

Predictive modeling and web-based tool for cervical cancer risk assessment: A comparative study of machine learning models



Ritu Chauhan^{a,*}, Anika Goel^a, Bhavya Alankar^b, Harleen Kaur^b

^a Artificial Intelligence and IoT Automation Lab, Center for Computational Biology and Bioinformatics, Amity University, Noida, Uttar Pradesh 201313, India

^b Department of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi 110062, India

ARTICLE INFO

Method name:

Predictive Modeling and Web-Based Tool for Cervical Cancer Risk Assessment: A Comparative Study of Machine Learning

Keywords:

Cervical cancer
Predictive modeling
Machine learning
XGBoost
Web-based tool
Risk assessment
Early detection

ABSTRACT

In today's digital era, the rapid growth of databases presents significant challenges in data management. In order to address this, we have developed and designed CHAMP (Cervical Health Assessment using machine learning for Prediction), which is a user interface tool that can effectively and efficiently handle cervical cancer databases to detect patterns for future prediction diagnosis. CHAMP employs various machine learning algorithms which include XGBoost, SVM, Naive Bayes, AdaBoost, Decision Tree, and K-Nearest Neighbors in order to predict cervical cancer accurately. Moreover, this tool also designates to evaluate and optimize processes, to retrieve the significantly augmented algorithm for predicting cervical cancer. Although, the developed user interface tool was implemented in Python 3.9.0 using Flask, which provides a personalized and intuitive platform for pattern detection. The current study approach contributes to the accurate prediction and early detection of cervical cancer by leveraging the power of machine learning algorithms and comprehensive validation tools, which aim to provide learned decision-making.

- CHAMP is a user interface tool which is designed for the detection of patterns for future diagnosis and prognosis of cervical cancer.
- Various machine learning algorithms are employed for accurate prediction.
- This tool provides personalized and intuitive data analysis which enables informed decision-making in healthcare.

* Corresponding author.

E-mail address: Rituchauha@gmail.com (R. Chauhan).

Specification Table

Subject Area	Bioinformatics
More Specific Subject Area:	Web-Based Tool for Cervical Cancer Risk Assessment using predictive modeling
Name of Method:	Predictive Modeling and Web-Based Tool for Cervical Cancer Risk Assessment: A Comparative Study of Machine Learning
Name and reference of the original method:	<ol style="list-style-type: none"> 1. Alsmariy, Riham, Graham Healy, and Hoda Abdelhafez. 2020. "Predicting Cervical Cancer Using Machine Learning Methods." <i>International Journal of Advanced Computer Science and Applications</i> 11(7): 173–84. 2. Deng, Xiaoyu, Yan Luo, and Cong Wang. 2019. "Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods." <i>Proceedings of 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems, CCIS 2018</i>: 631–35. 3. Jones, M., Lee, S., & Smith, R. (2022). Building Interactive Dashboards with Flask and Python. <i>Journal of Web Development</i>, 15(3), 45–62.
Resource Availability:	Python 3.9.0

Introduction

Artificial Intelligence (AI) has emerged as a transformative technology in the healthcare sector, revolutionizing the way medical data is analysed, diagnoses are conducted, and treatments are administered. These technologies have the potential to improve patient outcomes, enhance efficiency, and reduce costs in healthcare systems worldwide. We can say that AI-based technology has algorithmic supremacy which can predict disease progression, treatment responses, and patient outcomes by considering various factors such as patient characteristics, genetic profiles, and treatment histories [1]. This predictive capability can assist healthcare providers in making informed decisions about treatment options, optimizing resource allocation, and designing personalized interventions. So, meaningful insights can be fetched from the above technology which has the potential to analyze vast amounts of medical data, including patient records, images, genetic information, and clinical research for decision-making. These insights can aid in early detection and diagnosis of diseases, identification of high-risk individuals, and personalized treatment planning. For example, ML algorithms can analyze medical images such as X-rays, MRIs, and CT scans to detect abnormalities or assist radiologists in making accurate diagnoses [2].

So, the assistive technology of AI is increasing at the unprecedented rate in healthcare industry in correspondence with exponential growth of data. However, the major focus of AI is to transform the healthcare data such as patient care data, Emergency data, insurance data, imaging data and into meaningful tools and technologies which can benefit the prognosis and the diagnosis of the disease. In addition, AI can not be termed as a single technology to handle the burden of data, but rather it tends to be integral form of several other technologies. Such technologies which can be of immediate relevance to healthcare is discussed below:

1. Deep Learning with Neural Network

Deep learning has changed the era of medical technology by determine the factors which might be responsible for the progression of the disease. Moreover, the deep learning concept is based upon the human brain structure where each neurons perform some individual tasks and confers the decision making. For example: the medical image data has several features such as the cancerous and non-cancerous lesions which are difficult to determine by human eye al alone. So, deep learning has unprecedentedly perceived the relevant clinical features which may be responsible for the automated patterns detection.

2. Medical Image Analysis

The diagnosis of medical data is a trivial task as there exists several features which may or may not be responsible for the prognosis of disease. Moreover, current high-end technology of image analysis such as CT scan, radio imaging, resonance imaging has certainly modified the diagnostic system in medical diagnosis. Such databases, generates huge number of images which constantly require a high-end automation technology to process and generate effective and efficient patterns for diagnosis of the disease. To deal, with this ongoing knowledge extraction machine learning technology has proven to be a mechanized technology which can deliver some extended knowledge which can be supportive to healthcare practioners. However, the complex databases cannot be alone formalised with machine learning technology hence assistive support is required from the healthcare researchers.

3. Natural Language Processing (NLP)

The Language processing is the powerful tool where the language is referred to as "speech". Undoubtedly, NLP has brought the revolution in the spectrum of healthcare databases where the speech recognition, voice recognition, translation speech has created impulsive contribution for patient diagnosis system. Moreover, the diagnostic system of NLP has been discussed with two technical approaches such as Statistical and Semantic NLP. The statistical approach is based on the concept of machine learning whereas the semantic NLP is based on the complex nature of text data which includes the words with different arrangements, clauses and words which have intermediate relation with each other.

However, AI technologies has evidently surfaced several application domains but certainly healthcare data has laid several challenges to extract knowledge due several encounters which include the complex and unstructured nature of data with privacy concerns to access the features which can be the most relevant for the prognosis of the disease. Also, several exertions are subjective to adopt comparative algorithmic technology to retrospect and investigate the different population samples and characteristics, but no foreseen results are disclosed for real diagnosis. So, it becomes really difficult for healthcare practioners to characterize and make decisions based on implicated studies suggested by AI for patient diagnosis. Moreover, data generated by these advent technologies must be

predicted in correspondence with emerging expertise to discover effective and efficient hidden patterns. In addition, the amount of data engendered is complex and available in an unstructured format which has led to complexity among the databases and difficulty to evaluate and extract knowledge that is hidden among the data. In the past, several studies have been elaborated on and discussed in correspondence with varied machine learning technology which can benefit healthcare practitioners and other working professionals to interpret data and predict results [3]. We can say, machine learning comes under the umbrella of Artificial Intelligence which amply works on generic databases to retrieve or discover effective patterns for decision-making [3]. In similar, machine learning algorithms are designed to extract knowledge from previously sampled data and try to interpret patterns from these data. Additionally, if we look at the healthcare sector, ML has resulted in an increase in the accuracy of predicting the occurrence of diseases, disorders, and their classification by practitioners [3,4]. Therefore, researchers are continuously working to build better-performing models to interpret healthcare datasets to provide better predictions that help medical practitioners give the right results in diagnosing diseases.

Moreover, in the healthcare domain, cervical cancer remains a significant concern worldwide. The expansion of cervical cancer can be attributed for various factors, including HPV infection, viral infection, and the persistent progression of infected cell clones, leading to pre-cancerous changes and invasive tumours. Cervical cancer is responsible for over 300,000 deaths, predominantly occurring in low-income or working-class countries which is due to limited access to screening and vaccination programs [4,5]. Due to limited screening and treatment leads to late-stage diagnosis. However, reasons may be the cultural taboos and stigma hinder discussions on reproductive health, while healthcare resource shortages and poverty impede prevention effort. On the other hand, developed countries have witnessed a reduction in cervical cancer cases and deaths through the implementation of formal screening programs. However, in countries like India, cervical cancer continues to pose a serious threat, accounting for the highest number of cases in Asia and contributing to a significant percentage of global deaths. Additionally, early detection of cervical cancer is challenging due to the absence of noticeable symptoms in its initial stages, emphasizing the importance of regular check-ups for timely diagnosis as its late detection can result in cancer spreading to other vital organs, intensifying the difficulties faced by patients. To address these issues, comprehensive strategies are required which includes improved healthcare infrastructure, awareness campaigns, vaccination programs, and effort to reduce socioeconomic disparities [6–8]. Therefore, by understanding the risk factors and their impact on the development of cervical cancer, healthcare practitioners can enhance their diagnostic capabilities and develop strategies for early detection and treatment. The present study aims to contribute to this knowledge by exploring predictive modeling techniques and developing a web-based tool for cervical cancer risk assessment, empowering medical professionals to make informed decisions and improve patient outcomes.

However, the Early detection of cervical cancer significantly impacts patient outcomes by leading to increased treatment success, reduced disease progression, improved survival rates, and preservation of fertility. Moreover, early detection prevents the cancer from advancing to more difficult-to-treat stages and reduces the need for extensive interventions, resulting in lower healthcare costs and a better quality of life for patients [6]. Additionally, early detection leads to less aggressive treatment options such as surgery or radiation therapy, which greatly enhance the chances of complete remission and cure. On the other hand, psychologically, early detection alleviates the emotional distress associated with a cancer diagnosis, while raising awareness about the benefits and encourages individuals to participate in regular screening programs. Ultimately, emphasizing the importance of early detection contributes to better patient outcomes and improved overall public health [9–11].

In the healthcare domain, machine learning has emerged as a trustworthy system for interpreting complex datasets and for clinical decision-making with strong data analysis capabilities [11–14]. This research focuses on applicability of machine learning algorithms on basis of their sensitivity and accuracy. Also, we have elaborated on varied statistical measures which are applied to determine a best performing model for predicting cervical cancer. Cervical cancer databases were extracted from public domain repository where it consisted of 36 features and 858 observations. This dataset is composed of the varied features such as demographic information such as the age of the individuals, the number of sexual partners in past and present, age at which they had their first sexual intercourse, and the number of pregnancies they have experienced. Additionally, the dataset incorporates lifestyle factors such as smoking habits, the use of hormonal contraceptives, and the presence of sexually transmitted diseases. The dataset was discussed on the factors which may or may not responsible for the risk of cervical cancer. This dataset focuses on capturing a wide range of clinical and demographic factors that are associated with the development and progression of cervical cancer in women. The features are crucial in understanding the complex nature of the disease and identifying potential risk factors that contribute to its occurrence.

This study focuses on the design and development of a user interface tool called CHAMP (Cervical Health Assessment using Machine learning for Prediction) that effectively handles cervical cancer databases and detects patterns for future diagnosis. The tool has been specifically designed to prioritize user interaction and accessibility for healthcare practitioners. By deploying machine learning algorithms such as XGBoost, SVM, Naive Bayes, AdaBoost, Decision Tree, and K-Nearest Neighbors study aims to measure the prediction accuracy of cervical cancer databases and optimize the algorithm selection process. The user interface tool, implemented in Python 3.9.0 using Flask, provides a personalized and intuitive platform for pattern detection and analysis. Through this approach, the study contributes to the accurate prediction and early detection of cervical cancer, empowering healthcare professionals with informed decision-making capabilities.

