifiers. By optimizing feature representation through diverse learners, stacking reduces bias and variance, improving generalization and predictive performance over traditional ensemble methods. We have designed different ensembles and show their result in the result section to demonstrate the purpose of choosing the stacking approach in this study. An overview block diagram of this ensemble classifier is in Figure 2.

The primary objective of an ensemble model is to diminish the variance in the data, thereby enhancing its suitability for ML models. Firstly, the LR operates. LR classifies the target as:

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

Here,  $\mu$  is a location parameter and  $s$  is a scale parameter. The expression can be present as,

$$p(x) = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}$$

Here,  $\beta_0 = -\mu/s$  is called the intercept.

We all know that boosting is a sequential ensemble learning technique where each model in the series aims to rectify the errors of its predecessor. Subsequent models are intricately linked to the performance of the preceding model, creating a cumulative improvement in predictive accuracy throughout the boosting process. To get this advantage, we add RF in the stacking classifier to get a good result.

To reduce the computation, prevent overly complex tree creation, and contribute to better model generalization, we add XGB in the stacking process. The framework supports parallel processing, leveraging computational resources effectively for scalability. Then, RF is used as the final estimator for the classification. The pseudocode is in Algorithm 2.

### Hyperparameter tuning using GridSearch

Hyperparameter tuning is a strategy used to determine the best combination of hyperparameters for an ML model. We utilized GridSearch<sup>56</sup> for this purpose, which returned the best configuration settings for our proposed SRXL

**Algorithm 2** Stacking SRXL built upon XGB, RF, and LR

<b>S:</b> Datasets ( <b>S</b> )
<b>M</b> : The number of classifiers ( <b>M</b> = [ <b>LR</b> , <b>XGB</b> , <b>RF</b> ])
<b>R</b> : The parameter that specifies the proportion of samples to be replaced during the training process.
<b>L</b> : The parameter employed to regulate the distance between synthetic samples and training samples.
<b>Output:</b> Classify the target
<b>procedure</b> Ensemble SXRL( <b>S</b> )
<b>for</b> i in <b>M</b> <b>do</b>
Get a copy of the dataset $D_i = D$
Get several training samples from the dataset and define it as <b>P</b> .
<b>for</b> j in <b>P</b> <b>do</b>
Train $M_i$ by $P_i$
<b>end for</b>
<b>Return</b> new data from the output of $M_i$
<b>end for</b>
<b>end procedure</b>

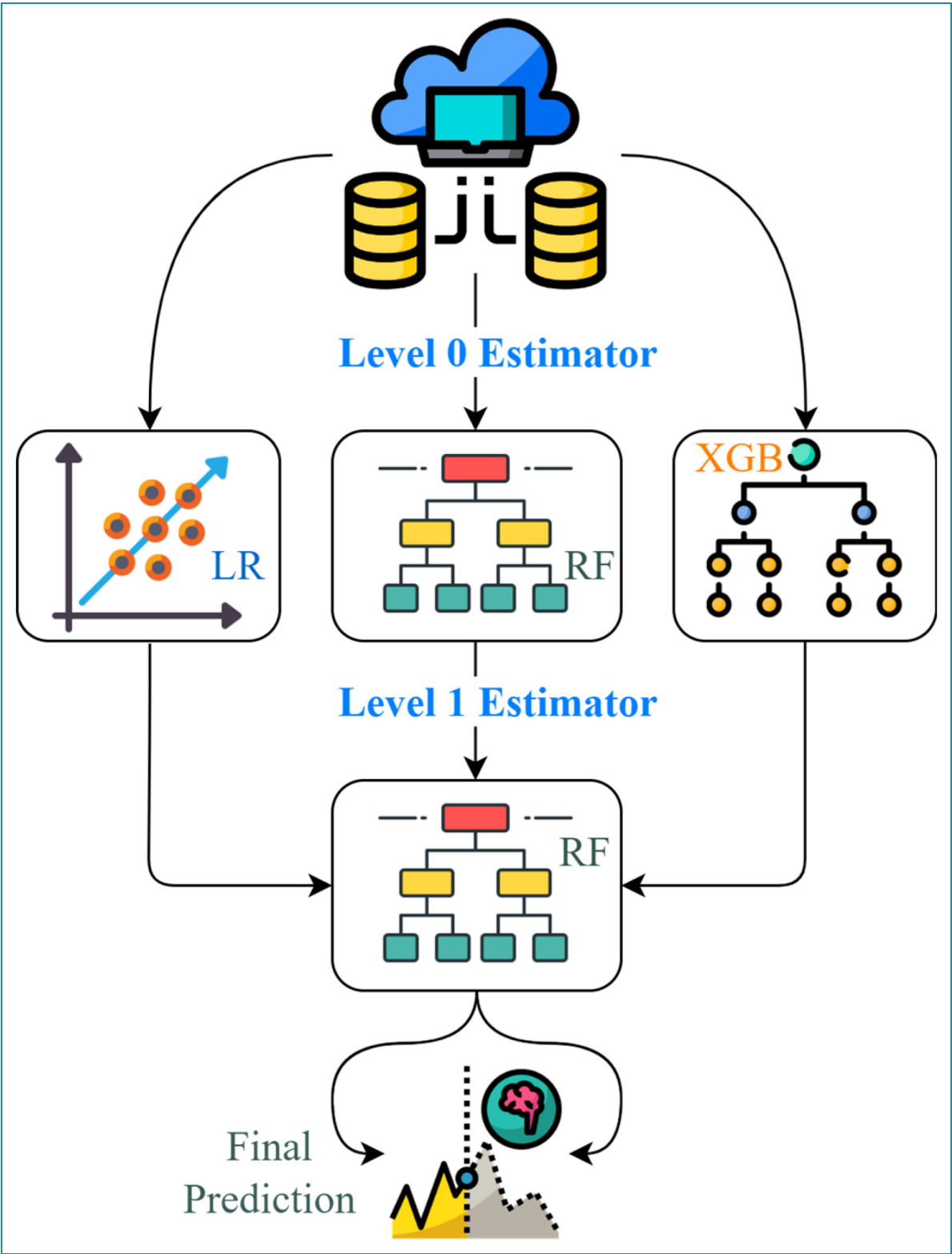


Figure 2. The proposed stacking SRXL ensemble classifier.

model to predict the presence of cervical cancer accurately. GridSearch systematically searches through a predefined grid of hyperparameters to find the optimal values that yield the best performance for the model.

### XAI techniques

XAI is designed as a solution to improve model transparency, changing AI from a mysterious black-box ML model into a more understandable one for the users.<sup>57</sup> The goal of XAI is to develop a variety of strategies that produce models with enhanced explainability while maintaining performance.<sup>58</sup> In this study, we use LIME, SHAP, ELI5 XAI tools, and DT surrogate model explainability to make the model more understandable.

**Shapley Additive Explanations (SHAP).** SHAP stands as the all-encompassing method for interpreting models,<sup>59</sup> serving both local and global explainability purposes. This approach, independent of specific models, aims to unravel the details of complex models such as ensemble methods and DNNs. Moreover, the SHAP framework employs SHAP values, a concept from coalitional game theory, to elucidate the individual contributions of each feature to the model output.<sup>60</sup> SHAP provides comprehensive and fair feature importance values, enhancing the interpretability and trustworthiness of ML models.

**Local interpretable model-agnostic explanations (LIME).** The LIME framework is designed to explain a single instance.<sup>61</sup> Its functionality provides a localized explanation of the classification by identifying the smallest features that significantly influence the likelihood of a specific class outcome for a given observation. In this study, we assess the efficacy of LIME explanations. LIME provides model-agnostic interpretability for ML predictions, enhancing transparency and trust by generating human-understandable explanations.

**ELI5.** ELI5, a Python module, aids in the debugging of ML classifiers and offers straightforward explanations for their predictions.<sup>62</sup> It is designed for non-experts and uses simple language, making it a handy tool for a wide range of applications. ELI5 offers simplified, intuitive explanations for ML models, promoting transparency and interpretability in complex decision-making processes.

**DT surrogate model.** Another popular method to explain the overall behavior of a “black box” model involves using a global surrogate model.<sup>63</sup> The aim of training a global surrogate model is to approximate the predictions of a black-box model. We can make conclusions about the black-box model by analyzing the surrogate model. This research uses DT as the surrogate model to make the tree-based visual representation. These models provide transparent and interpretable

insights into complex ML models, aiding in the understanding and trustworthiness of predictions.

### Performance measurement metrics

ML models’ accuracy and efficacy are assessed using performance-measurement methods. These methods offer measurements and metrics that evaluate a model’s performance and can be used to inform model selection, model improvement, and parameter tuning. The below equations display commonly used techniques in this study as performance measures of the classifiers.

$$\text{Accuracy(Acc)} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision(P)} = \frac{TP}{TP + FP}$$

$$\text{Recall(R)} = \frac{TP}{TP + FN}$$

$$\text{F1 - score(F1)} = \frac{2PR}{P + R}$$

$$\text{Specificity(S)} = \frac{TN}{TN + FP}$$

**Receiver operating characteristics (ROC).** The area under the rate plot of the true positive versus false positive.

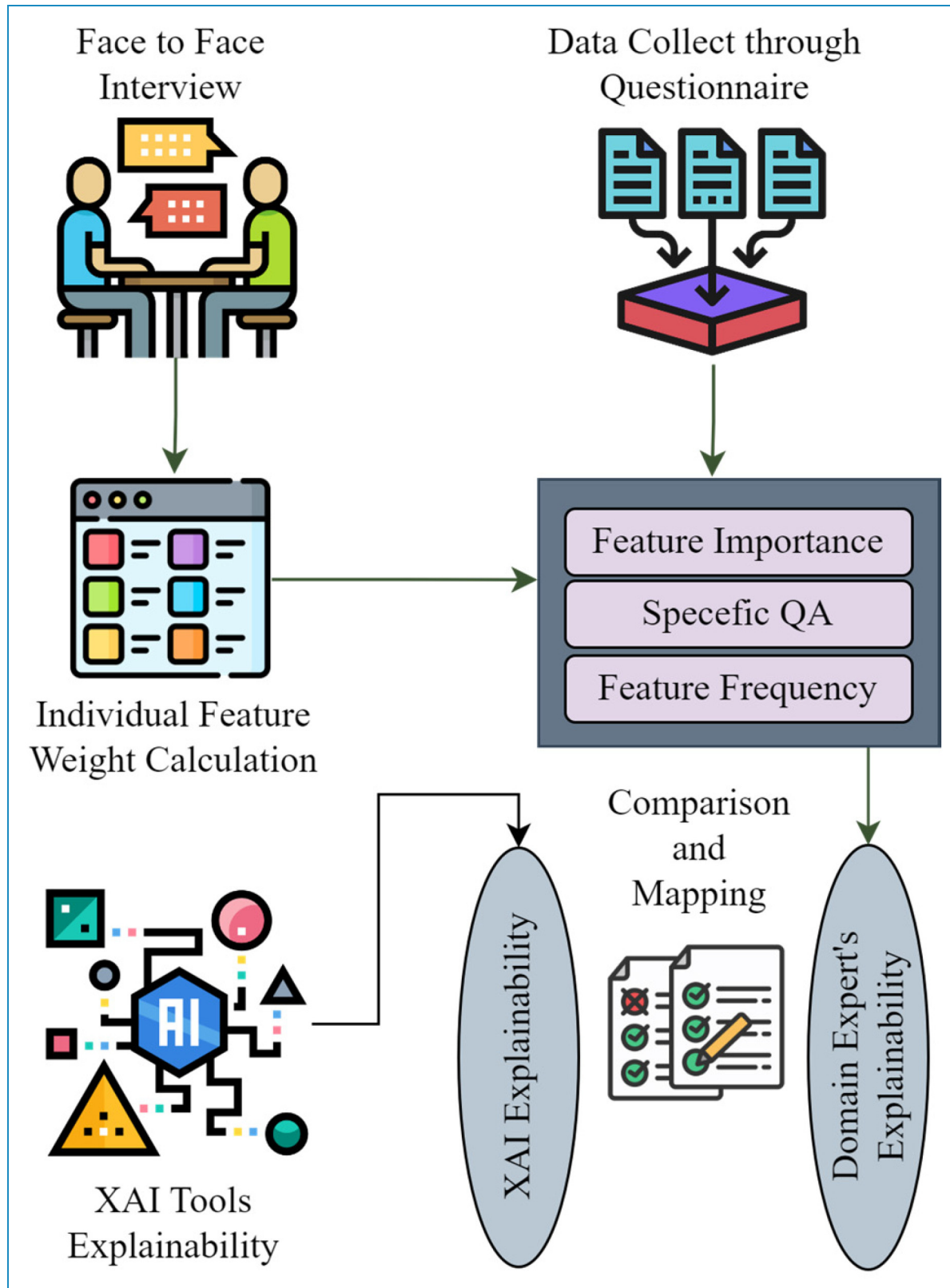
**Area under the curve (AUC).** The ability of the classifier to discriminate between positive and negative classifications.