However, machine learning algorithms offer valuable applications in predicting and preventing cervical cancer. By analyzing various risk factors, patient data, and HPV infection status, these algorithms can provide risk assessments that help healthcare providers identify individuals at a higher risk of developing cervical cancer. Moreover, machine learning can enhance the accuracy of HPV risk assessments by considering multiple parameters and historical data. In the context of cervical cancer screening, machine learning can assist in the analysis of Pap smear slides, improving the accuracy of diagnoses by classifying cell samples as normal, precancerous, or cancerous. Additionally, machine learning can aid in the analysis of cervical images, detecting abnormal tissue patterns and lesions indicative of cervical cancer. Moreover, early detection models can predict the likelihood of developing cervical cancer within a spe-

cific timeframe, facilitating timely interventions and personalized treatment plans. On a broader scale, machine learning can analyze population health data to inform public health strategies, vaccination campaigns, and screening programs. However, it also has the potential to enable remote monitoring and patient education, particularly in underserved or remote areas. While machine learning holds great promise, its integration into clinical practice must consider ethical, legal, and privacy concerns to ensure its responsible use in cervical cancer prevention efforts.

Moreover, this tool plays a pivotal role in advancing the accurate prediction and early detection of cervical cancer by incorporating a comprehensive array of risk factors, such as age, HPV infection status, Pap smear results, sexual behavior, and smoking history, the CHAMP tool offers a holistic perspective on an individual's risk profile. This personalized approach is invaluable for optimizing preventive measures and focusing resources where they are most needed.

Furthermore, the CHAMP tool's ability to identify high-risk individuals is closely intertwined with the early detection of cervical cancer and by efficiently flagging those at elevated risk, the tool empowers healthcare professionals to recommend more frequent screenings, employ advanced diagnostic tests, and initiate follow-up protocols promptly. This strategic approach significantly enhances the likelihood of catching precancerous changes or early-stage cancers before they progress, thus improving treatment outcomes and reducing the burden of the disease. In effect, the CHAMP tool's precision and tailored approach contribute to a shift from a one-size-fits-all strategy to a more individualized and effective cervical cancer prevention paradigm. In addition, the objective of the current study of approach is to design and develop an interface which can handle the features of Cervical cancer database with machine learning algorithms for prediction and prognosis of the disease. The intuitive CHAMP tool is a personalized tool to enable informed decision making using machine learning algorithms such as XGBoost, SVM, Naive Bayes, AdaBoost, Decision Tree, and K-Nearest Neighbors for retrieval of information which can be relevant for the prognosis of the disease.

However, the impact of the CHAMP tool extends beyond individual patient care. Aggregated data from its assessments can inform broader public health initiatives by offering insights into cervical cancer trends, risk factor prevalence, and efficacy of prevention strategies. This data-driven approach guides the development of targeted awareness campaigns, vaccination programs, and screening efforts, leading to a holistic improvement in population health outcomes. In essence, the CHAMP tool's multifaceted contributions, ranging from accurate risk prediction and early detection to supporting public health initiatives, position it as a crucial asset in the ongoing battle against cervical cancer.

Finally, rest of the paper is organized as follows Section 2 previous literature reviews suggest the context of cervical cancer with new adoptive technology. The overall approach of the study is elaborated in Section 3, and the results are explained in Section 4, and in the last conclusion is discussed

Literature review

With advancements in machine learning and web development, there is a growing interest in utilizing these technologies to improve cervical cancer prediction and visualization. In this section, we explore existing research and studies that have focused on integrating machine learning algorithms with web development techniques for cervical cancer prediction and visualization. Although, several studies have demonstrated the potential of machine learning algorithms in accurately predicting cervical cancer outcomes by leveraging various features such as demographic information, medical history, and biomarkers to train predictive models [14–17]. In parallel, the development of interactive web-based tools has gained grip as an effective way to present complex data and facilitate user engagement. These tools provide a user-friendly interface where clinicians and healthcare professionals can input patient data and obtain real-time predictions for cervical cancer risk. Additionally, visualization techniques such as pie charts, bar graphs, and heatmaps have been employed to effectively communicate the predictive insights derived from the machine learning models. While previous studies have individually explored machine learning algorithms and web development techniques for cervical cancer prediction and visualization, there is a lack of comprehensive research that integrates both aspects into a unified framework. This research aims to bridge this gap by combining machine learning algorithms with modern web development practices to create a robust and user-friendly predictive and visualization tool for cervical cancer. In earlier studies, we can see many ML methodologies have been utilized for cervical cancer diagnosis as in one of the studies, researchers showed that an oversampling technique known as SMOTE increased the performance of traditionally working models. In this work, the researcher builds a model using a random forest that performs well after applying two feature-selection techniques [1]. The two techniques were based on specificity and positive predictive accuracy. The two methods used were Principal component analysis (PCA) and recursive feature elimination. Although, it was not clear how this technique would not result in increasing the accuracy [1]. In [18] another work on the same dataset used by the above researcher used two modeling techniques voting and deep neural network which gives the accuracy of 97% and 99%. However, Voting performed well in comparison to deep neural networks and in this study, it was also shown that feature extraction helps in better performance of the model. In, a new model was introduced in this study called the gene sequence model which will not be applied in our paper. Although, the data they used were both public and private. The private dataset they used was from a hospital in China which was obtained from the questionnaire survey from a population of 472 and the public dataset was from a Universitario de Caracas Hospital in Venezuela. In one of the study [19], random forest is used and its efficacy in various genomic and proteomics datasets is described. However, the random forest was unable to classify related traits based on this dataset. Additionally, another research was taking place in Makassar, Indonesia used the data from 38 patients of cancer sufferers and applied the random forest to the sample data they obtained but the problem was they did not get good accuracy, it was just 50% [20]. Researchers Bandyopadhyay and Nasipuri in their study [3] did the work on the preprocessed images and did the k-clustering and, then the Herlev analysis.

Cervical health assessment using machine learning for prediction CHAMP tool

The proposed tool is designed and developed with focus on cervical cancer datasets for prognosis and diagnosis of the disease. The proposed framework is designed on 3 phase layer structure where first phase corresponds to pre-processing to remove missing values with correlated patterns. The second phase conceptualizes on prediction module where the feature selection and best fit machine learning algorithm were enabled to measure the accuracy among the classifiers and in the final phase the visualization interface has been designed for cervical cancer databases to perpetual discover the prediction of the disease.

Principle of CHAMP

The underlying principle of this tool is to develop an advanced and user-friendly platform dedicated to cervical cancer prediction which is designed around a three-layer framework in which each layer serving a specific purpose. The initial layer, known as the Data pre-processing layer, focuses on preparing the data for analysis. This involves careful selection of a comprehensive dataset that encompasses relevant cervical cancer risk factors. Key variables and features known to impact the risk of cervical cancer are identified through data extraction. The data is then preprocessed, addressing missing values, encoding categorical variables, and normalizing numerical data. Once the pre-processing phase complete, the dataset is divided into training and testing sets for subsequent model evaluation. Further, Prediction layer comprises of classification models which are employed to predict the risk of cervical cancer using the preprocessed data. Various classifiers, including machine learning algorithms such as decision trees, support vector machines (SVM), K-nearest neighbors (KNN), AdaBoost, XGBoost, and Naive Bayes, have demonstrated their utility in predicting the risk of cervical cancer as these techniques have been found to be effective in accurately predicting and diagnosing cervical cancer, making them valuable tools in cancer-related applications [21–23]. Then the performance of each model is assessed using appropriate evaluation metrics, including accuracy, precision, and recall and selection of best performing model is selected as the final predictive model for cervical cancer risk assessment. The final layer, the Visualization layer, focuses on delivering an interactive and user-friendly interface through a dashboard layout. Moreover, to enhance the user experience widgets are incorporated and call-back functions are defined to update the dashboard in response to user inputs or the results obtained from the prediction module. Widgets are implemented to allow users to select specific variables or parameters for prediction. The predictions generated by the model are displayed through text which provide prediction to user according to user input. Moreover, the conceptualization of a three-layer framework for data pre-processing, prediction, and visualization was part of the CHAMP' send-to-end development process. The first layer of the model works with databases where each factors were identified in correspondence with risk of cervical cancer for determining the relevant data and then features were selected. Subsequently, in the second layer which is the Prediction layer, many categorization models were trained and assessed for identifying the best model to be applied for the particular database. Lastly, the layer would entitle the users to enter parameters and receive risk projections using the best model, which was integrated into an interactive dashboard in the Visualization layer. However, before deployment, accuracy and user-friendliness were verified by iterative testing and improvement. This tool aims to provide an advanced and intuitive platform for cervical cancer prediction through its three-layer framework, it enables efficient data pre-processing, accurate prediction using machine learning models, and a user-friendly visualization interface. By leveraging these capabilities, the tool empowers users to make informed decisions and gain valuable insights into cervical cancer risk assessment.

Prototype of CHAMP

CHAMP has been designed and developed on systematic approach of accessibility among data with machine learning models integrated with a visualization interface web application using Flask. The prototype involved importing relevant libraries and loading the dataset, followed by comprehensive data cleaning and pre-processing to ensure optimal data quality for model training. The databases were integrated with nominal and categorical variables, where the categorical data was encoded using techniques such as one-hot encoding or label encoding, while numerical variables were scaled to achieve uniformity across features. Further, data was classified using feature extraction and selection methods which were utilized to explore and identify the most informative features for model training. The preprocessed data was further split into training and testing sets to evaluate the model's performance. Various machine learning algorithms were considered and trained using the training data [24]. Predictions were made on the testing data, and performance evaluation was conducted using metrics such as accuracy, precision, and recall.

Further, if the CHAMP performance fell below a specified threshold, hyperparameter tuning, algorithm experimentation, and ensemble methods were employed for refinement of the same. Also, the CHAMP model efficacy is evaluated with trained data using confusion matrix, precision, recall and ROC curve. If the model's performance falls below the specific threshold, then the model ensemble methods were employed for refinement. Also, as the performance lay backs constantly the model is unable to retrieve the effective and efficient patterns with similar time complexity. Moreover, it will lately not serve any purpose. Otherwise, the current model was retained. A Flask application was initialized to create a web interface, define routes, and implement view functions for data rendering [25–27].

Moreover, dynamic data rendering was achieved using Flask's templating engine, and form submissions were handled, including data retrieval and processing, error handling, and user prompts for correction [28–31]. The application was tested extensively to ensure functionality and address other encountered issues for prognosis and diagnosis of disease. Finally, a Visualization-based interface was designed and implemented, incorporating interactive elements and a user-based experience. The user experience inputs were implemented through call-backs to update the interface designed, based on model results, with the option to display selected

Pseudocode for CHAMP
<p>Input the cervical cancer datasets retrieval from the UCI Repository</p> <ol style="list-style-type: none"> 1. Initialize various hyperparameters 2. pre-processing steps for a dataset 3. For the dataset do 4. Extract features 5. end for 6. feature selection 7. EDA ← Shape, Column, summary table, Descriptive analysis 8. For each set of features 9. Combine 10. //ML Model selection and turning steps 11. Models ← are chosen best (XGBoost, AdaBoost, DT, KNN, NB, SVM) for chosen appropriate features 12. Results ← apply the model to the cervical cancer dataset 13. Create a dashboard 14. Run the dashboard application 15. Returns (message, results)

Fig. 1. Pseudocode to create CHAMP.

visualizations or default ones. Moreover, the best-performing machine learning model was integrated into the Flask web application by serializing the model and incorporating it within the application's backend logic. However, data was preprocessed to ensure compatibility when users provided input using a web form, mimicking the procedures used during model training. Then, the serialized model was then used to make predictions using the preprocessed data, and the outcome was published on the website. Furthermore, through a user-friendly and interactive online interface, this integrated approach allowed for smooth user involvement while delivering forecasts and insights. A structured methodology for end-to-end development, encompassing data pre-processing, model training, evaluation, and web application creation, resulting in a user-friendly interface for data exploration and analysis demonstrated in Fig. 1.

Architecture of CHAMP

The architecture of the tool involves several layers that work together to enable the prediction of cervical cancer and the visualization of the results. The framework consists of three layers, each serving a specific purpose. However, the first layer is the data pre-processing layer, which comprises the focus on refining the dataset by addressing missing values and inconsistencies which aims to identify relevant patterns within the data that can contribute to accurate predictions. The second layer, is termed as the prediction layer which emphasizes on selecting the most suitable classifier model that can accurately predict the risk of cervical cancer. Various classification models, such as Decision tree, Naïve Bayes, AdaBoost, XGBoost, and KNN are applied to the preprocessed data and the performance of each classifier is evaluated using appropriate metrics to determine their effectiveness. Optimization techniques are utilized to evaluate the best-performing classifiers. The third layer focuses on visualization, providing an interactive user interface for effective exploration and interpretation of the results. The components within this layer are designed to facilitate user interaction and provide a seamless experience, providing user access to different attributes and classifiers through user inputs, enabling them to explore various scenarios and evaluate the predictions [32,33]. The visualization layer aims to enhance the usability and accessibility of the tool, enabling users to gain valuable insights from the generated predictions. By combining these three layers, the proposed architecture aims to provide an effective and user-friendly tool for predicting the risk of cervical cancer. The tool leverages the pre-processing layer to refine the data, the data classifier layer to identify the best-performing models, and the visualization layer to present the results intuitively and interactively. This architecture enables healthcare professionals and researchers to make informed decisions based on the predictions and gain a better understanding of cervical cancer risks. The structured architecture is shown in Fig. 2.

Moreover, the designed webpage consists of two interconnected pages. Firstly, the home page acts as an entry point to the application, offering information about the best-performing model and the attributes used. Although, this page not only summarizes the attributes employed in the tool but also highlights the best model based on evaluation metrics and it presents a clear overview of each attribute's significance and their role in the prediction process. Although, the "Predictor" button featured on the home page serves as the pivotal link that seamlessly connects users to both the analytical and interactive realms of our platform. As this button acts as a dynamic gateway, allowing users to input their data interact with the predictive algorithms and offer transition from the home page to the prediction interface, users are empowered to harness the predictive capabilities of our platform.

Moving to the second page, known as the prediction page, users can input their information to obtain a prediction result for cervical cancer risk and this page features attribute text boxes where users enter details such as age and number of sexual partners. Subsequently, by clicking the "Predict" button, the prediction process is initiated. The trained prediction model, powered by advanced machine learning algorithms, analyzes the input data and generates a prediction outcome.

However, the prediction outcome is presented to users as either 0 or 1. A prediction of 0 signifies a low risk of cervical cancer, while a prediction of 1 indicates a higher risk. This information empowers users to gain insights into their potential risks and make proactive decisions about their health.

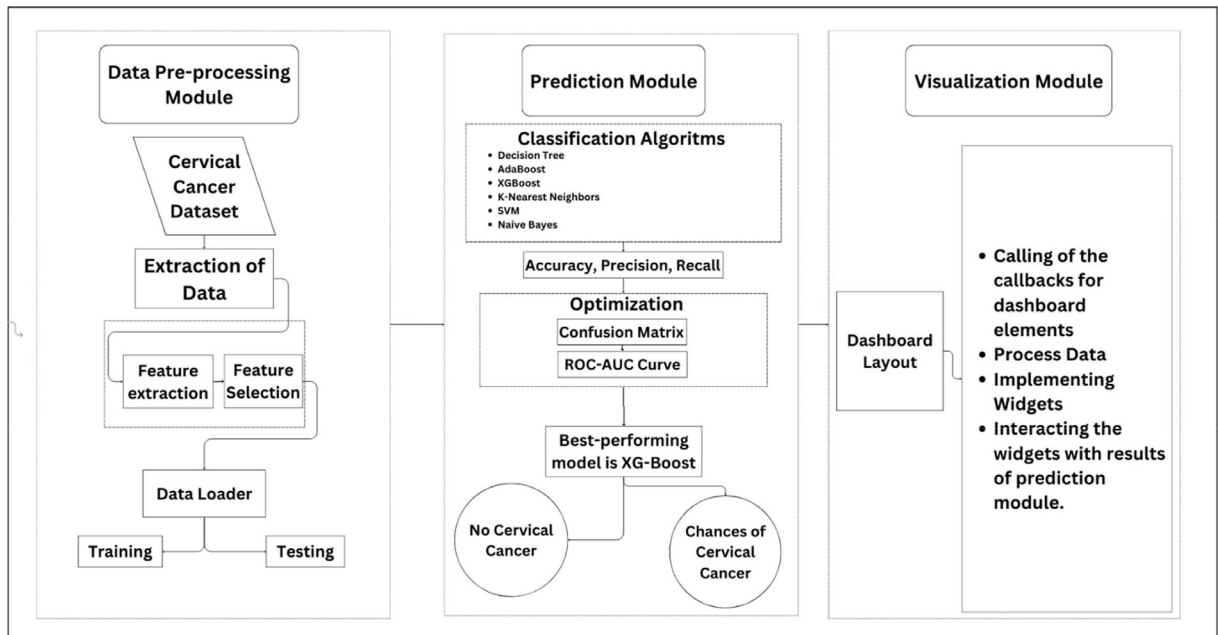


Fig. 2. Architecture of CHAMP.

CHAMP overall in layers

The proposed tool for the diagnosis of cervical cancer is structured into three layers: data pre-processing, prediction, and visualization. Each layer serves a specific purpose and incorporates machine learning algorithms. The data pre-processing layer handles tasks such as dataset selection, data extraction, and data cleaning to ensure the dataset is comprehensive and relevant. The prediction layer utilizes machine learning algorithms to analyze the pre-processed data and make accurate predictions. Finally, the visualization layer provides an interactive interface for users to explore and interpret the results effectively.

Data pre-processing layer

The data-pre-processing layer in the proposed tool for cervical cancer prediction involves various steps such as dataset selection, data extraction, and data cleaning. It focuses on preparing the selected comprehensive and relevant dataset by handling missing values, encoding categorical variables, and normalizing numerical data. The preprocessed dataset is then split into training and testing sets for evaluation and further analysis.

However, the data pre-processing layer plays a pivotal role in refining the raw dataset before it enters the prediction phase as it encompasses several crucial steps, including handling missing data points, removing inconsistencies, and addressing outliers. Moreover, by ensuring data quality and integrity, this layer sets the foundation for accurate predictions and it also involves normalization and scaling to bring all data attributes to a similar range, preventing any particular attribute from dominating the analysis due to its magnitude. Furthermore, feature selection techniques might be employed to identify the most relevant attributes that contribute significantly to the prediction process. However, this layer's focus on data enhancement and preparation ensures that the subsequent prediction models receive high-quality input, thereby increasing their reliability.

Cervical cancer dataset. The cervical dataset utilized in this study serves as a foundational element for our research focused on cervical cancer prediction and diagnosis. The data is derived from real-world clinical data collected at the Hospital Universitario de Caracas, situated in Caracas, Venezuela comprising a comprehensive range of variables and attributes, the dataset is specifically curated to capture relevant risk factors and characteristics associated with cervical cancer [34].

The dataset's structure follows a meticulous approach to data collection, ensuring the inclusion of pertinent features that hold the potential for predicting cervical cancer outcomes. Moreover, the proposed study approach has utilized the cervical cancer databases which are comprised of 36 features and 858 observations. The data quantifies the varied features which tend to be accountable for the prognosis of the disease. The disease dataset focuses on the risk of cervical cancer in the population studied. The datasets capture a wide range of clinical and demographic factors that are associated with the development and progression of cervical cancer in women. Variables such as age, HPV test results, cytology outcomes, and various clinical parameters are meticulously recorded, contributing to the dataset's richness and complexity. This comprehensive scope enables us to encompass diverse aspects of cervical cancer risk assessment.

Moreover, these factors are crucial in understanding the complex nature of the disease and identifying potential risk factors that can contribute to its occurrence. Further, datasets encompass various attributes which were extensively studied in correspondence to cervical cancer.

However, the dataset's diversity is essential in enabling our comparative study of machine learning algorithms, we employ a range of techniques, including decision trees, support vector machines, Naive Bayes, AdaBoost, and K-Nearest Neighbors, to predict cervical cancer risks accurately. Each algorithm undergoes rigorous evaluation, considering metrics such as accuracy, precision, and recall, to identify the best-performing model.

Data pre-processing. Data pre-processing involves data cleaning, feature extraction, and feature selection. Whereas, data cleaning has been done for identifying and correcting or removing any errors or inconsistencies in the data, such as missing values, duplicate data, and outliers. Firstly, missing data are handled using methods such as imputation, which replaces missing values with statistically derived estimates. Next, inconsistencies in the dataset are addressed to rectify contradictory values within attributes. Furthermore, outliers, data points significantly deviating from the norm, are detected and managed through statistical techniques and to create a uniform analysis platform, normalization and scaling techniques are applied, preventing attribute magnitude from skewing results.

Furthermore, to handle categorical and numerical data during the model training process, a technique called one-hot encoding is applied. This method transforms categorical variables into binary vectors, with each category represented as a unique binary attribute and this ensures that the machine learning algorithms can interpret categorical data appropriately, without imposing any ordinal relationships among categories. On the other hand, numerical data is subjected to normalization and scaling. Normalization transforms numerical attributes to a common scale, often between 0 and 1, to prevent attributes with larger values from disproportionately influencing the models. Furthermore, scaling standardizes attribute magnitudes to similar ranges, enhancing the algorithms' convergence and efficiency, and by treating both categorical and numerical data with these tailored techniques, CHAMP ensures that all relevant information is integrated into the predictive models without introducing bias or inaccuracies due to data type variations. So, this approach contributes to the accuracy and effectiveness of the machine learning models in predicting cervical cancer risks.

In order to transform raw data into a set of features that are more meaningful and informative for the machine learning models we use feature extraction which is an important step in machine learning because the quality and relevance of the features used to train a model can significantly impact its performance and accuracy. Further, this technique has been used to make a new column named 'total_tests' which is for the Number of tests done for diagnosis. Additionally, one of the important processes, which entails selecting a subset of pertinent characteristics from a broader set in order to increase the precision and effectiveness of a model includes feature selection. The goal is to reduce the dimensionality of the dataset by removing redundant, irrelevant, or noisy features while retaining the most informative ones. The techniques can be broadly categorized into three categories: filter methods, wrapper methods, and embedded methods [35]. Filter methods use statistical measures to rank features based on their relevance to the target variable whereas wrapper approaches assess the performance of various feature subsets using a machine learning algorithm. Lastly, the embedded methods incorporate feature selection as part of the model-building process, where the feature selection and model training are performed simultaneously [18–20,36–47].

However, the size and complexity of the dataset, the kind of machine learning algorithm being employed, and the research issue under consideration all influence the selection of the best feature selection technique. So, the effectiveness of the feature selection approach and the resulting model may be assessed using several performance metrics, including accuracy, precision, recall, and F1 score. For feature selection, we employ the Chi-square technique, which is a filter-based feature selection technique that works best when the target variable and the majority of the characteristics are both categorical. Utilizing the chi-square test, categorical features are ranked according to how well they are associated with the target variable [35]. The more important a feature is to the objective variable the higher the chi-square score, and the more likely it is to be chosen for further investigation. Moreover, it is generally faster and less computationally expensive in comparison with wrapper-based or embedded-based methods. However, they do not consider the interaction between features and may not perform as well as wrapper-based or embedded-based methods in complex datasets. After performing this we get to know eight columns have higher chi scores and other columns were eliminated to get the best result out of the models. The selected eight columns and target variables are explained in Table 1.

Prediction layer

The prediction layer of the proposed framework incorporates various machine learning algorithms, including decision trees, support vector machines (SVM), K-nearest neighbors (KNN), AdaBoost, XGBoost, and Naive Bayes. Moreover, these algorithms are utilized to predict the risk of cervical cancer based on preprocessed data, enabling accurate and reliable predictions for diagnosis and prognosis purposes.