In the equations, true positive, true negative, false positive, false negative, positive rate, precision (P), and recall (R) are used as TP, TN, FP, FN, PR, P, and R, respectively. These performance metrics provide insightful information about the operation of ML models, assisting in assessment and advancement. The individual task, dataset, and goals of the ML project determine the appropriate technique.

### Domain knowledge integrated explainability

In our research, we also conducted another experiment that correlates the domain expert’s opinion with the XAI tool’s output. To assess and compare the explainability of XAI and the expert’s viewpoint, we engaged in an extensive process, illustrated in Figure 3 as a block diagram.

This experiment established a significant alignment between the domain expert’s opinion and the explanations provided by the XAI tool. We assigned varying weight values to distinguish between features based on their significance in the medical field. We met with several medical experts to discuss the dataset’s features and gather their insights. The weight for individual features was determined according to the recommendations of these domain experts.



**Figure 3.** An overview of domain knowledge-driven explainability mapping and comparison with XAI tools explainability.

Subsequently, we developed an online questionnaire to collect. This section is subdivided into several subsections to convey a comprehensive grasp of the study's contributions on a five-degree Likert scale. Furthermore, we granted experts the flexibility to express their opinions independently. After the data collection procedure, we computed the importance of features using our proposed explainable method, outlined in Algorithm 3.

## Results

This section is subdivided into several subsections to convey a comprehensive grasp of the study's contributions. Initially, we present the dataset's EDA to comprehend its inherent nature. Subsequently, we showcase the classifier's performance across various stages and objectives. The results of XAI tools are presented to enhance interpretability. Lastly, we align the domain experts' opinions with the outcomes of XAI tools, establishing a connection between expert insights and explainability measures. This structured approach provides a holistic overview of the study's findings and their significance in different dimensions.

### Algorithm 3 Feature importance

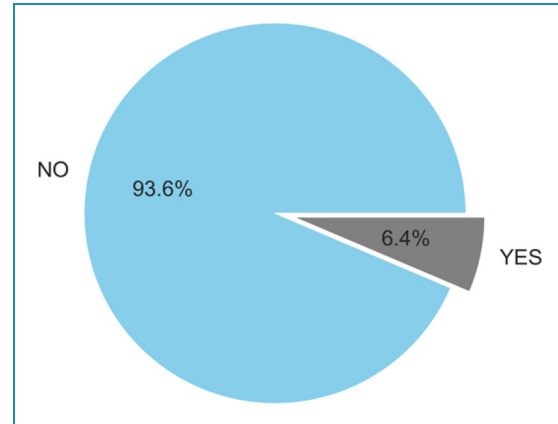
<b>S</b> : Datasets ( <i>S</i> )
<b>F</b> : Number of features
<b>W</b> : Weight of individual features
<b>N</b> : Number of instances
<b>Output</b> : Sorted features according to importance
<b>procedure</b> <i>Feature Importance (S)</i>
for <i>i</i> in <i>F</i> do
<b>for</b> <i>j</i> in <i>N</i> do
$X_i = \sum_{j=1}^N w_{ji}$
$\bar{X}_i = \frac{X_i}{N}$
$D = \bar{X}_i$
<b>end for</b>
<b>end for</b>
sort <i>D</i> in descending order
Return sorted features
<b>end procedure</b>

### EDA to understand the nature of the dataset

One of the primary concerns encountered when dealing with this dataset was effectively managing the imbalanced distribution of data samples across different cases for our target class "biopsy." Figure 4 illustrates the clear imbalance within the dataset. A mere 6.4% of the total data records indicate the presence of cancer-positive classes, highlighting the need for further investigation and attention to this critical issue.

This study has effectively addressed this problem by implementing well-established balancing techniques, including SMOTE, ADASYN, NearMiss, and SMOTETomek. After applying the data balancing techniques, a significant improvement in the data distribution for each class can be observed in Table 2. This ensures a more balanced representation of the classes in our dataset.

The following figures provide a comprehensive overview of the dataset's behavior. Figure 5 clearly illustrates that the majority of patients belong to the age group between 20 and 40. It demonstrates that a significant portion of the



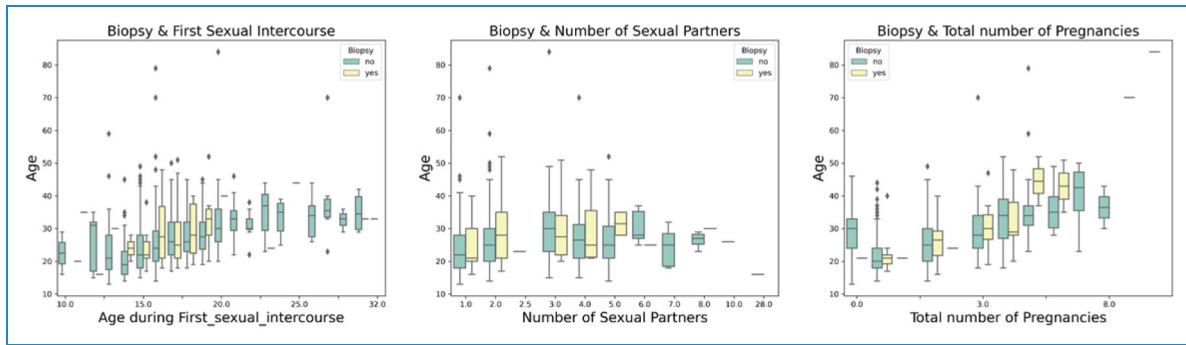
**Figure 4.** Percentage of biopsy of individual class.

**Table 2.** Distribution of individual classes before and after applying the data balancing techniques.

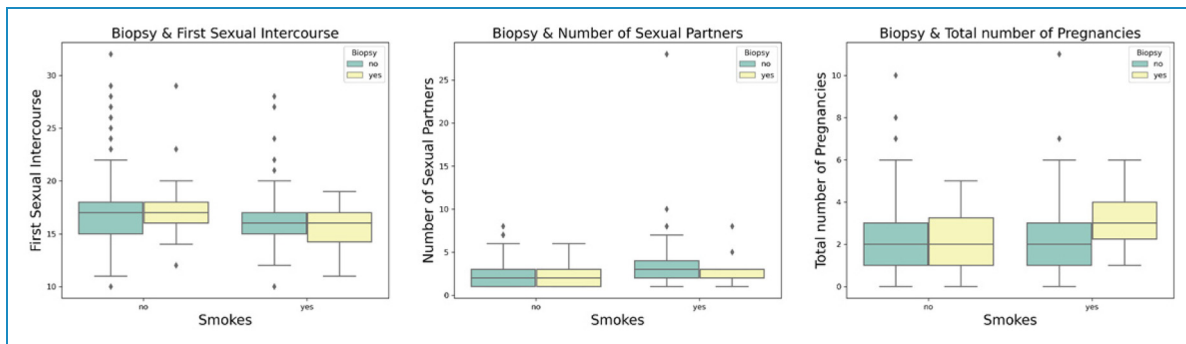
Method	Class 0	Class 1	Class 0 (%)	Class 1 (%)	Total
Normal	803	55	93.5	6.5	858
SMOTE	803	803	50	50	1606
ADASYN	803	823	49	51	1626
NearMiss	55	55	50	50	110
SMOTETomek	802	802	50	50	1604

ADASYN: adaptive synthetic; SMOTE: synthetic minority oversampling technique.

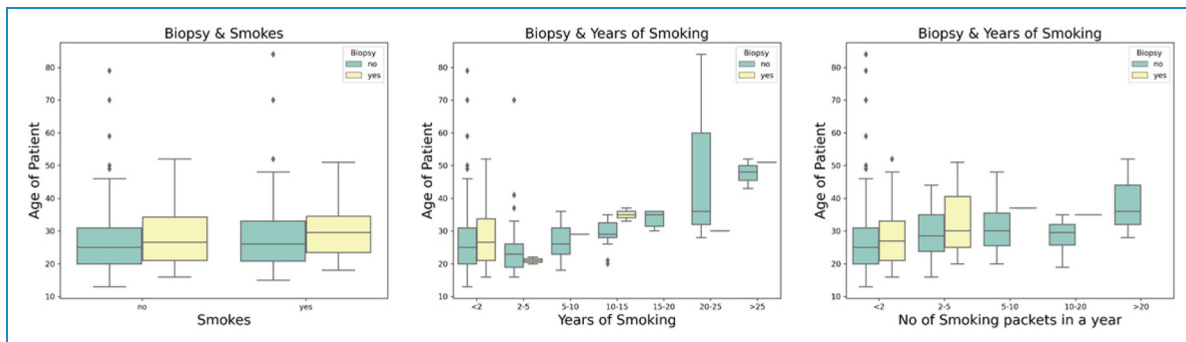




**Figure 5.** Multivariate analysis on age and sexual habits versus biopsy.



**Figure 6.** Multivariate analysis on smoking and sexual habits versus biopsy.



**Figure 7.** Multivariate analysis on age and smoking habits versus biopsy.

patients underwent pregnancy between one and three times. Moreover, the majority of individuals engaged in their initial sexual experience between the ages of 15 and 20, and had been exposed to zero to five sexual partners. Therefore, it is evident from the plots that those who had their first sexual intercourse between 15 and 18 years of their life and experienced sexual interaction with multiple partners are at a higher risk of testing positive on the biopsy test.

Figures 6 and 7 establish the correlation between a patient's smoking habits, sexual behaviors, and the target variable. The graphical visualization highlights that approximately 88% of the patients (700 samples) have a

non-smoking background. This information primarily suggests that smoking may not be strongly correlated with the target variable, as the majority of patients do not have a smoking background. One important point to be noticed here is that individuals who smoke and have early age first sexual intercourse are at a higher risk of testing positive in a biopsy. Moreover, the average age of non-smoking cancer-positive patients is less than the average age of cancer-positive patients with a prior smoking history. This indicates that older women with a prolonged smoking history are more prone to be tested cancer-positive compared to young women who do not smoke.

**Table 3.** Hyperparameter values of the ML classifiers in both without preprocessed and with preprocessed data.

Model	Raw data	Pre-process data
SVC	C = 0.1, gama = 0.1, kernel = "linear"	C = 0.1, gama = 0.01, kernel = "linear"
DT	min_samples_leaf = 20, random_state = 42	min_samples_leaf = 25, random_state = 42
RF	max_depth = 3, min_samples_leaf = 7, min_samples_split = 12	max_depth = 19, min_samples_leaf = 9, min_samples_split = 10
GBC	n_estimators = 100, learning_rate = 0.01, max_depth = 3, subsample = 1.0, min_samples_split = 2, min_samples_leaf = 1	n_estimators = 120, learning_rate = 0.01, max_depth = 3
LR	C = 1.0, penalty = "l2," solver = "liblinear," max_iter = 100, random_state = 42	C = 1.0, penalty = "l2," solver = "liblinear," max_iter = 100, random_state = 42
KNN	n_neighbors = 19, weights = "uniform," metric = "minkowski," p = 2	n_neighbors = 5, weights = "uniform," metric = "minkowski," p = 2
CB	depth = 8, learning_rate = 0.01	depth = 10, learning_rate = 0.01
XGB	objective = "reg:linear," n_estimators = 20	objective = "reg:linear," n_estimators = 42
SRXL	kernel = "rbf," gamma = "scale," nu = 0.5	Kernel = "rbf," random_state = 42

ML: machine learning; SVC: support vector classifier; DT: decision tree; RF: random forest; GBC: gradient boosting classifier; LR: logistic regression; KNN: K-nearest neighbor; CB: CatBoost; XGB: extreme gradient boosting.

### Algorithms hyperparameters setting

In this study, we utilized a secondary dataset for binary classification purposes. Following the preprocessing steps, we employed eight ML models and one ensemble ML model for classification. Here, we presented the associated hyperparameters of the ML classifiers in Table 3. The values of the hyperparameters change for different data distributions and data properties.