However, the prediction layer is the heart of the architecture, where advanced machine learning algorithms are employed to analyze the pre-processed data and generate predictions. Each selected classifier, such as Decision Tree, Naive Bayes, AdaBoost, XGBoost, and K-Nearest Neighbors, comes with its own strengths and limitations. Moreover, the focus here is on implementing and fine-tuning these models to achieve optimal performance. Furthermore, hyperparameter tuning, cross-validation, and ensemble methods might be utilized to ensure the classifiers generalize well to unseen data and mitigate overfitting. However, the prediction layer's ultimate objective is to produce accurate predictions of cervical cancer risk based on the input data.

Algorithms. **Decision Tree Classifier** A supervised machine learning approach used for both classification and regression tasks is a decision tree classifier which is used to develop a model that can be used for prediction in which the dataset is recursively divided into smaller subsets based on the values of one of the characteristics [48].

Table 1
Selected eight columns and target variable column.

Age	This refers to the age of the individual. Age is an important risk factor for cervical cancer, as both young age (due to increased exposure to risk factors) and older age (due to changes in hormonal balance and cellular function) can contribute to the risk.
Smokes	This feature likely indicates whether the individual is a smoker or not. Smoking tobacco can increase the risk of cervical cancer due to its harmful effects on the immune system and cervical cells.
STDs	This could represent the presence or absence of any sexually transmitted diseases (STDs) in the individual. Certain STDs, particularly human papillomavirus (HPV), are the primary cause of cervical cancer.
STDs: pelvic inflammatory disease	This indicates the presence or absence of pelvic inflammatory disease (PID) as a complication of STDs. PID can lead to inflammation of the reproductive organs and potentially increase the risk of cervical cancer.
STDs:AIDS:	This likely indicates the presence or absence of AIDS as a complication of STDs. Having a compromised immune system due to AIDS can elevate the risk of cervical cancer.
STDs:Hepatitis B	This indicates the presence or absence of hepatitis B as a complication of STDs. Hepatitis B can potentially contribute to an increased risk of cervical cancer.
STDs:Number of Diagnosis	This could represent the number of STD diagnoses an individual has had. Multiple STD diagnoses might indicate higher exposure to risk factors.
total_tests	This feature could refer to the total number of tests conducted. It's unclear what specific tests these refer to, but they could be relevant in assessing overall health and risk.
Dx:Cancer	This likely indicates whether the individual has been diagnosed with cancer. It's a critical feature for prediction since individuals with a cancer diagnosis would have a significantly higher risk of cervical cancer.

Decision trees make the resulting model simple to see and comprehend which is one of their key features. Additionally, decision trees can handle missing values as well as categorical and numerical features [49]. They are also useful in ensemble approaches like Random Forests and Gradient Boosting and perform well on huge datasets.

K-Nearest Neighbors (KNN) KNN is a non-parametric algorithm that does not assume anything about the distribution of the data at its core. However, a new data point is classified using KNN by looking at its k nearest neighbors in the training set and being assigned to the class that is most prevalent among them [36]. Further, the user selects the value of k, which establishes how many neighbors should be taken into account when producing a prediction. KNN can be used for regression problems by taking the average or median of the k nearest neighbors, as well as for binary and multi-class classification tasks which makes it beneficial for being simple to use and requiring no training time. However, when working with sizable datasets, it might be computationally expensive and is regarded as a lazy learning algorithm because it does not create a model from the training set of data [37]. Instead, it only maintains the training data and generates predictions based on how far the new data point is from the training set's previous data points [38].

Adaptive Boosting (AdaBoost) AdaBoost is to iteratively train a sequence of weak learners, each focusing on the examples that the previous models misclassified, and then combine the weak learners into a final strong model. In each iteration, the algorithm assigns higher weights to the misclassified examples so that the next model in the sequence focuses on these difficult examples [18]. The final model is then a weighted combination of the weak learners, with each weak learner's weight determined by their performance on the training set. AdaBoost is particularly useful when the base learners are simple and weak, such as decision stumps (simple decision trees with only one split). Although, AdaBoost can also handle datasets with a large number of features, as it can identify the most relevant features for the task at hand [39]. Additionally, in AdaBoost, each weak learner is assigned a weight based on their performance in classifying the training data. The weight of each instance in the training set is adjusted based on whether it was correctly or incorrectly classified by the weak learner. In Eq. (1), $F(x)$ is the final prediction of the AdaBoost model for the input feature vector x , α_t is the weight or coefficient assigned to the t -th weak classifier $f_t(x)$, $f_t(x)$ represents the t -th weak classifier's prediction for the input feature vector x and T is the total number of weak classifiers in the ensemble.

$$F(x) = \sum_t^T \alpha_t f_t(x) \quad (1)$$

XGBoost (extreme Gradient Boosting) XGBoost is a state-of-the-art machine learning algorithm known for its exceptional performance and versatility which belongs to the family of gradient-boosting algorithms and is widely used in various domains for classification, regression, and ranking tasks, optimizes the loss function using gradient descent. It computes gradients and adjusts the model's parameters in an iterative manner to minimize the loss and improve predictions which enables the algorithm to handle complex patterns and capture subtle relationships in the data. It provides several regularization techniques to control model complexity and prevent overfitting. It includes methods such as shrinkage (learning rate) and L1 and L2 regularization [40]. These techniques help in reducing variance, improving generalization, and enhancing the model's ability to generalize well to new, unseen data. Although, it has built-in capabilities to handle missing values in the dataset which can automatically learn the optimal direction to assign missing values during the tree construction process, eliminating the need for manual imputation. It also provides a feature importance metric that ranks the importance of each feature in the dataset. This information helps in understanding the relative contribution of features to the model's predictions, identifying key factors driving the outcomes, and performing feature selection and is highly optimized for efficiency and speed. And it also supports parallel processing and can leverage multiple cores on a single machine or distributed computing frameworks to train models faster, making it suitable for large-scale datasets [41]. Moreover, it offers a wide range of hyperparameters that can be tuned to optimize model performance which has been widely adopted in both research and industry, winning numerous machine-learning competitions and demonstrating its effectiveness in real-world scenarios.

In Eq. (2), \hat{y}^i is the predicted output for the i th sample, $f_k(x_i)$ represents the k -th base learner’s prediction for the i th sample, K is the total number of base learners (individual decision trees) in the ensemble model and x_i denotes the feature vector for the i th sample.

$$\hat{y}^i = \sum_{k=1}^K f_k(x_i) \tag{2}$$

Naïve Bayes Naive Bayes is a popular and widely used machine learning algorithm known for its simplicity and effectiveness in classification tasks which is particularly suited for text classification and spam filtering applications. The algorithm is referred to as “naive” because it makes a strong assumption of independence among the features in the data, meaning that each feature contributes independently to the final classification decision. Although, this assumption simplifies the calculations and makes the algorithm computationally efficient, even with large datasets [42]. The algorithm uses conditional probabilities and probability density functions to make these calculations [14]. And it’s one of the key advantages is its simplicity and fast training and prediction times which requires a relatively small amount of training data compared to other algorithms and can handle high-dimensional data efficiently. Moreover, Naive Bayes performs well even when the independence assumption is violated to some extent and other advantage of this algorithm is its interpretability [15]. Additionally, this algorithm provides clear probabilities and conditional probabilities for each class, allowing for a better understanding of the underlying patterns and relationships in the data. However, it may not perform optimally when the independence assumption is strongly violated or when there are complex dependencies among the features. It may also struggle with rare events or imbalanced datasets where one class dominates the others. Furthermore, Naive Bayes assumes that features are conditionally independent, which may not always hold true in real-world scenarios. Although despite these limitations, Naive Bayes remains a valuable algorithm in various domains, including text classification, spam filtering, sentiment analysis, and recommendation systems. Its simplicity, efficiency, and interpretability make it a popular choice for many classification tasks, especially when dealing with large volumes of textual data. The equation for the Naive Bayes classifier is shown in Eq. (3). The abbreviation used is $P(y|x_1, x_2, \dots, x_n)$ is the posterior probability of class y given the features x_1, x_2, \dots, x_n , $P(y)$ is the prior probability of class y , $P(x_1|y)$, $P(x_2|y), \dots, P(x_n|y)$ are the conditional probabilities of feature x_i given class y , $P(x_1, x_2, \dots, x_n)$ is the probability of observing the features x_1, x_2, \dots, x_n .

$$P(y|x_1, x_2, \dots, x_n) = P(y) * P(x_1|y) * P(x_2|y) * \dots * P(x_n|y) / P(x_1, x_2, \dots, x_n) \tag{3}$$

Support Vector Machines (SVM) Support Vector Machines (SVM) are a potent and adaptable supervised learning technique which enhances generalization and increases its capacity to correctly categorize brand-new, untainted data. Moreover, by transforming the input features into a higher-dimensional space using kernel functions, SVM can handle both linearly separable and non-linearly separable data [16]. As a result, SVM can recognize subtle patterns in the data and locate complex decision boundaries provides a number of benefits, including the capacity to handle high-dimensional data, robustness to outliers, and efficacy even with little training data. It has been extensively used in a variety of fields, including bioinformatics, text classification, and image classification. However, the performance of SVM could depend on the hyperparameters selected and its capacity to scale to big datasets. Nevertheless, SVM is still a popular option for many machine learning tasks when properly tuned and taken into account. The equation shown in Eq. (4) used abbreviations is $f(x)$ represents the predicted class label for the input vector x , w is the weight vector, T denotes the transpose operation, b is the bias term.

$$f(x) = \text{sign}(w^T * x + b) \tag{4}$$

Six algorithms—AdaBoost, KNN, Naive Bayes, SVM, XGBoost, and Decision tree—were used in this experiment. This experiment was compared and validated using the criteria recall, accuracy, precision, and ROC curve.

Accuracy measures. Accuracy measures are the fundamental indicators which represent how well machine learning models perform in reference to the given dataset. The accuracy measures work on the protocol of counting the number with examples in a dataset which were successfully predicted as a percentage of all the instances. In other words, accuracy refers to how well a model can categorize or forecast the desired variable with low and high scores. However, a low accuracy score suggests that the model is producing inaccurate predictions, a high accuracy score suggests that the model is making accurate predictions. Due to this, the accuracy models are widely applied globally for its simplicity and interpretability which offer insights into a straightforward understanding of the specification of model performance [21].

Further, while measuring accuracy, it is vital to keep in mind the context nature of the problem domain. The desired level of accuracy may vary depending on the application. For example, in some domains, such as medical diagnosis, achieving high accuracy is crucial to avoid misdiagnosis and ensure patient safety. On the other hand, in certain scenarios, a slightly lower accuracy may be acceptable if it is accompanied by higher precision or recall for specific classes of interest. In eq. (5) the overall measure of accuracy can be discussed with TP for True Positive and TN for True Negative.

$$\text{Accuracy} = \frac{((TP + TN))}{(\text{Total no of samples})} \tag{5}$$

Further, the precision or correctness of a model’s positive predictions is quantified by exactness making it a crucial evaluation parameter in machine learning. It works on the protocol in which it calculates the percentage of genuine positive predictions from all the projected positive instances. In other words, the precision of positive predictions is the main emphasis of precision, which also sheds light on the model’s capacity to filter out false positives. A high precision score means that the model is producing fewer

false positives and accurate positive predictions. It means that the model has a good chance of being accurate when it predicts a favorable result. Precision is particularly important in situations where false positives can have significant consequences or costs, such as in medical diagnoses or fraud detection. In these cases, ensuring high precision is crucial to minimize false alarms and unnecessary actions [17]. However, it's important to note that precision alone may not provide a complete picture of a model's performance, especially when dealing with imbalanced datasets or when the focus is on identifying all positive instances. In some cases, maximizing precision may lead to a decrease in recall (the proportion of true positive instances identified), resulting in missed positive cases. The balance between precision and recall depends on the specific objectives and requirements of the problem at hand. It is often necessary to strike a trade-off between precision and recall by adjusting the model's threshold or using techniques such as precision-recall curves or F1 score, which combine both precisions and recall into a single metric. It is a crucial metric in evaluating the performance of machine learning models, especially when the focus is on minimizing false positives. However, it should be interpreted alongside other evaluation metrics and consider the trade-off between precision and recall based on the specific needs and context of the problem being addressed. In Eq. (6) the overall measure of accuracy can be discussed with TP for True Positive and FP for False Positive.

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (6)$$

Further, recall is a crucial evaluation parameter in machine learning that assesses a model's accuracy in correctly identifying all positive cases. It measures the percentage of true positive forecasts and is concerned with capturing the pertinent positive events and preventing false negatives from the dataset's actual positive cases, A high recall score indicates that the model is effectively identifying most of the positive instances and minimizing the number of false negatives. It implies that when a positive instance exists in the dataset, the model is likely to detect it. Moreover, recall is particularly crucial in situations where missing positive cases can have significant consequences, such as in medical diagnoses or anomaly detection. In some cases, maximizing recall is vital to ensure that all relevant positive instances are captured and addressed [21], and in others, it is vital to maximize recall may lead to an increase in false positives, which can result in unnecessary actions or alarms. The trade-off between recall and other metrics, such as precision, depends on the specific goals and requirements of the problem. However, achieving a balance between recall and precision is often necessary to optimize the model's performance. Techniques like precision-recall curves or F1 scores can provide a more comprehensive evaluation by considering both precision and recall simultaneously. In Eq. (6) the overall measure of accuracy can be discussed with TP for True Positive and FN for False Negative.

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (7)$$

Roc curve is an exactness which is a key evaluation criterion in machine learning, quantifies the accuracy or correctness of a model's positive predictions which determines the percentage of accurate positive predictions out of all the anticipated positive cases. In other words, the emphasis on precision is on making accurate positive predictions, which also reveals how well the model can weed out false positives. When the model predicts a positive outcome, it suggests that the model has a good likelihood of being accurate [22]. The ROC curve provide valuable insights into the model's discriminative power, robustness, and overall performance. Moreover, a model with a higher score is considered better at distinguishing between positive and negative instances and used as a summary statistic to compare different models or to determine an optimal threshold for classification. It is important to note that the ROC curve for effective evaluation metrics, particularly in scenarios where the cost of false positives and false negatives may vary [23]. Although, they are commonly used in medical diagnostics, fraud detection, and other applications where the balance between sensitivity and specificity is crucial. By analyzing the ROC curve and AUC, practitioners can make informed decisions about model performance and select the appropriate threshold for their specific needs [50].

Model building. Model building comprises of step-by-step approach which first starts with the loading of the Scikit-learn machine learning library that features numerous ML algorithms for the Python programming language. Further, to make the data comprehensible for ML to read, data need to be converted to numerical values of 1 s and 0 s. Moreover, in the data 'Cancer' was the dependent feature and the rest were independent features which were separated and divided into training and testing subsets where the distribution ratio was 70:30 and 80:20, respectively. After splitting the data in ratio, the programs for transferring the data and using the algorithm on the data were used to fit the data into each algorithm model. Following that, only training data was used to train the models, and the remaining data, often known as the testing data, are now presented to the model. The prediction of the model is evaluated on this subset of the data which allows us to evaluate the predictive accuracy by running algorithms and recording the accuracy findings. The accuracy results were noted for further research.

Visualization layer

The visualization layer of the proposed tool using Flask plays an important role in presenting and interpreting the data and results of cervical cancer prediction models which leverages various visualization techniques and interactive features to provide a comprehensive and user-friendly interface for healthcare professionals [51,52]. Moreover, the visualization layer presents the prediction results generated by the machine learning models and it displays the predicted probabilities. Although, proposed tool CHAMP incorporates a user-input component for predicting the likelihood of cervical cancer which includes eight attributes that allow users to input specific information related to a patient's health and medical history. These attributes, such as age, and smoking status, serve as input fields where users can enter the corresponding values. Once the user has provided the necessary information, they can click the "Predict" button to initiate the prediction process which is then passed to the underlying machine learning models, which

Table 2
All columns in dataset and their datatype.

Column Name	Datatype
Age	Integer
Number of sexual partners	Object
First sexual intercourse	Object
Num of pregnancies	Object
Smokes	Object
Smokes(years)	Object
Smokes(packs/year)	Object
Hormonal Contraceptive	Object
Hormonal Contraceptive (years)	Object
IUD	Object
IUD (years)	Object
STDs	Object
STDs (number)	Object
STDs: condylomatosis	Object
STDs: cervical condylomatosis	Object
STDs: vaginal condylomatosis	Object
STDs: vulva-perineal condylomatosis	Object
STDs: syphilis	Object
STDs: pelvic inflammatory disease	Object
STDs: genital herpes	Object
STDs: molluscum contagiosum	Object
STDs: AIDS	Object
STDs: HIV	Object
STDs: Hepatitis B	Object
STDs: HPV	Object
STDs: Number of diagnoses	Object
STDs: Time since first diagnosis	Object
STDs: Time since last diagnosis	Object
Dx: Cancer	Integer
Dx: CIN	Integer
Dx: HPV	Integer
Dx	Integer
Hinselmann	Integer
Schiller	Integer
Citology	Integer
Biopsy	Integer

have been trained on the cervical cancer dataset. Further, the models analyze the input attributes and generate a prediction output indicating the probability or likelihood of cervical cancer for the given input which is then displayed to the user, providing valuable insights into the potential risk of cervical cancer based on the provided attributes. Although, this functionality empowers healthcare professionals to quickly assess the risk of cervical cancer for individual patients, aiding in early detection and intervention. Moreover, by integrating the user-input component with the prediction process, the visualization module of the tool using Flask allows for a seamless and user-friendly experience which enables healthcare professionals to leverage the power of machine learning algorithms and predictive models to make informed decisions and provide personalized care to their patients. Although, based on the information provided by the user, the model then generates a prediction outcome. Specifically, if the resulting prediction is 0, it signifies that the model has assessed a low risk of cervical cancer associated with the provided data. Conversely, if the prediction outcome is 1, this indicates a higher risk of cervical cancer as evaluated by the model. Through this process, users gain valuable insights into their potential risk level, enabling them to make informed decisions and take proactive steps toward managing their health and well-being.

However, the visualization layer aims to make the prediction results accessible and interpretable for healthcare professionals and researchers and it offers a user-friendly interface where users can interact with the models and their outcomes. Moreover, this layer may utilize graphical representations, such as Pie charts, bar graphs and confusion matrix, to present the predictions and associated metrics visually. Furthermore, the user can input various scenarios and parameters to observe how different classifiers perform under different conditions. Additionally, this layer may allow users to compare the predictions, facilitating informed decision-making, the emphasis here is on creating an intuitive and engaging platform that enhances the user's understanding of the prediction results and encourages exploration.

Results

In the current, scope of the study, we have explored the basic statistics and characteristics among the data with relevant features. Each feature was statistically calculated to other features to determine the dependency of features within the data. The statistical measures which were calculated to measure the correlation were summary statistics such as mean, median, standard deviation, and quartiles for numerical variables like Age, Number of sexual partners, First sexual intercourse, Num of pregnancies, Smokes

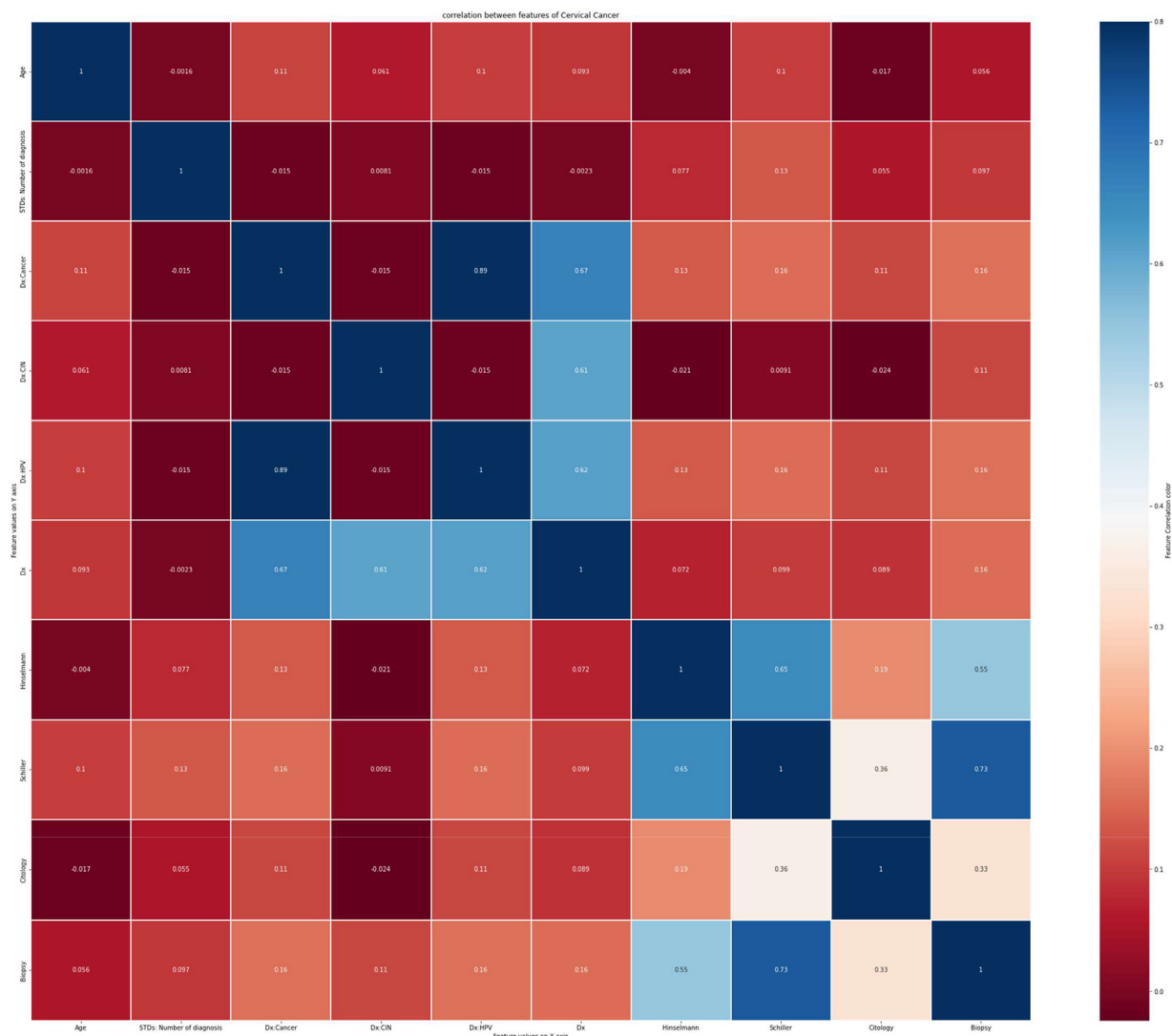


Fig. 3. Correlation between features of the dataset.

(years), Smokes (packs/year), Hormonal Contraceptives (years), IUD (years), STDs: Time since first diagnosis, STDs: Time since last diagnosis and also computing the mode and frequency distribution for categorical variables like Smokes, Hormonal Contraceptives, IUD, STDs, STDs (number) STDs:condylomatosis, STDs: cervical condylomatosis, STDs: vaginal condylomatosis, STDs: vulva-perineal condylomatosis, STDs: syphilis, STDs: pelvic inflammatory disease, STDs: genital herpes, STDs: molluscum contagiosum, STDs: AIDS, STDs: HIV, STDs: Hepatitis B, STDs: HPV, STDs: Number of diagnoses, Dx: Cancer, Dx: CIN, Dx: HPV, Dx Hinselmann, Schiller, Cytology, Biopsy. All the columns and their datatype are represented in Table 2.

Furthermore, the correlation matrix has been utilized to measure the dependency of features with the corresponding features. The results signify the multicollinearity among the features with the focus on identifying patterns that might be strongly correlated with the target variable mentioned in Fig. 3.