### Results of different classifiers

We examined the behavior of the data and made predictions for cervical cancer by applying nine (9) classifiers on the risk factors dataset. Table 4 demonstrates the performance comparison between nine algorithms and different oversampling and under-sampling techniques studied in this experiment. The presence of biases in the model is evident. While the overall accuracy may reach 94% to 96%, it is concerning that the performance for the minority class is significantly poorer and shows a lack of consistency without applying any preprocessing techniques. Moreover, the minority class exhibits significantly lower values in terms of precision, recall, and F1 score compared to the majority class. These imbalances highlight the urgent need to

address the bias problem to obtain more accurate predictions for cervical cancer. To handle the imbalanced dataset, we applied the SMOTE balancing technique. SMOTE is a highly effective technique for oversampling that creates synthetic samples in the minority class by interpolating between neighboring examples. After applying SMOTE, we achieved a more balanced dataset, which improved the performance of different ML and deep learning models. The subsequent findings suggest the dominance of our proposed ensemble model over other classifiers in terms of every performance standard. ADASYN is an oversampling technique that balances the dataset by increasing the sample count for the minority class. The results demonstrate that by implementing ADASYN along with the hybrid feature selection method, we were able to enhance the overall performance of the models significantly. This improvement is especially noteworthy compared to their previous performance on the initial imbalanced dataset. However, it is noticeable that the ML classifiers could not surpass their performance with SMOTE. SVC showed the least accuracy among them, achieving only 62%, while the proposed model is proven to be the most capable one with 97% accuracy.

After applying an undersampling technique, NearMiss, the results are reported as an improvement of the overall model prediction compared to the initial findings. However, the

**Table 4.** Performance comparison of accuracy and FI-score across algorithms and preprocessing techniques.

Model	Without preprocessing		SMOTE		ADASYN		NearMiss		SMOTETomek	
	Acc.	FI-score	Acc.	FI-score	Acc.	FI-score	Acc.	FI-score	Acc.	FI-score
SVC	0.95	0.98	0.88	0.89	0.62	0.69	0.91	0.92	0.87	0.88
		0.60		0.87		0.52		0.89		0.85
DT	0.95	0.97	0.91	0.91	0.87	0.88	0.91	0.92	0.91	0.91
		0.56		0.92		0.86		0.89		0.91
RF	0.95	0.97	0.97	0.97	0.96	0.97	0.91	0.92	0.96	0.96
		0.40		0.97		0.96		0.89		0.96
GBC	0.96	0.98	0.97	0.97	0.95	0.95	0.86	0.89	0.97	0.97
		0.67		0.97		0.94		0.82		0.97
LR	0.96	0.98	0.89	0.90	0.61	0.62	0.93	0.94	0.87	0.89
		0.59		0.88		0.59		0.92		0.86
KNN	0.94	0.97	0.86	0.84	0.81	0.79	0.80	0.84	0.87	0.85
		0.00		0.87		0.83		0.71		0.88
CB	0.95	0.97	0.96	0.96	0.96	0.96	0.91	0.92	0.96	0.96
		0.56		0.96		0.95		0.89		0.96
XGB	0.95	0.98	0.97	0.97	0.96	0.96	0.82	0.84	0.97	0.97
		0.62		0.97		0.96		0.79		0.97
SRXL	0.96	0.98	0.98	0.98	0.97	0.97	0.95	0.96	0.98	0.98
		0.67		0.97		0.97		0.95		0.98

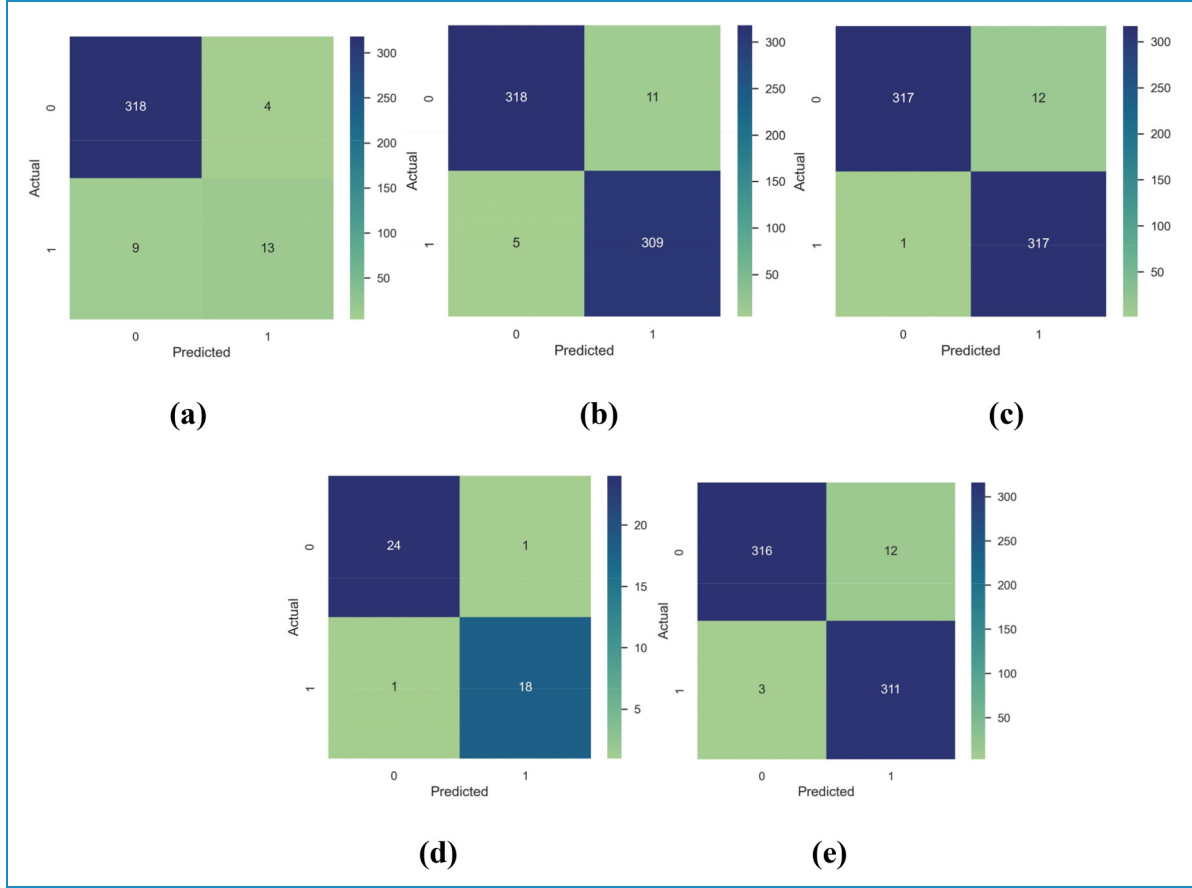
Acc.: accuracy; SMOTE: synthetic minority oversampling technique; ADASYN: adaptive synthetic; SVC: support vector classifier; DT: decision tree; RF: random forest; GBC: gradient boosting classifier; LR: logistic regression; KNN: K-nearest neighbor; CB: CatBoost; XGB: extreme gradient boosting.

performance considerably deteriorated for most of the classifiers this time. Once more, the top-performing model is the proposed ensemble model, achieving a 95% accuracy, while the poorest performer is XGB, with an accuracy of 82%. On the other hand, SMOTETomek significantly enhanced the performance metrics for all classifiers, effectively addressing bias for both majority and minority classes. SMOTETomek is a hybrid sampling technique combining both up and downsampling. Despite achieving impressive results, SMOTETomek still fell short of outperforming SMOTE. From the above discussion, we can conclude that oversampling techniques can be extremely useful when dealing with imbalanced data.

### Confusion matrix of the proposed method in different state

The confusion matrix depicted in Figure 8 denotes the model performance at different stages using different oversampling and undersampling techniques. It is visible that the model performance has been significantly improved and is stable when the hybrid feature selection method was applied combined with the oversampling techniques. The proposed model's performance improved significantly, utilizing the SMOTE oversampling technique to handle the imbalanced dataset.

Figure 9 of the AUC-ROC curves further establishes the superiority of our proposed stacking ensemble model. The



**Figure 8.** Confusion matrices of different stages (a) normal setup, (b) feature selection with SMOTE, (c) feature selection with ADASYN, (d) feature selection with NearMiss, and (e) feature selection with SMOTETomek. SMOTE: synthetic minority oversampling technique; ADASYN: adaptive synthetic.

ensemble model outperformed the traditional ML and deep learning models in accurately classifying cervical cancer, demonstrating remarkable performance with an AUC of 0.995 with SMOTE. The graph clearly shows that the ROC curve becomes unstable when using ADASYN, NearMiss, and SMOTETomek balancing techniques with other ML and deep learning classifiers. In contrast, stability is achieved in Figure 9(b) when using SMOTE. This indicates the superior classification performance of the proposed model, along with the robustness and consistency of the proposed classification pipeline.

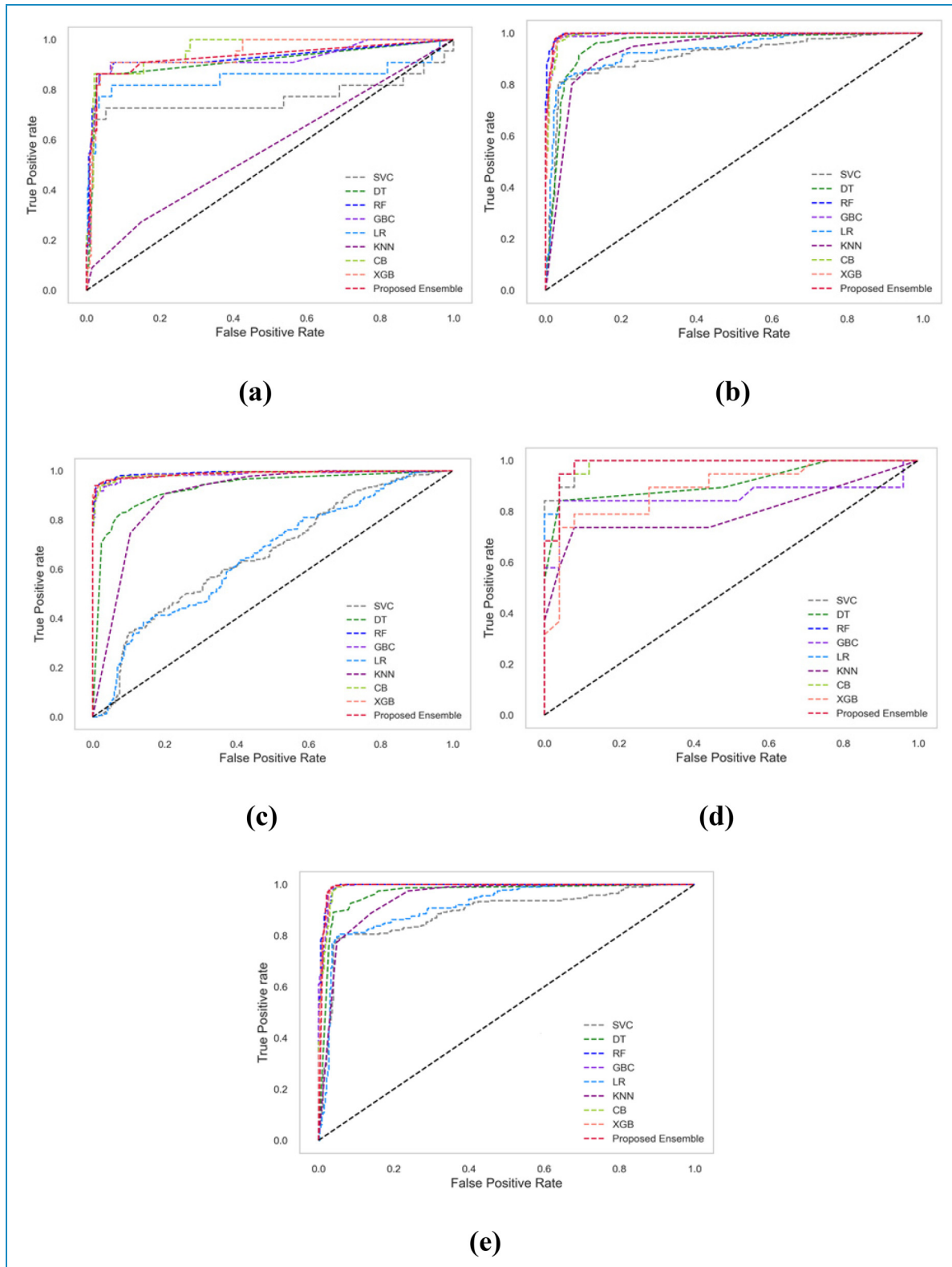
#### Cross validation and different ensemble approach's result

To verify the robustness and generalizability of our proposed ensemble model, we conducted a 10-fold cross-validation. The results are demonstrated in Table 5. This iterative process helps smooth out variations in model performance, providing a more stable and reliable assessment while minimizing the potential scope of overfitting. The results indicate a consistency in stable performance across all the ML algorithms used in this study. However, the

proposed SRXL model consistently outperforms them, achieving higher predictive accuracy which demonstrates its effectiveness in handling complex unseen patterns within the dataset. This superior performance highlights its potential for real-world applications where reliability and robustness are crucial.