To determine the correlation between different features in the data, the statistical measure employed is the Chi-square technique. This technique is particularly effective when both the target variable and the majority of the characteristics are categorical. The Chi-square test was utilized to rank categorical features based on their association with the target variable. Features that are more crucial to the target variable received higher Chi-square scores and were more likely to be selected for further investigation. Moreover, the Chi-square method is generally known for its efficiency, as it is faster and less computationally intensive when compared to other statistical approaches. Also, it is a filter-based method suitable for categorical features and target variables. Moreover, it ranks features based on their association with the target variable, with higher chi-square scores indicating greater important. So, after applying the Chi-square technique, it was determined that eight columns exhibited higher Chi-scores, while other columns were eliminated from consideration. These selected eight columns, along with the target variables with their chi-square value, are detailed in Table 3.

Table 3
Columns with their highest chi-square value.

S.No	Column Name	Chi-square value
1	Age	7.225
2	Smokes	6.110
3	STDs	7..679
4	STDs:pelvic inflammatory disease	7.247
5	STDs:AIDS	7.347
6	STDs:Hepatitis B	7.247
7	STDs: Number of diagnosis	8.522
8	Dx:Cancer	8.739
9	Total_tests	8.729

Proportion of women across age categories with a diagnosis of Cancer, HPV

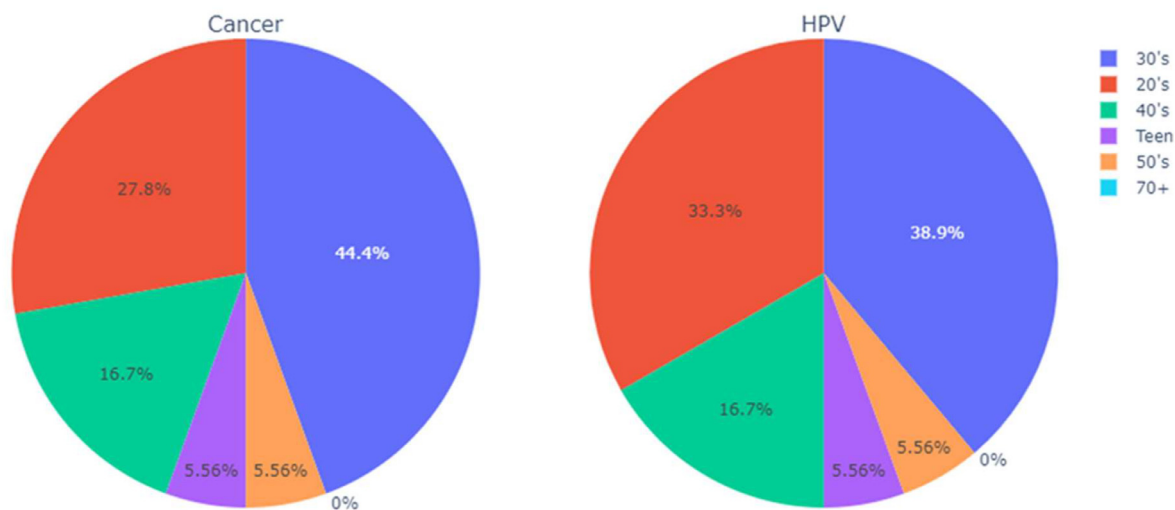


Fig. 4. Proportion of women across age categories with a diagnosis of Cancer, or HPV.

Table 4
K-Fold score for each split of each model.

Classifier	Train-test split	K-fold Score
Ada-Boost	70:30	97.3
Ada-boost	80:20	97.2
K-Nearest Neighbour	70:30	98.3
K-Nearest Neighbour	80:20	97.6
Naïve Bayes	70:30	24.6
Naïve Bayes	80:20	25.8
Support Vector Machine	70:30	97.2
Support Vector Machine	80:20	98.1
XGBoost	70:30	97.5
XGBoost	80:20	98.9
Decision Tree	70:30	97.8
Decision Tree	80:20	96.8

Additionally, data was visualized using varied plots and charts to identify the significant patterns and information from the data. Further, database has been visualized to understand the proportion of women across age categories with a diagnosis of Cancer, or HPV in Fig. 4.

Validation of model

To validate the model performance, prediction accuracy was calculated for each model, along with this receiver operating characteristic (ROC), and the Area under Curve (AUC) score was also determined to evaluate the model's effectiveness.

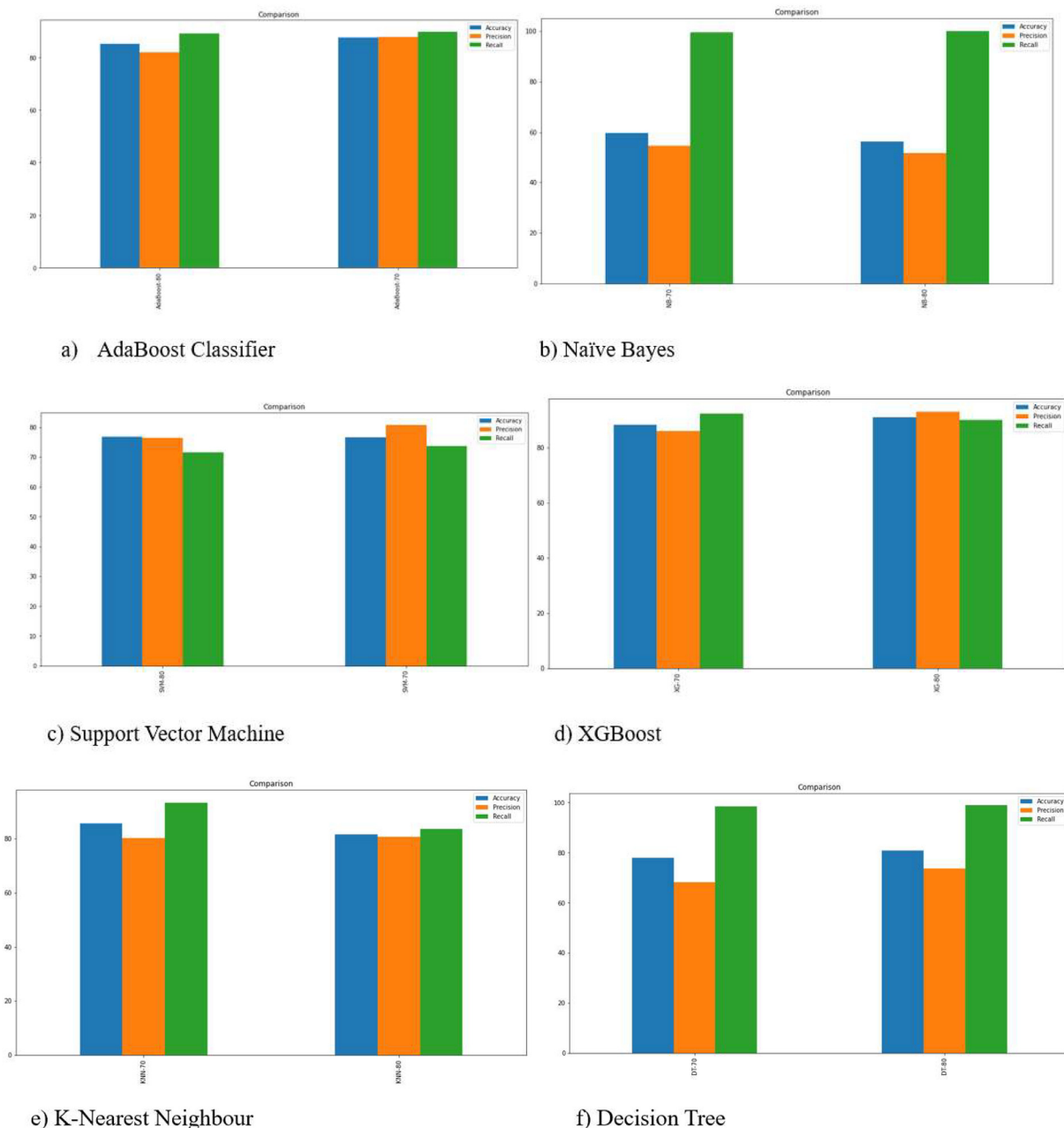


Fig. 5. Comparison of accuracy, precision, and recall between both train-test split of (a) AdaBoost Classifier (b) Naïve Bayes (c) Support vector machine (d) XGBoost Classifier (e) K-Nearest Neighbour (f) Decision tree.

In addition, to ensure the reliability and robustness of the model evaluations, a K-fold cross-validation technique has been employed. Moreover, this technique involves partitioning the dataset into K subsets, training the model on K-1 subsets, and validating it on the remaining subset. This process is iterated K times, with each subset being used as the validation set once, and by averaging the performance metrics across all K iterations, a more accurate estimate of the model’s performance can be obtained.

In our current study of approach, a systematic procedure has been followed to evaluate and compare the performance of different models. Firstly, the databases have been split into two sets of training and test datasets where 80% of the data was synthesized for training whereas 20% was for testing in one case. Further, the data was reconsidered with the following, 70% for training, and 30% for testing in another case. For each model, data has been trained on the training set, and its accuracy, precision, and recall on both the 80:20 and 70:30 splits have been evaluated. Moreover, by comparing the performance metrics between the two splits, the impact

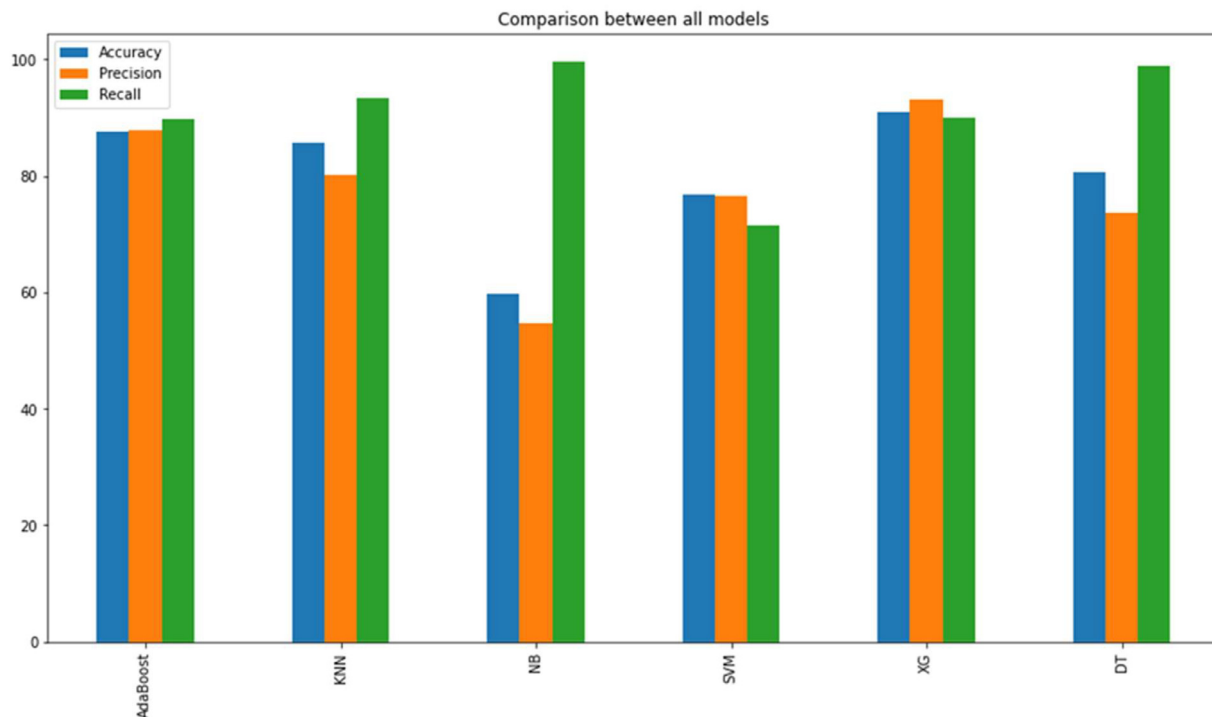


Fig. 6. Comparison of accuracy, precision, and recall between the best test-train split of each model.