We present a comparative analysis of different ensemble techniques, evaluating their performance based on accuracy, precision, recall, and F1-score in Table 6. The ensemble models compared include voting ensemble, blending ensemble, stacking with basic ML models, and the proposed SRXL model.

The voting ensemble, which combines LR, DT, KNN, and SVC, achieves an accuracy of 0.95. It maintains balanced performance across precision, recall, and F1-score, each reaching 0.94. Although this technique leverages multiple models, its performance is limited by the averaging nature of voting, which may not optimally capture complex patterns in the data. The blending ensemble, incorporating LR, DT, and SVC, slightly improves performance over voting, reaching 0.96 accuracy and 0.95 across other metrics. This improvement suggests that



**Figure 9.** ROC curve of different stages (a) normal setup, (b) feature selection with SMOTE, (c) feature selection with ADASYN, (d) feature selection with NearMiss, and (e) feature selection with SMOTETomek. ROC: receiver operating characteristics; SMOTE: synthetic minority oversampling technique; ADASYN: adaptive synthetic.



**Table 5.** Result of the 10-fold cross validation of the models.

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
SVC	0.955	0.978	0.9693	0.9639	0.9462	0.9462	0.9423	0.9746	0.964	0.9683
DT	0.9408	0.9788	0.9733	0.9485	0.9473	0.9473	0.9522	0.961	0.9573	0.9516
RF	0.9645	0.9456	0.9517	0.9547	0.9582	0.9714	0.948	0.9606	0.9637	0.9419
GBC	0.9643	0.9468	0.9426	0.978	0.9786	0.9723	0.9522	0.9439	0.9674	0.9576
LR	0.9449	0.9598	0.9414	0.9764	0.9504	0.9665	0.9525	0.9608	0.9619	0.9474
KNN	0.9788	0.971	0.9776	0.9758	0.9639	0.9769	0.9435	0.9478	0.9418	0.953
CB	0.9555	0.9509	0.9731	0.9543	0.9512	0.9617	0.9456	0.9721	0.943	0.9795
XGB	0.9709	0.9479	0.9402	0.9726	0.9683	0.9692	0.9709	0.943	0.9543	0.9446
SRXL	0.9745	0.9649	0.9532	0.9625	0.9724	0.9853	0.9692	0.9755	0.9755	0.9689

SVC: support vector classifier; DT: decision tree; RF: random forest; GBC: gradient boosting classifier; LR: logistic regression; KNN: K-nearest neighbor; CB: CatBoost; XGB: extreme gradient boosting.

**Table 6.** Results of different ensemble techniques.

Model	Algorithm	Accuracy	Precision	Recall	F1-score
Voting ensemble	LR, DT, KNN, SVC	0.95	0.94	0.94	0.94
Blending ensemble	LR, DT, SVC	0.96	0.95	0.95	0.95
Stacking (basic ML models)	LR, DT, KNN, SVC	0.97	0.96	0.96	0.96
Proposed SRXL	RF, XGB, LR	0.98	0.98	0.97	0.98

LR: logistic regression; DT: decision tree; KNN: K-nearest neighbor; SVC: support vector classifier; RF: random forest; XGB: extreme gradient boosting; ML: machine learning.

weighted blending of models is more effective than simple voting, potentially due to better capturing the strengths of individual classifiers. However, it still lags behind the more advanced ensemble techniques. The stacking ensemble, which utilizes LR, DT, KNN, and SVC, significantly enhances performance, achieving 0.97 accuracy along with precision, recall, and an F1-score of 0.96. Stacking generally outperforms voting and blending because it leverages a meta-model that learns from the predictions of base models, improving decision boundaries and reducing biases inherent in individual classifiers. The proposed SRXL model, which integrates RF, XGB, and LR, exhibits the highest performance, achieving an accuracy of 0.98. It also outperforms other methods in precision (0.98), recall (0.97), and F1-score (0.98). This suggests that SRXL benefits from the powerful feature selection and ensemble learning capabilities of RF and XGB while maintaining the interpretability of LR. The superior results indicate that this combination is more effective in capturing complex

data relationships and minimizing errors compared to traditional ensemble techniques.

The results highlight that while all ensemble models improve classification performance, SRXL is the most effective due to its advanced ensemble learning approach, leveraging the strengths of RF, XGB, and LR. This makes it a promising technique for applications requiring high accuracy and robustness.

### Result of different generative AI approaches

Table 7 presents the results obtained by applying variational autoencoder (VAE) to generate artificial data and then utilizing RF as the ultimate classifier. On the other hand, Table 8 depicts the GTNs' performance. These models are evaluated across different data balancing techniques and batch sizes.

The tables present the performance of the VAEs-enabled RF and GTNs model, respectively, using batch sizes of 32

**Table 7.** Performance of VAEs-enabled RF.

State	Batch size = 32				Batch size = 64			
	Acc	P	R	F1	Acc	P	R	F1
Without balancing	0.93	0.20	0.11	0.14	0.94	1.00	0.09	0.17
SMOTE	0.59	0.53	0.59	0.56	0.64	0.58	0.67	0.62
ADASYN	0.67	0.66	0.69	0.67	0.66	0.64	0.73	0.68
NearMiss	0.45	0.50	0.58	0.54	0.54	0.57	0.67	0.62
SMOTETomek	0.64	0.59	0.68	0.63	0.67	0.61	0.78	0.68

VAE: variational autoencoder; RF: random forest; Acc: accuracy; P: precision; R: recall; SMOTE: synthetic minority oversampling technique; ADASYN: adaptive synthetic.

**Table 8.** Performance of GTNs.

State	Batch size = 32				Batch size = 64			
	Acc	P	R	F1	Acc	P	R	F1
Without balancing	0.94	0.25	0.15	0.20	0.94	0.00	0.00	0.00
SMOTE	0.49	0.45	0.64	0.53	0.52	0.47	0.64	0.55
ADASYN	0.47	0.46	0.42	0.44	0.48	0.48	0.58	0.53
NearMiss	0.36	0.42	0.42	0.42	0.41	0.45	0.42	0.43
SMOTETomek	0.47	0.44	0.66	0.52	0.47	0.45	0.62	0.52

GTN: generative teaching network; Acc: accuracy; P: precision; R: recall; SMOTE: synthetic minority oversampling technique; ADASYN: adaptive synthetic.

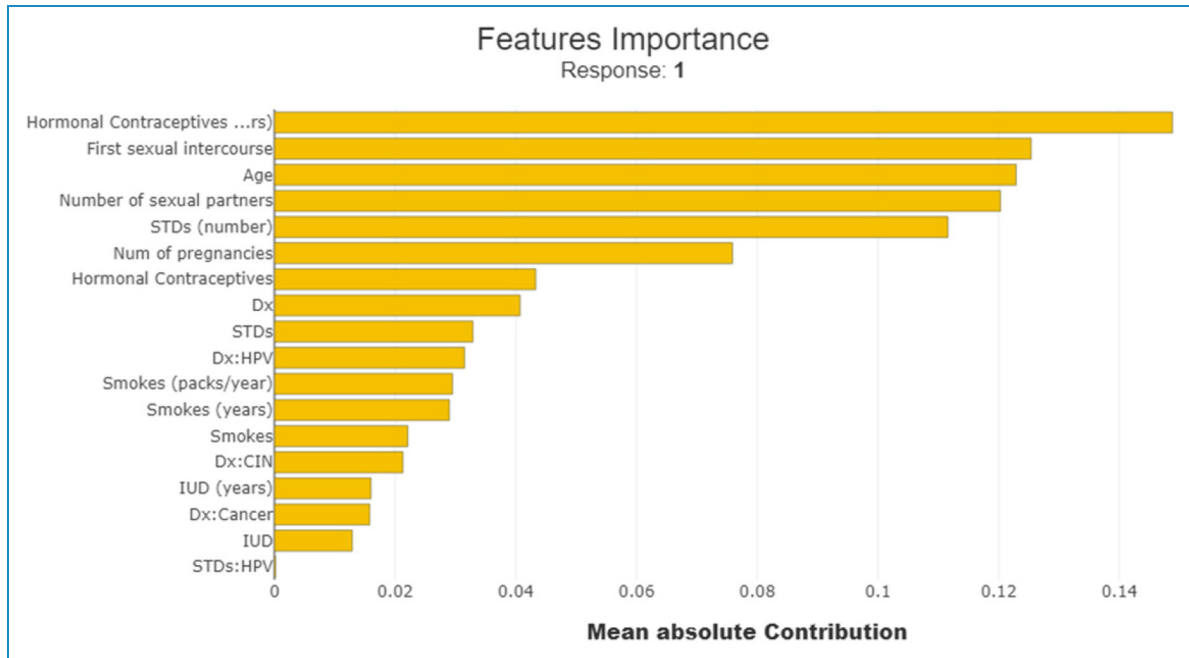
and 64, 500 epochs, and stochastic gradient descent optimizer across different data balancing techniques. The first state indicates the model performance on the original imbalanced dataset, showing high accuracy but poor precision, recall, and F1 scores, particularly for the minority class. Using SMOTE and SMOTETomek, the model shows improved precision, recall, and F1 scores, indicating better handling of the imbalanced data. ADASYN also improves these metrics, resulting in lower accuracy, reinforcing that aggressive undersampling might be detrimental due to significant information loss. Comparing the two batch sizes, a batch size of 64 generally provides slightly better results than a batch size of 32 across most metrics and techniques.

Although VAEs and GTNs excel at capturing the intricate probability distribution of data, they may not be ideal for achieving precise classification on tabular data. Generative models lack the discriminative capability to differentiate between classes based on input features. Moreover, they encounter challenges in handling categorical data, a common occurrence in tabular datasets. In

addition, VAEs and GTNs frequently entail intricate training processes and are susceptible to mode and posterior collapses. This significantly reduces their effectiveness in classification tasks. Conversely, discriminative ML models such as LR, DT, RF, gradient-boosting machine (GBM), and SVM are generally more suitable for classification tasks because they handle nonlinear relationships and interactions between features. These models can effectively capture the structure of tabular data, better handle categorical features, and make more accurate and reliable predictions for tabular data classification problems.

### Explainability results

Model explainability in ML ensures transparency by revealing how models make predictions. Techniques such as SHAP, LIME, and ELI5 provide insights into the importance of features, helping users understand complex algorithms. XAI is vital for trust and ethical deployment in critical areas such as healthcare and finance, allowing stakeholders to validate decisions, identify biases, and enhance



**Figure 10.** Feature importance using Shapley additive explanations (SHAP).

model performance. The results of the model's explainability using different tools are given below.

**Global explainability.** Explainability adds the essence of transparency to the models' prediction. Figure 10 ranks the top 10 features contributing most to the overall prediction of the classification model. The results indicate that the feature "Hormonal Contraceptives (years)" has the highest importance, followed by "First Sexual Intercourse" and "Age." These features provide valuable insights into the decision-making process of the model. The mean absolute contribution reveals that hormonal contraceptives have the most significant impact on the model's predictions.

In Table 9, we can see a comprehensive breakdown of each feature and their corresponding weights, listed in descending order of importance. The number of years spent consuming hormonal contraceptives is given the highest priority, while the number of smoking years is given the lowest priority. The numerical value shown in the table is the result of combining the individual feature weight with its corresponding tolerance. For instance, the individual feature weight for age is 0.0240, meaning that even the smallest alteration in the age value could impact the model prediction by 0.0064. Furthermore, in Figure 10, we can observe that most features correspond to the table representation. This confirms the validity of our findings.

### Local explainability

**Impact of individual features.** This section explains the effect of the individual feature's impact on the final prediction of

the classification model. Figure 11(a) and (b) shows that the longer a woman uses hormonal contraceptives, the higher her probability of being diagnosed with cervical cancer. The subsequent two figures unveil the significant impact of a woman's sexual habits on the classifiers' predictions. Figure 11(c) illustrates the impact of early sexual exposure on the classifier's prediction, while Figure 11(d) depicts the influence of sexual preferences in terms of the count of sexual partners. These insights shed light on the significance of the predictive model. Based on these findings, it can be concluded that women who engage in early sexual intercourse and have multiple sexual partners are at a significantly higher risk of being diagnosed with cervical cancer. This further emphasizes the importance of comprehensive sex education and access to reproductive health services for young women.

Figure 11(e) to (j) portrays the significant factors that contribute to the patient's condition, including age, the number of STDs detected, the number of pregnancies, and the presence of HPV. The data indicates that women diagnosed with STDs and other conditions such as cancer, CIN, or HPV are at a higher risk of developing cancer. Therefore, healthcare professionals must take into account these factors when evaluating an individual's cancer risk. Furthermore, we can observe the impact of the woman's age on the prediction accuracy, with younger women tending to have a lower likelihood of the predicted outcome. More precisely, older women with similar conditions are more susceptible to risk compared to younger women.

**Explainability on individual instances.** Figure 12 shows the impact of the features on individual patient records. It is

**Table 9.** Individual feature's weight and tolerance to the classifiers.

Feature name	Weight
Hormonal contraceptives (years)	$0.0259 \pm 0.0061$
First sexual intercourse	$0.0247 \pm 0.0048$
Age	$0.0240 \pm 0.0064$
Number of sexual partners	$0.0207 \pm 0.0027$
Number of pregnancies	$0.0193 \pm 0.0079$
Hormonal contraceptives	$0.0100 \pm 0.0043$
STDs (number)	$0.0096 \pm 0.0045$
STDs	$0.0084 \pm 0.0023$
IUD	$0.0037 \pm 0.0009$
Smokes	$0.0037 \pm 0.0009$
IUD (years)	$0.0035 \pm 0.0021$
Smokes (packs/years)	$0.0028 \pm 0.0024$
Dx:HPV	$0.0026 \pm 0.0009$
Dx:Cancer	$0.0023 \pm 0.0000$
Dx:CIN	$0.0023 \pm 0.0000$
Dx	$0.0023 \pm 0.0015$
Smokes (years)	$0.0016 \pm 0.0011$

STD: sexually transmitted disease; IUD: intrauterine device; Dx:HPV: presence of human papilloma virus after the diagnosis; Dx:Cancer: presence of cancer after the diagnose; Dx:CIN: presence of cervical intraepithelial neoplasia after the diagnosis.

evident from the figure that when our extracted top features from Figure 10 contribute inversely to the patient's condition, it is less likely to be diagnosed with cancer. On the contrary, when the extracted top features contribute mostly to the outcome, the patient is at high risk of cervical cancer (see Figure 13(a) and (b)).

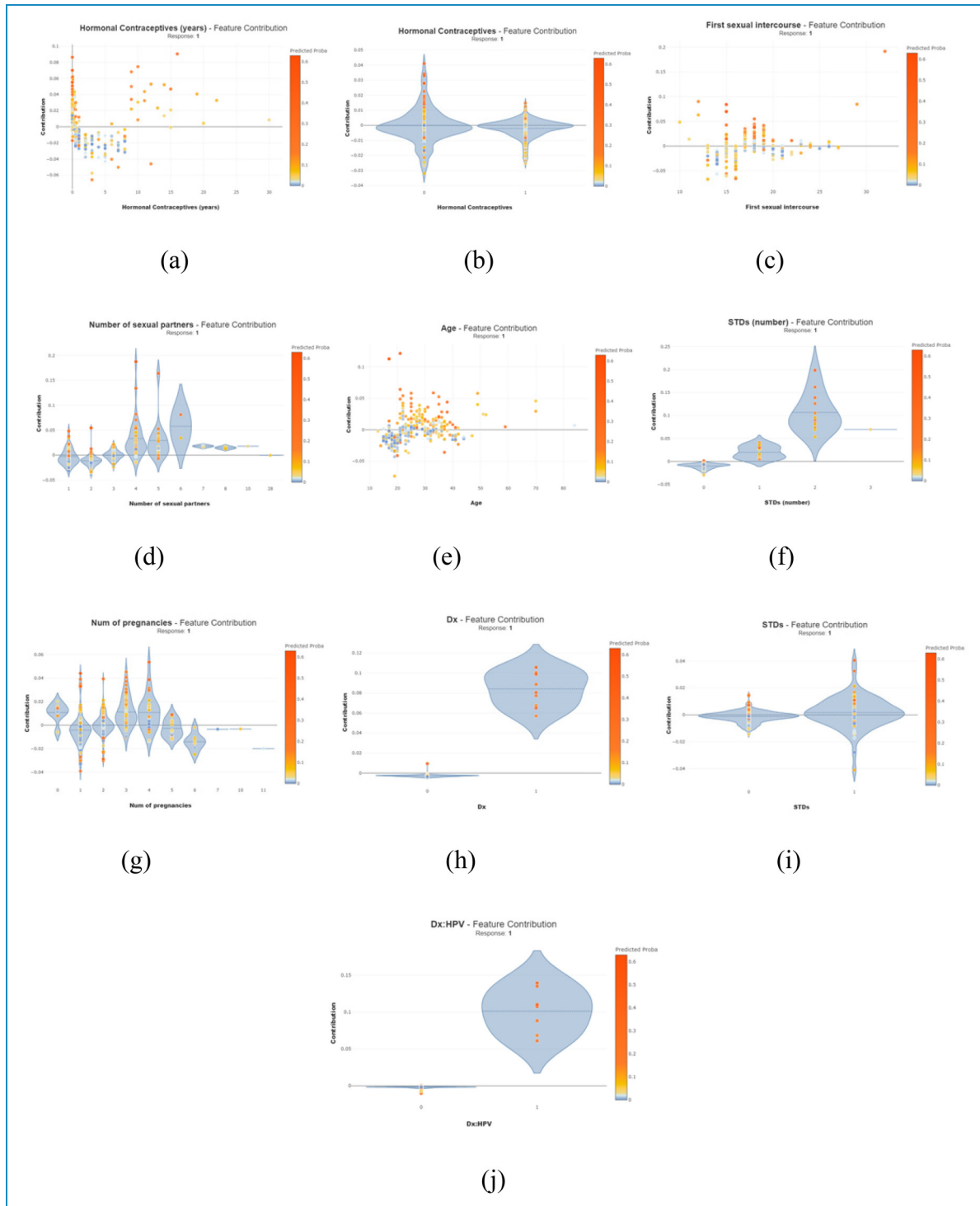
Figure 14 displays a comprehensive comparison of the significance of various features and their respective contributions to the prediction of the outcome. We have selected 10 random data samples from the dataset to analyze the contribution of different features to the model's predictions. As shown in the compare plot, features such as "Hormonal Contraceptives (years)," "Number of pregnancies," and "Number of sexual partners" exhibit greater variation in their contributions (more scattered

lines) compared to others. This indicates a higher standard deviation for these features, suggesting they have a more significant and variable impact on the model's final predictions compared to features with smaller deviations. As the top features contribute more, the models' predictions also improve. This highlights the significance of selecting and incorporating the most influential features in our predictive models. It further establishes the superiority of these top features to achieve more accurate and reliable predictions.

**DT surrogate model explainability.** A surrogate model is the graphical representation of the predictions made by the classification model at each step. It explains the complex decision-making task by splitting the overall process based on several features and their respective values. In Figures 15 and 16, two surrogate models have been shown at different depth levels. In these figures, the internal or parent nodes represent the features that serve as the conditions for splitting, while the terminal nodes represent the ultimate prediction results based on the specified criteria for splitting. This approach allows for a more transparent understanding of the decision-making process and provides insights into how the model arrives at its predictions. Visualizing the surrogate models at different depth levels makes it easier to comprehend the impact of various features on the final prediction. In these figures, the DT surrogate model uses different features obtained from the patient's medical records such as "Hormonal Contraceptives (years)," "STDs," "Num of pregnancy," etc. to determine the probability of cervical cancer. If the conditions for splitting exceed a certain threshold, the model accurately predicts a positive cancer diagnosis for the patient. Conversely, if the conditions fall below the threshold, the patient is not at a high risk for cancer. Furthermore, the DT surrogate model considers various factors such as age, smoking habits, and sexual activity duration to obtain a more accurate prediction of cervical cancer risk. By analyzing these features, the model aims to assist in determining the top contributing features and better prediction of the disease.

### Result of domain knowledge integrated explainability and XAI tools mapping

In this study, we compared the insights provided by domain experts with the outcomes revealed by utilizing integrated XAI tools. This approach allowed us to identify areas of agreement and potential discrepancies between the model and human experts. Figure 17(a) depicts the weighted average calculated for individual features. The proposed feature selection method categorizes the top features into three weighted categories, W1, W2, and W3, and assigns labels to the features based on their contribution to the final

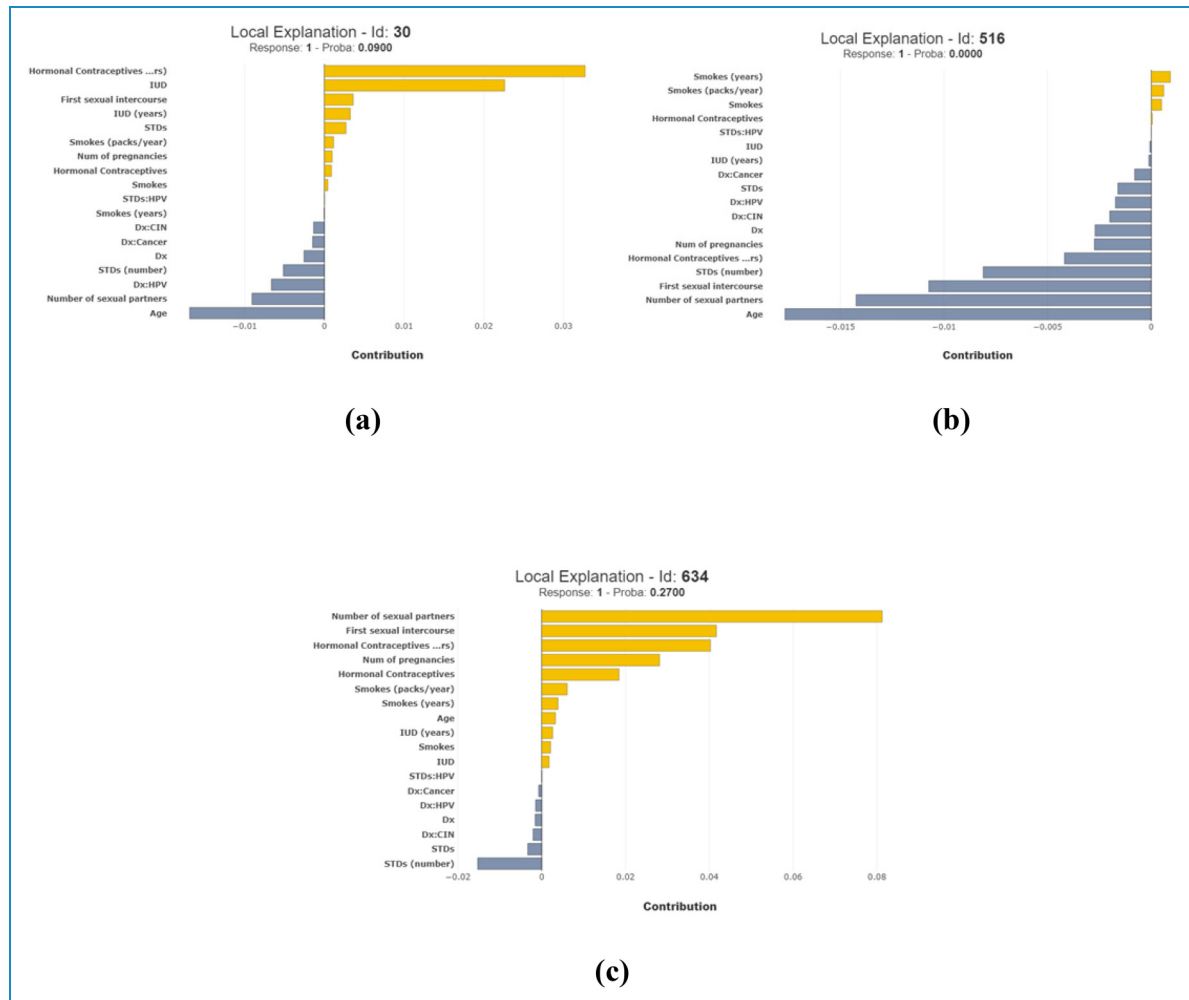


**Figure 11.** Individual features explainability.

predicted outcome. The two main factors that have the greatest impact on the likelihood of developing cervical cancer are HPV infection and having multiple sexual partners. These factors are given the highest level of importance, indicated by a weighting of W3. On the other hand, age, hormonal contraceptives, and smoking are assigned

the least weighting of W1, indicating a lesser impact on cancer risk. The domain experts' response exhibits an indexing pattern similar to the previous figure. It is illustrated in Figure 17(b). Experts have confirmed that the main factors contributing to the risk of developing cervical cancer in the future are engaging in sexual activity with multiple partners