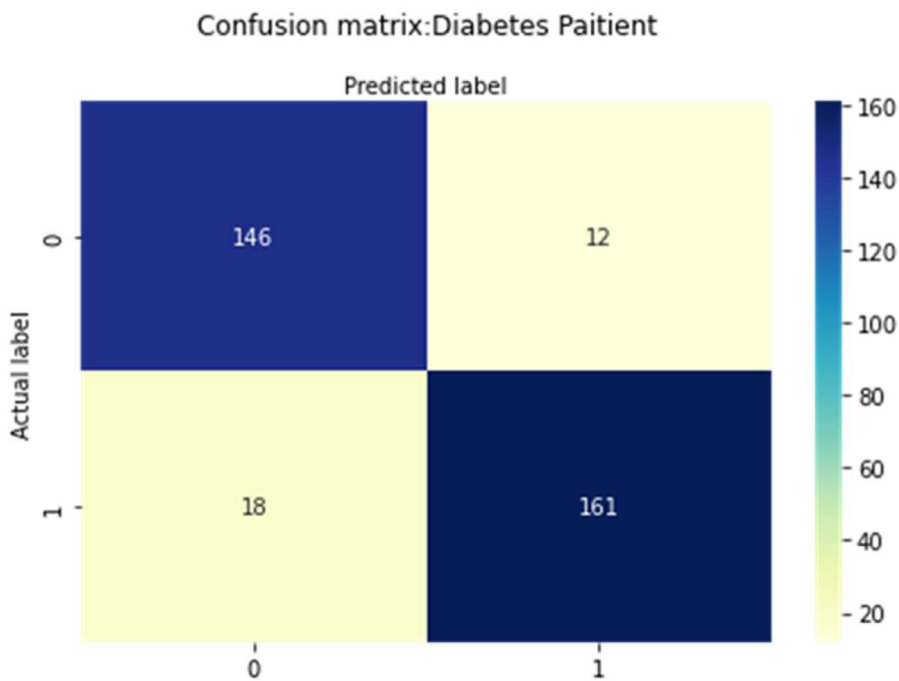
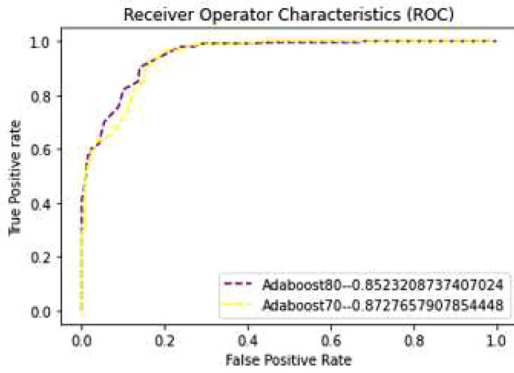


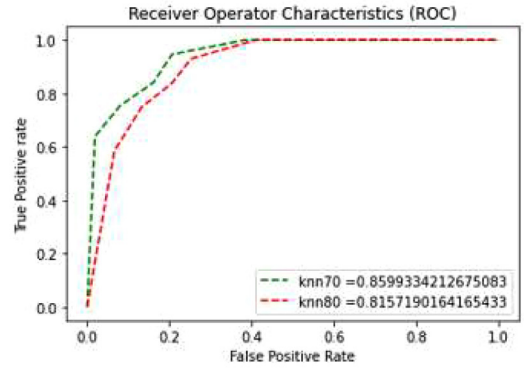
Fig. 7. Confusion matrix of XGBoost classifier.

of the training-test data ratio on the model’s performance has been assessed which allowed us to identify the split that resulted in better accuracy, precision, and recall for each model. A comparison between both train-test splits of each model is shown in Fig. 5.

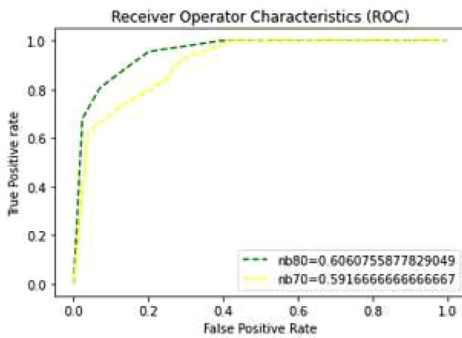
Moreover, for every model under scrutiny, the training data subset was employed to hone its predictive capabilities. Key performance metrics like accuracy, precision, and recall were assessed for both the 80:20 and 70:30 splits, as well as for each iteration



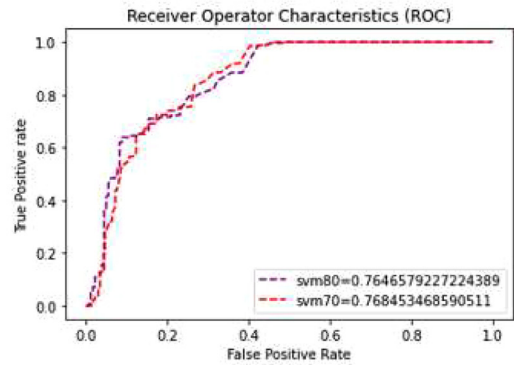
a) AdaBoost Classifier



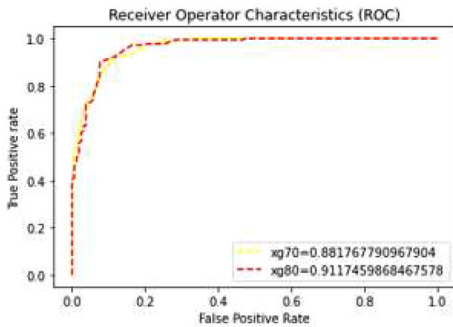
b) K-Nearest Neighbor



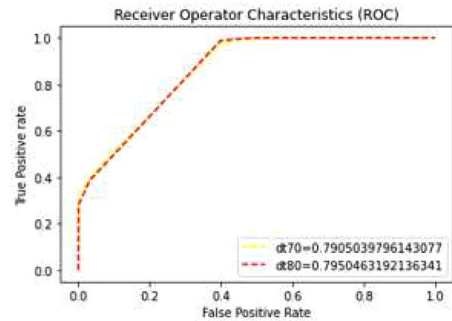
c) Naïve Bayes



d) Support Vector Machine



e) XG-Boost



f) Decision Tree

Fig. 8. Comparison between ROC between the test-train split of each model.

of K-fold cross-validation. However, this enabled a comprehensive understanding of the model's efficacy under diverse training-test ratios and cross-validation folds. And, the performance metrics from all scenarios were meticulously compared, unraveling insights into the impact of the distribution and cross-validation on the models' predictive prowess.

As a result of this methodical investigation, the optimal training-test split and the most effective K value for cross-validation for each model were discerned. This identification was grounded in the evaluation of accuracy, precision, and recall across various scenarios. So, the K-fold score for each split is shown in Table 4.

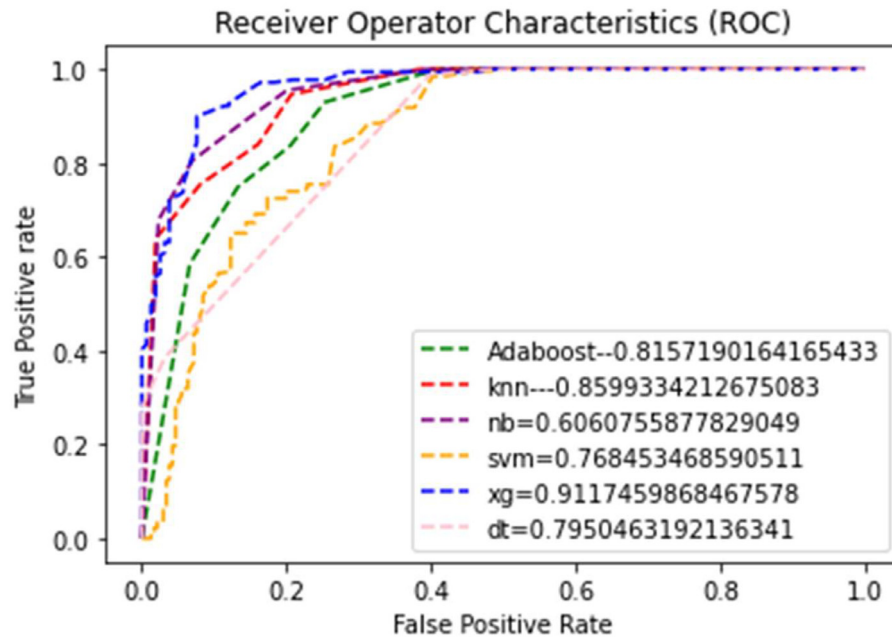


Fig. 9. Comparison between ROC of the best test-train split of each model.

Further, the split that yielded the highest values for these metrics has been selected and there is a comparison between the best test-train split of each model with each other as shown in Fig. 6.

The CHAMP tool utilizes XGBoost for the purpose of evaluating their performance in disease prognosis and diagnosis. To validate this, various metrics are employed, including prediction accuracy, k-fold score, and receiver operating characteristic (ROC) analysis. ROC analysis is particularly important as it helps gauge a model's ability to distinguish between positive and negative cases. Moreover, to ensure the reliability and robustness of the model evaluations, a K-fold cross-validation technique is applied. However, this technique entails dividing the dataset into K subsets, training the model on K-1 subsets, and validating it on the remaining subset. By averaging performance metrics across all K iterations, a more accurate assessment of the model's performance is obtained. However, the study systematically explores different training-test data splits, such as an 80:20 split and a 70:30 split, to evaluate how data distribution impacts model performance and the comparison of performance metrics between these splits helps identify the split that yields superior accuracy, precision, and recall for each model.

In summary, we employ a range of machine learning algorithms and comprehensive evaluation techniques, including ROC analysis and cross-validation, to rigorously assess model performance. XGBoost, in particular, stands out as the best-performing classifier with 91% accuracy in the 80:20 train-test split. Additional analyses, such as confusion matrix and ROC curve analysis, are conducted to gain deeper insights into its performance characteristics. This comprehensive approach allows for data-driven decisions in disease prognosis and diagnosis. However, the confusion matrix provided insights into the model's classification performance and allowed to identification of any potential imbalances or misclassifications in the predictions which are shown in Fig. 7.

Moreover, to optimize the performance of each predictive model in CHAMP is evaluated with the accuracy measures such precision and recall. The varied model is assessed with target features to measure the overall accuracy of the models using ROC curve analysis. The ROC curve provides valuable insights into the model's discriminative power, robustness, and overall performance. Moreover, a model with a higher score is considered better at distinguishing between positive and negative instances and used as a summary statistic to compare different models or to determine an optimal threshold for classification.

In addition to the confusion matrix, the ROC curve has been utilized to evaluate the model's performance shown in Fig. 9 and comparison between each test train split of each model is shown in Fig. 8. However, a model that has better performance in comparison to distinguishing between the positive and negative classes has a bigger area under the ROC curve. By incorporating the confusion matrix and ROC curve analysis, a comprehensive understanding of the model's performance characteristics has been gained. This allowed me to make informed decisions about selecting the best-performing model among the different options. Moreover, the combination of accuracy, precision, recall, confusion matrix, and ROC curve analysis provided a robust evaluation framework, enabling us to assess the models' effectiveness and make data-driven decisions for prognosis and diagnosis of disease.