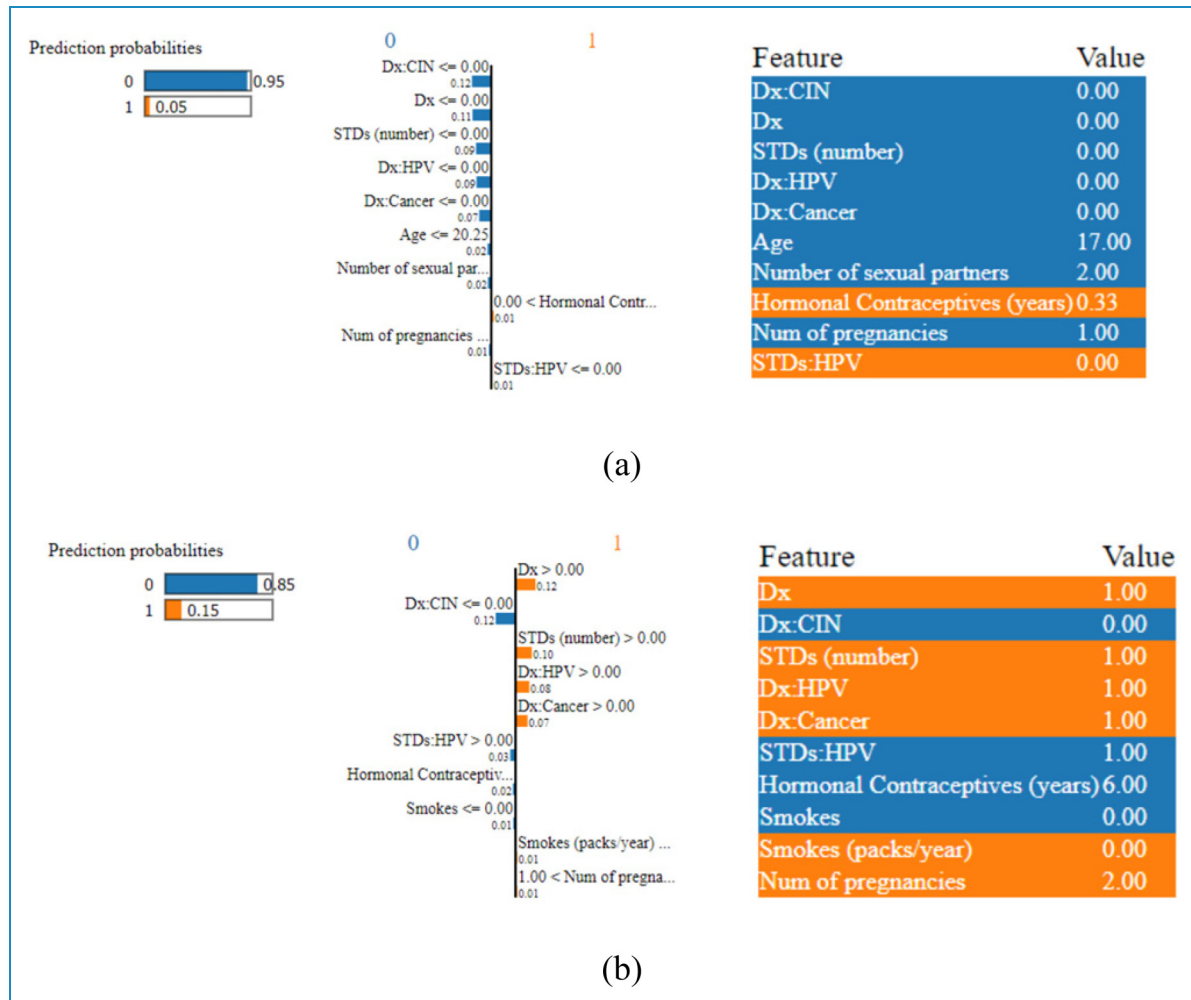


**Figure 12.** Individual instance explainability.

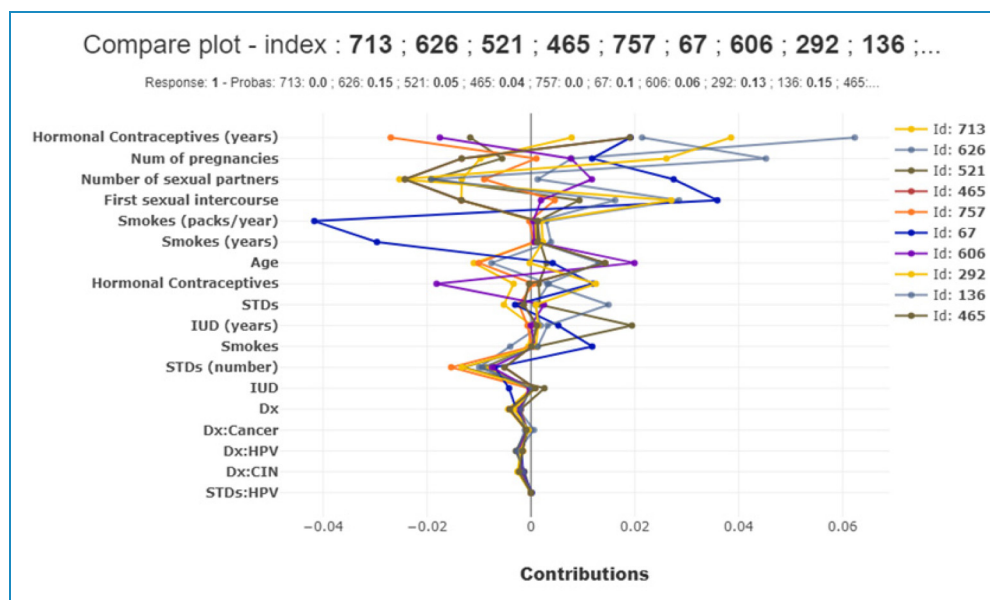
and contracting the HPV. Furthermore, they identified early sexual activity, a long history of hormonal contraceptive use, and the presence of other STDs as major risk factors for the development of cervical cancer.

Figure 18 visualizes the primary factors leading to cervical cancer as identified by domain experts. These reasons include HPV infection, smoking, hormonal contraceptives, multiple sexual partners, and poor hygiene practices. Further analysis of the data reveals that these factors often interact with one another, exacerbating the risk of cervical cancer. A clear dominance of HPV infection and engagement with multiple sexual partners can be observed with 90% and 55% scores, respectively. This further confirms that our proposed model, integrated with XAI, accurately detected the indicators of cervical cancer. The majority of experts agree that the prolonged use of hormonal contraceptives for over 10 years, particularly during a woman's reproductive years, significantly increases the risk of developing cervical cancer (see Figure 19).

The model's predicted feature importance differs and may not always align perfectly with the risk factors identified by the experts. This is due to various factors such as data quality, sample size, and modeling assumptions. Figures 17 and 18 clearly illustrate that although experts assert oral hormonal contraceptives as a major contributor to cervical cancer, our proposed method assigned it the least weight, W1. Additionally, experts emphasized the significance of early sexual intercourse, a crucial aspect that was not addressed in the proposed model. This indicates that even with the impressive capabilities of advanced AI and explainability tools in uncovering hidden patterns in patient data and identifying key features that influence model outcomes, there is still room for improvement to match the expertise of human domain experts. Further research and refinement are essential to ensure these AI models reach their full potential in the medical field. However, this comparison provided detailed insights into the decision-making processes of both the model and the experts, effectively bridging the gap.



**Figure 13.** Prediction probability of an individual instant and features impact on it using local interpretable model-agnostic explanations (LIME).



**Figure 14.** Comparison of 10 random instants feature impact.

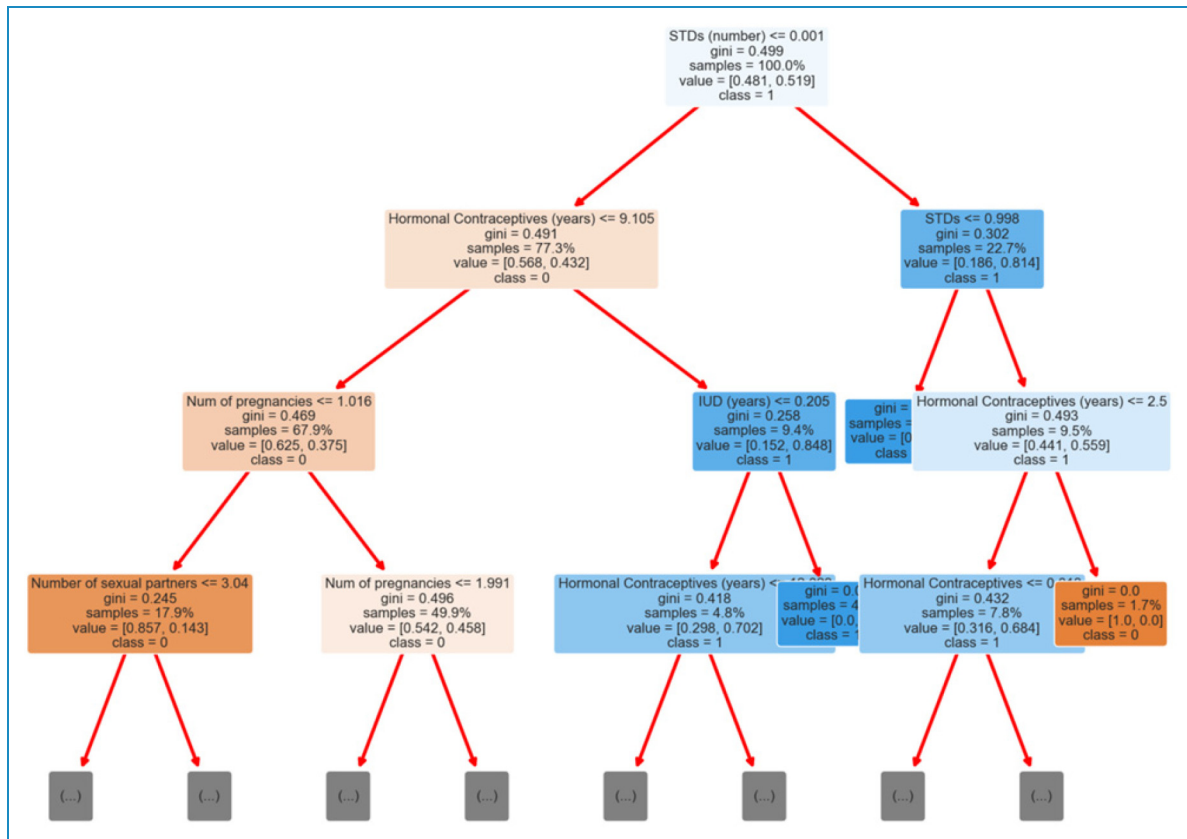


Figure 15. Decision tree (DT) surrogate model explainability in depth 3.

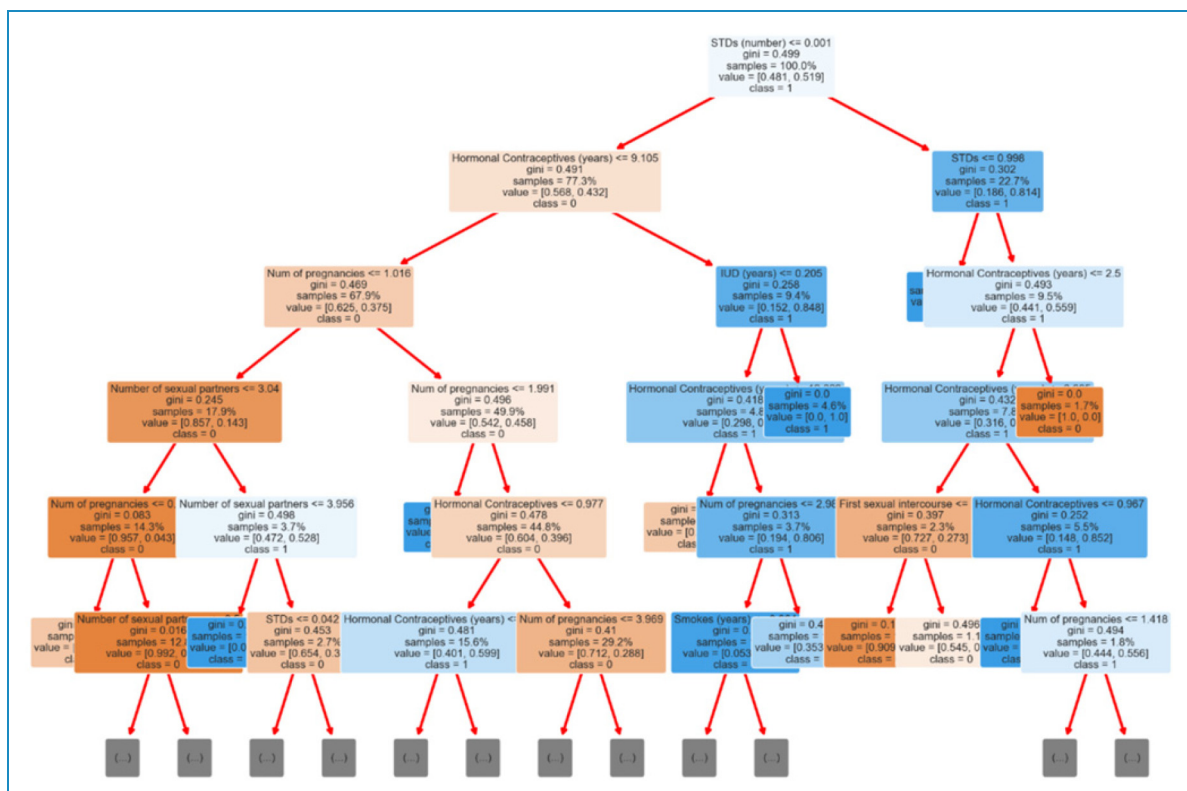
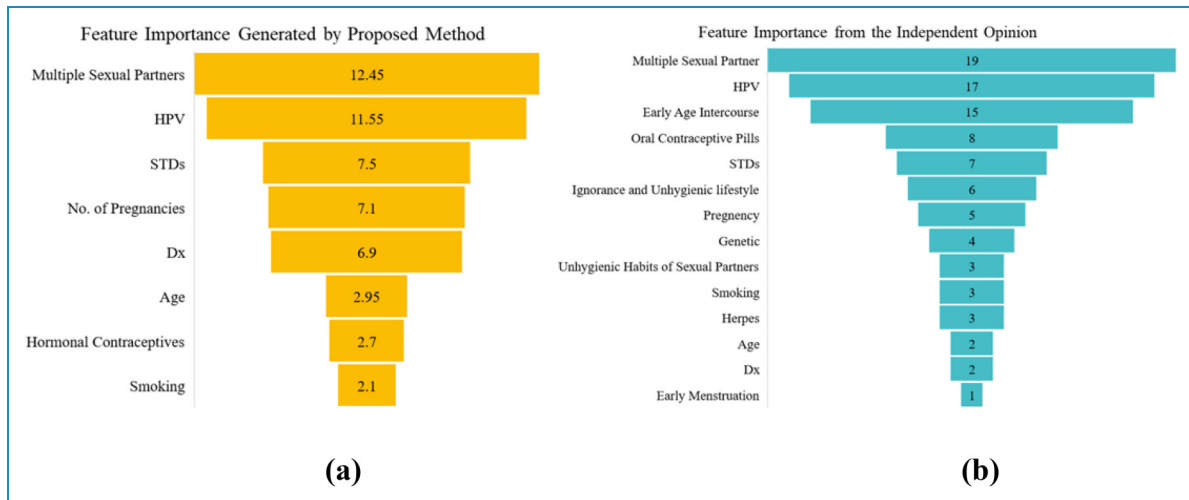
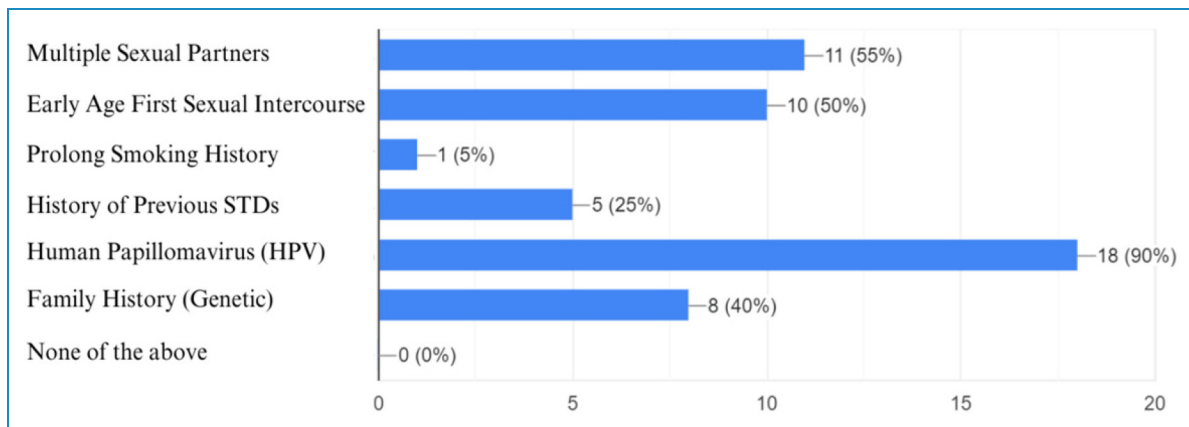


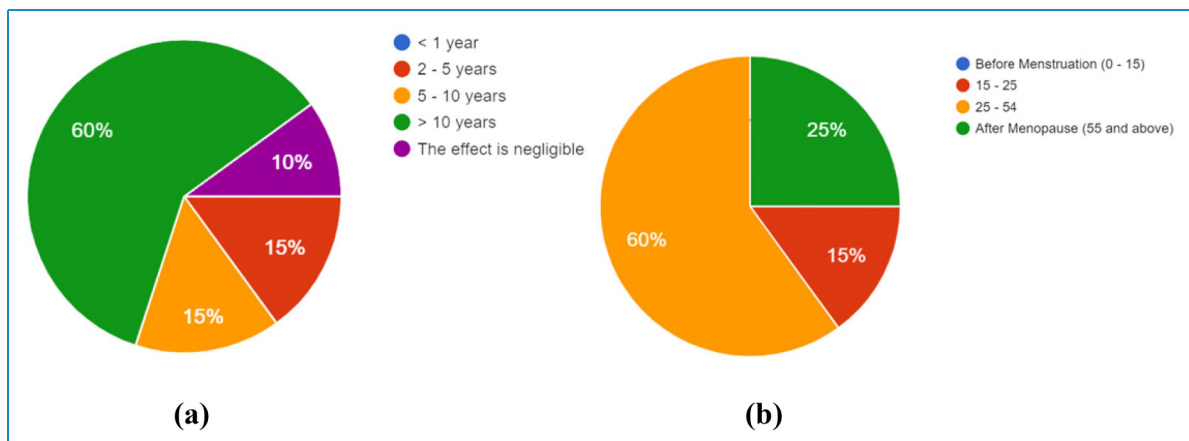
Figure 16. Decision tree (DT) surrogate model explainability in depth 5.



**Figure 17.** Importance of individual features from the domain experts.



**Figure 18.** Major risk factors for cervical cancer listed by domain experts.



**Figure 19.** Visualizing the insights provided by domain experts: (a) How many years of having oral contraceptive pills can lead to cervical cancer? (b) What age range do you consider women more susceptible to contracting cervical cancer?.

**Table 10.** Comparison with recent works.

Article	Year	Dataset	Method	Accuracy
Hariprasad et al. <sup>16</sup>	2023	UCI risk factors dataset	RF with oversampling	98.7%
Ali et al. <sup>18</sup>	2024	Both UCI datasets	RF-based ensemble model	98.06% on (DS-I) and 94.45% on (DS-II)
Kumari et al. <sup>30</sup>	2023	UCI risk factors dataset	DNN with PCA	87.4%
Khanam and Mondal <sup>17</sup>	2021	UCI risk factors dataset	Stacked Ensemble	94.4%
Kumawat et al. <sup>22</sup>	2023	UCI risk factors dataset	XGB	94.94%
Ilyas and Ahmad <sup>21</sup>	2021	Both UCI datasets	Voting ensemble	94%
<b>This study</b>	<b>2024</b>	<b>UCI risk factors dataset</b>	<b>Stacked ensemble with RF as meta-learner</b>	<b>98%</b>

RF: random forest; DNN: deep neural network; PCA: principal component analysis; XGB: extreme gradient boosting.

## Discussion

Medical diagnosis data is used to create highly accurate advanced machine-learning models that can make precise predictions. These models serve as valuable tools for healthcare professionals, empowering them to make more accurate and informed diagnoses for their patients. However, the primary hurdles lie in the limited access to real-world datasets and their imbalanced nature. It negatively affects the ML model's overall prediction, leading to a bias toward one class and subpar performance when classifying the minority class. The proposed study introduces a two-level pipeline that combines a hybrid feature selection method with an ensemble model and domain knowledge. This approach is designed to address the complexities of this challenging task effectively. By incorporating a hybrid feature selection method and stacking ensemble model, the proposed methodology aims to address and mitigate the issue of biased predictions and improve the model's performance in classifying both the majority and the minority class. This comprehensive approach ensures that the model not only guarantees accuracy for individual classes but also exhibits the ability to recognize both positive and negative outcomes. Table 10 presents the overall comparison of our proposed pipeline with other recent research.

We experimented with various oversampling and under-sampling techniques such as SMOTE, SMOTETomek, ADASYN, and NearMiss to assess their effects and compare the predictive results with the original imbalanced dataset. By applying these techniques, this study observed how each method affected the model's performance, allowing the selection of the most appropriate approach for the

enhanced performance and stability of the ML models. In this case, it has been shown that oversampling techniques outperformed undersampling techniques, with SMOTE being the most effective. Along with our proposed SRXL model, other ML models performed better, such as DT, LR, CB, etc. This emphasizes the importance of preprocessing, as proposed in the study, which can significantly enhance the performance of ML models. Moreover, this study conducted a thorough analysis of the data, harnessing the potential of ML to uncover concealed patterns within the dataset. The findings revealed novel insights and correlations that were previously unrecognized. Additionally, the results obtained from the analysis of different generative AI approaches indicate their incapability in handling tabular datasets. The proposed hybrid feature selection method effectively ranked the most important features contributing to the model's predictions. The features were subsequently utilized to develop a predictive model, showcasing remarkable accuracy in forecasting the target outcome. Furthermore, the study conducted an in-depth analysis of the importance of features that further facilitate the early prediction of cervical cancer and ensure timely treatment.

Furthermore, tools for explainability, such as SHAP, ELI5, and LIME, have increased the transparency and reliability of the pipeline. This incorporated advanced automated AI algorithms in the professional health domain. SHAP and LIME are powerful tools for providing global and local explanations. We integrated the expert opinion of the domain and healthcare professionals to map the findings. We facilitate one-to-one interviews to assign weights to individual features and rank their contribution. By conducting personalized interviews, this study evaluates the



importance of different factors and establishes their impact on the final ranking. Therefore, healthcare professionals can better understand the decision-making process. This potentially increases the acceptance and adoption of the developed predictive model in clinical practice.

### Limitations

This study is based on a single publicly available dataset. Additionally, we collected expert opinions from domain experts in a single country. Ideally, gathering insights from domain experts across different regions would enhance the study's robustness, but due to resource constraints, this was not feasible. Incorporating more diverse perspectives from experts and collecting additional data would further improve the study's explainability and practical applicability.

### Conclusion and future work

In conclusion, our study addresses the intricate challenges of real-world data analysis, particularly focusing on the "Cervical Cancer (Risk Factors)" dataset. The prevalence of imbalanced class distributions, noise, missing values, and duplicate records in healthcare datasets emphasizes the need for a meticulous preprocessing pipeline. Our proposed stacking ensemble model has exhibited exceptional performance, achieving a remarkable 98% accuracy and an outstanding AUC score of 0.995, surpassing traditional ML and deep learning models.

Applying SHAP, LIME, ELI5, and DT surrogate models enhances our model's interpretability. We have successfully ranked influential features such as hormonal contraceptives, women's age, sexual partners, sexually transmitted disease (STD) infections, and other habits, providing valuable insights into cervical cancer predictions. Validation through expert interviews reinforces the robustness of our model.

Further work will focus on integrating more datasets in this domain in a distributed manner where datasets are not gathered centrally to enhance security. More relevant features will be added to analyze cervical cancer more deeply. Additionally, we will explore ways to mitigate the mismatch in domain experts' opinions on the explainability of the XAI tools.


### Acknowledgments

We thank the Computer Science and Engineering Department of Hajee Mohammad Danesh Science and Technology University, Bangladesh, for supporting the study process. Additionally, we extend our deepest gratitude to the Center for Multidisciplinary Research and Development (CeMRD) for their resources, guidance, and support. The expertise and encouragement from CeMRD members have been invaluable. This research was made possible by the collaborative environment and cutting-edge facilities at CeMRD.

### Guarantor

MPU

### ORCID iD

Md Palash Uddin  <https://orcid.org/0000-0002-4429-6590>

### Statements and declarations

#### Ethical considerations

This study was conducted following the ethical guidelines. Prior to participation, all individuals received comprehensive information about the study and written informed consent was necessary.

#### Informed consent

All participants provided written informed consent before taking part in the study.

#### Author contributions/CRedit

Priyanka Roy: Conceptualization, methodology, software, validation, data curation, data analysis, visualization, and manuscript drafting. Mahmudul Hasan: Conceptualization, methodology, investigation, validation, supervision, and manuscript review. Md Rashedul Islam: Conceptualization, methodology, resources, investigation, and manuscript drafting. Md Palash Uddin: Methodology, investigation, manuscript review, manuscript finalization, and supervision.

#### Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

#### Conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

#### Data availability

The dataset utilized in this study is available in Kelwin and Fernandes.<sup>19</sup>

### References

- Schulz WA. *Molecular Biology of Human Cancers*. Berlin/Heidelberg, Germany: Springer, 2023 Feb 28.
- World Health Organization. Cervical cancer. <https://www.who.int/news-room/fact-sheets/detail/cervical-cancer> (2023, accessed 19 December 2023).
- Sung H, Ferlay J, L Siegel R, et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2021; 71: 209–249.
- Shrestha AD, Neupane D, Vedsted P, et al. Cervical cancer prevalence, incidence and mortality in low and middle income countries: a systematic review. *Asian Pac J Cancer Prev* 2018; 19: 319.
- Canfell K, J Kim J, Brisson M, et al. Mortality impact of achieving who cervical cancer elimination targets: a comparative modelling analysis in 78 low-income and lower-middle-income countries. *Lancet* 2020; 395: 591–603.

6. Kessler TA. Cervical cancer: prevention and early detection. *Semin Oncol Nurs* 2017; 33: 172–183. Cancer screening and early detection.
7. Gheisari M, Ghaderzadeh M, Li H, et al. Mobile apps for COVID-19 detection and diagnosis for future pandemic control: multidimensional systematic review. *JMIR Mhealth Uhealth* 2024; 12: e44406.
8. Roy P, Sriyon FMS, Hasan M, et al. An ensemble machine learning approach with hybrid feature selection technique to detect thyroid disease. In: *Proceedings of the 2nd International conference on Big Data, IoT and machine learning*, pp.379–394. Singapore: Springer Nature Singapore, 2024.
9. Ghaderzadeh M, Shalchian A, Irajian G, et al. Artificial intelligence in drug discovery and development against antimicrobial resistance: a narrative review. *Iran J Med Microbiol* 2024; 18: 135–147.
10. Xue P, Ng MTA and Qiao Y. The challenges of colposcopy for cervical cancer screening in LMICs and solutions by artificial intelligence. *BMC Med* 2020; 18: 1–7.
11. Hafsa H and Suril G. Machine learning in healthcare. *Curr Genomics* 2021; 22: 291–300 .
12. Tarawneh AS, Hassanat AB, Altarawneh GA, et al. Stop oversampling for class imbalance learning: a review. *IEEE Access* 2022; 10: 47643–47660.
13. Hasan M, Haque A, Islam MM, et al. How much do the features affect the classifiers on UNSW-NB15? An XAI equipped model interpretability. In: Abedin MZ and Hajek P (eds) *Cyber security and business intelligence*. Routledge, 2023, pp.137–153.
14. Hassan MM, Abrar MF and Hasan M. An explainable AI-driven machine learning framework for cybersecurity anomaly detection. In: Abedin MZ and Hajek P (eds): *Cyber security and business intelligence*. Routledge, 2023, pp.197–219.
15. Wu W and Zhou H. Data-driven diagnosis of cervical cancer with support vector machine-based approaches. *IEEE Access* 2017; 5: 25189–25195.
16. Hariprasad R, Navamani TM, Rote TR, et al. Design and development of an efficient risk prediction model for cervical cancer. *IEEE Access* 2023; 11: 74290–74300.
17. Khanam F and Mondal MRH. Ensemble machine learning algorithms for the diagnosis of cervical cancer. In: *2021 International conference on science & contemporary technologies (ICSCT)*, Dhaka, Bangladesh, 5–7 August 2021 pp.1–5. IEEE . DOI: 10.1109/ICSCT53883.2021.9642612.
18. Ali MdS, Hossain MdM, Kona MA, et al. An ensemble classification approach for cervical cancer prediction using behavioral risk factors. *Healthc Anal* 2024; 5: 100324.
19. Kelwin FCJ and Fernandes J. Cervical cancer (Risk Factors). UCI machine learning repository, 2017. DOI: 10.24432/C5Z310.
20. Sobar, Machmud R and Wijaya A. Behavior determinant based cervical cancer early detection with machine learning algorithm. *Adv Sci Lett* 2016; 22: 3120–3123.
21. Ilyas QM and Ahmad M. An enhanced ensemble diagnosis of cervical cancer: a pursuit of machine intelligence towards sustainable health. *IEEE Access* 2021; 9: 12374–12388.
22. Kumawat G, Vishwakarma SK, Chakrabarti P, et al. Prognosis of cervical cancer disease by applying machine learning techniques. *J Circuits Syst Comput* 2023; 32: 2350019.
23. Song Y, Tan EL, Jiang X, et al. Accurate cervical cell segmentation from overlapping clumps in pap smear images. *IEEE Trans Med Imaging* 2017; 36: 288–300.
24. Lee H and Kim J. Segmentation of overlapping cervical cells in microscopic images with superpixel partitioning and cell-wise contour refinement. In: *2016 IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, Las Vegas, NV, USA, 26 June–01 July 2016, pp.1367–1373. IEEE. DOI: 10.1109/CVPRW.2016.172.
25. Su J, Xu X, He Y, et al. Automatic detection of cervical cancer cells by a two-level cascade classification system. *Anal Cell Pathol* 2016; 2016: 9535027.
26. Devi MA, Sheeba J and Joseph KS. Neutrosophic graph cut-based segmentation scheme for efficient cervical cancer detection. *J King Saud Univ-Comput Inf Sci* 2022; 34: 1352–1360.
27. Marinakis Y, Marinaki M, Dounias G, et al. Intelligent and nature inspired optimization methods in medicine: the pap smear cell classification problem. *Expert Syst* 2009; 26: 433–457.
28. Ali MM, Ahmed K, Bui FM, et al. Machine learning-based statistical analysis for early stage detection of cervical cancer. *Comput Biol Med* 2021; 139: 104985.
29. Ratul IJ, Al-Monsur A, Tabassum B, et al. Early risk prediction of cervical cancer: a machine learning approach. In: *2022 19th international conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON)*, Prachuap Khiri Khan, Thailand, 24–27 May 2022, pp.1–4. IEEE. DOI: 10.1109/ECTI-CON54298.2022.9795429.
30. Kumari CM, Bhavani R, Padmashree S, et al. Automated cervical cancer classification using deep neural network classifier. *Int J Model Simul Sci Comput* 2024; 15: 2450008.
31. Glučina M, Lorencin A, Anelić N, et al. Cervical cancer diagnostics using machine learning algorithms and class balancing techniques. *Appl Sci* 2023; 13: 1061.
32. AlMohimeed A, Saleh H, Mostafa S, et al. Cervical cancer diagnosis using stacked ensemble model and optimized feature selection: an explainable artificial intelligence approach. *Future Syst Based Healthc* 2023; 12: 200.
33. Curia F. Cervical cancer risk prediction with robust ensemble and explainable black boxes method. *Health Technol (Berl)* 2021; 11: 875–885.
34. Hasan M, Rabbi MF, Sultan MN, et al. A novel data balancing technique via resampling majority and minority classes toward effective classification. *TELKOMNIKA (Telecom Comput Electron Control)* 2023; 21: 1308–1316.
35. Mokoatle M, Coleman T and Mokilane P. A comparative study of over-sampling techniques as applied to seismic events. In: *Southern African conference for artificial intelligence research*, Muldersdrift, South Africa, December 4–8 2023, pp.331–345. Springer.

36. Hasan M, Islam MM, Sajid SW, et al. The impact of data balancing on the classifier's performance in predicting cesarean childbirth. In: *2022 4th International conference on electrical, computer & telecommunication engineering (ICECTE)*, Rajshahi, Bangladesh, 29–31 December 2022, pp.1–4. IEEE.
37. Sahid MA, Hasan M, Akter N, et al. Effect of imbalance data handling techniques to improve the accuracy of heart disease prediction using machine learning and deep learning. In: *2022 IEEE region 10 symposium (TENSYP)*, Mumbai, India, 01–03 July 2022, pp.1–6. IEEE.
38. Rousseeuw PJ and Hubert M. Robust statistics for outlier detection. *Wiley Interdiscip Rev: Data Min Knowl Discov* 2011; 1: 73–79.
39. Sánchez-Marono N, Alonso-Betanzos A and Tombilla-Sanromán M. Filter methods for feature selection—a comparative study. In: *International conference on intelligent data engineering and automated learning*, Birmingham, United Kingdom, 16–19 December 2007, pp.178–187. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.
40. Maldonado S and Weber R. A wrapper method for feature selection using support vector machines. *Inf Sci (NY)* 2009; 179: 2208–2217.
41. Bhavsar H and Panchal MH. A review on support vector machine for data classification. *Int J Adv Res Comput Eng Technol* 2012; 1: 185–189.
42. Somvanshi M, Chavan P, Tambade S, et al. A review of machine learning techniques using decision tree and support vector machine. In: *2016 International conference on computing communication control and automation (ICCUBEA)*, Pune, India, 12–13 August 2016, pp.1–7, IEEE. DOI: 10.1109/ICCUBEA.2016.7860040.
43. Charbuty B and Abdulazeez A. Classification based on decision tree algorithm for machine learning. *J Appl Sci Technol Trend* 2021; 2: 20–28.
44. Musa AB. Comparative study on classification performance between support vector machine and logistic regression. *Int J Mach Learn Cybern* 2013; 4: 13–24.
45. Mucherino A, Papajorgji PJ, Pardalos PM, et al. K-nearest neighbor classification. *Data Min Agric* 2009; 34: 83–106.
46. Breiman L. Random forests. *Mach Learn* 2001; 45: 5–32.
47. Liu Y, Wang Y and Zhang J. New machine learning algorithm: random forest. In: *Information computing and applications: third international conference, ICICA 2012*, Chengde, China, September 14–16, 2012. Proceedings 3, pp.246–252. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
48. Bentéjac C, Csörgo A and Martínez-Muñoz G. A comparative analysis of gradient boosting algorithms. *Artif Intell Rev* 2021; 54: 1937–1967.
49. Ibrahim AA, Ridwan RL, Muhammed MM, et al. Comparison of the catboost classifier with other machine learning methods. *Int J Adv Comput Sci Appl* 2020; 11.
50. Prokhorenkova L, Gusev G, Vorobev A, et al. CatBoost: unbiased boosting with categorical features. In: *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada, 3–8 December 2018, 2018.
51. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2001; 29: 1189–1232.
52. Such FP, Rawal A, Lehman J, et al. Generative teaching networks: accelerating neural architecture search by learning to generate synthetic training data. In: *Proceedings of the 37th international conference on machine learning*, Vienna, Austria, 13–18 July 2020, Paper No. 854, 119: 9206–9216. 2020.
53. Kalule R, Abderrahmane HA, Alameri W, et al. Stacked ensemble machine learning for porosity and absolute permeability prediction of carbonate rock plugs. *Sci Rep* 2023; 13: 9855.
54. Wang M, Qian Y, Yang Y, et al. Improved stacking ensemble learning based on feature selection to accurately predict warfarin dose. *Front Cardiovasc Med* 2024; 10: 1320938.
55. Hw Zhang, Yr Wang, Hu B, et al. Using machine learning to develop a stacking ensemble learning model for the CT radiomics classification of brain metastases. *Sci Rep* 2024; 14: 28575.
56. Shuai Y, Zheng Y and Huang H. Hybrid software obsolescence evaluation model based on PCA-SVM-GridSearchCV. In: *2018 IEEE 9th international conference on software engineering and service science (ICSESS)*, Beijing, China, 23–25 November 2018, pp.449–453. IEEE, 2018.
57. Tjoa E and Guan C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Trans Neural Netw Learn Syst* 2020; 32: 4793–4813.
58. Hasan M, Roy P and Nitu AM. Cervical cancer classification using machine learning with feature importance and model explainability. In: *2022 4th International conference on electrical, computer & telecommunication engineering (ICECTE)*, Rajshahi, Bangladesh, 29–31 December 2022, pp.1–4. IEEE.
59. Lundberg SM and Lee SI. A unified approach to interpreting model predictions. In: *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 4–9 December 2017, pp.1–10, 2017.
60. Shapley L. A value for n-person games, contributions to the theory of games. *Ann Math Stud* 1953; 28: 307–317.
61. Ribeiro MT, Singh S and Guestrin C. “Why should I trust you?” Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco, CA, USA, 13–17 August 2016, pp.1135–1144.
62. Melo E, Silva I, Costa DG, et al. On the use of explainable artificial intelligence to evaluate school dropout. *Educ Sci* 2022; 12: 845.
63. Jia S, Lin P, Li Z, et al. Visualizing surrogate decision trees of convolutional neural networks. *J Vis* 2020; 23: 141–156.