Tool interface

The best-performing model, XGBoost was selected and saved by using Python's object serialization module which is further used to develop and design a web application using Flask, which is a Python-based microweb framework., Web implementation is a complex

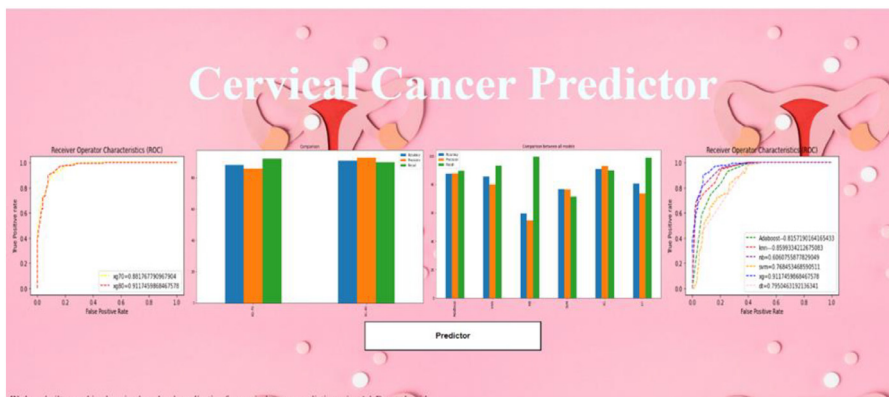


Fig. 10. The user interface of the Home page of CHAMP.

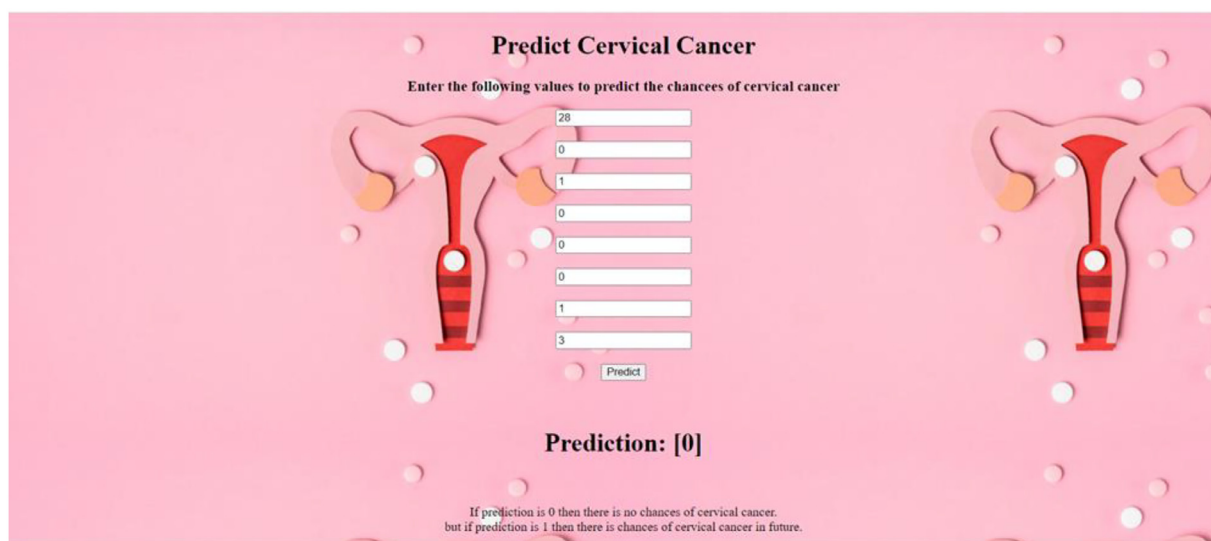


Fig. 11. The user interface of the Prediction page of CHAMP when the result is negative.

process that includes designing a framework using Flask which is used to implement the machine learning algorithm on the web page and it is required for the web page to predict the results of the prediction, as it did in the local systems on which the machine learning algorithms has been built. Further, HTML files are used to bring basics of a web page like adding the text and buttons which can perform clicking actions, and also allow users to interact with the webpage allowing them to provide inputs. However, the process started with creating a virtual environment with Python version: 3.9.0 and configuring the PyCharm Integrated development environment to set up the project. Further, there is the incorporation of folders that contain CSS files & images, and HTML files respectively which all result in a full-fledged web page that was designed and is capable of prognosis and diagnosis of cervical cancer. Although, designed webpage consists of two interconnected pages which first page is the home page which serves as a gateway to the application and provides the information that allows the user to gain insight regarding the best-performing model and the attributes used to show in Fig. 10. Moreover, this page presents a concise overview of the predictive models employed in the tool and highlights the best-performing model based on evaluation metrics which also allow users to distinguish the significance of each attribute and understand their contribution to the prediction process. Additionally, the home page offers a user-friendly interface with intuitive navigation, allowing users to seamlessly transition to the prediction page. With its clear and concise presentation, the home page sets the stage for users to explore the tool and make informed decisions regarding cervical cancer risk assessment.

Further, the second page, the prediction page of the cervical cancer prediction tool is where users can input their information and obtain the prediction result for cervical cancer risk. Moreover, this page features a set of eight attribute text boxes, where users can enter their relevant details such as age, number of sexual partners, and other required fields which enable users to click the “Predict” button to initiate the prediction process. However, the prediction model, trained on a comprehensive dataset and using advanced machine learning algorithms, analyses the input data and generates a prediction outcome. However, If the result is 0, it indicates that the model predicts a low risk of cervical cancer based on the provided information and on the other hand, if the result is 1, it

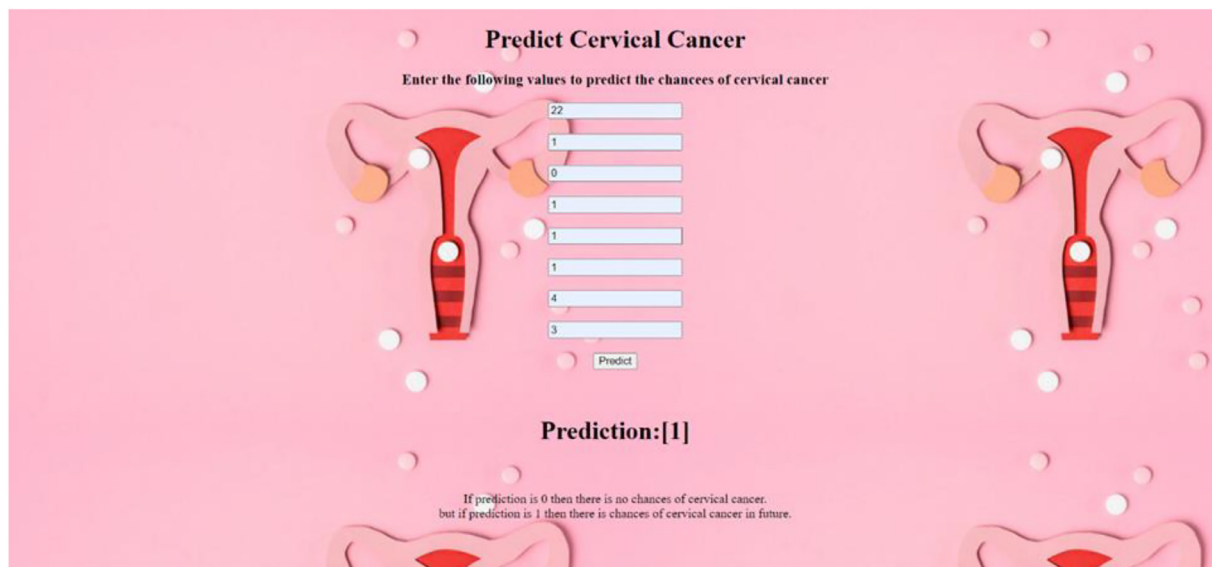


Fig. 12. The user interface of the Prediction page of CHAMP when results positive.

suggests a higher risk of cervical cancer as shown in Figs. 11,12. So, the prediction page provides users with valuable insights into their potential risks and empowers them to take proactive steps towards their health and well-being.

Conclusion

Our study focused on predicting cervical cancer using a dataset and comparing the performance of six different models. After careful evaluation, it was found that XGBoost exhibited the best performance among the models, demonstrating its effectiveness in accurately predicting cervical cancer outcomes. Moreover, to provide a user-friendly interface for utilizing the predictive model, a Flask tool was developed. This tool allows users to input relevant information and obtain real-time predictions regarding their risk of cervical cancer. By creating this interactive platform, the project aimed to enhance accessibility and facilitate informed decision-making for individuals concerned about cervical cancer. The study not only emphasized model performance but also prioritized user experience and convenience. The integration of Flask, a web development framework, enabled the development of a dynamic and responsive tool that can be easily accessed and utilized by users. The tool provides a seamless and intuitive interface, making it straightforward for individuals to obtain personalized predictions and gain insights into their potential risk of cervical cancer.

Overall, this project successfully employed machine learning techniques, model evaluation, and web development to create a predictive model for cervical cancer and deliver it through a user-friendly Flask tool. By leveraging advanced algorithms and providing an accessible platform, the project aimed to contribute to early detection and prevention efforts, ultimately enhancing cervical cancer awareness and supporting healthcare decision-making processes.

Ethics statements

Not applicable.

Funding

This study was conducted without any specific funding or financial support.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRedit authorship contribution statement

Ritu Chauhan: Conceptualization, Formal analysis, Supervision, Project administration, Writing – review & editing. **Anika Goel:** Conceptualization, Methodology, Software, Writing – original draft. **Bhavya Alankar:** Conceptualization, Formal analysis, Writing – review & editing. **Harleen Kaur:** Conceptualization, Formal analysis, Supervision, Project administration, Writing – review & editing.

Data availability

No data was used for the research described in the article.

Acknowledgments

The authors would like to express their gratitude to all individuals who have contributed to the study. We appreciate the valuable input and support received from colleagues, mentors, and institutions during the course of this research.

References

- [1] S.F. Abdo, M.A.B.O. Rizka, F.A. Maghraby, Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques, *IEEE Access* 6 (2018) 59475–59485, doi:10.1109/ACCESS.2018.2874063.
- [2] R. Alsmariy, G. Healy, H. Abdelhafez, Predicting cervical cancer using machine learning methods, *Int. J. Adv. Comput. Sci. Appl.* 11 (7) (2020) 173–184, doi:10.14569/IJACSA.2020.0110723.
- [3] H. Bandyopadhyay, Segmentation of pap smear images for cervical cancer detection, *Int. J. Comput. Sci. Eng.* 8 (6) (2020) 30–33.
- [4] M. Schiffman, J. Doorbar, N. Wentzensen, S. de Sanjose, C. Fakhry, B.J. Monk, P.E. Castle, Carcinogenic human papillomavirus infection, *Nat. Rev. Dis. Primers* 7 (1) (2021) 19 1– Schiffman.
- [5] M. Arbyn, E. Weiderpass, L. Bruni, S. de Sanjosé, M. Saraiya, J. Ferlay, F. Bray, Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis, *Lancet Glob. Health* 8 (2) (2020) e191–e203.
- [6] K. Chatterjee, U. Maulik, A. Ghosh, R. De, A comprehensive analysis of cervical cancer prediction using machine learning techniques, *J. Med. Syst.* 45 (1) (2021) 1–18.
- [7] C. Simoens, F. Goffin, P. Simon, J.C. Noel, Cervical cancer and HPV infection: ongoing therapeutic research to counteract the action of E6 and E7 oncoproteins, *Drug Discov. Today* 23 (5) (2018) 1089–1098.
- [8] M.R. McCredie, K.J. Sharples, C. Paul, J. Baranyai, G. Medley, R.W. Jones, D.C. Skegg, Natural history of cervical neoplasia and risk of invasive cancer in women with cervical intraepithelial neoplasia 3: a retrospective cohort study, *Lancet Oncol.* 9 (5) (2008) 425–434.
- [9] S. Vaccarella, S. Franceschi, D. Zaridze, M. Poljak, P. Veerus, M. Plummer, G. Clifford, Preventable fractions of cervical cancer via effective screening in six Baltic, central, and eastern European countries 2017–40: a population-based study, *Lancet Oncol.* 14 (7) (2013) 687–695.
- [10] R. Sankaranarayanan, J. Ferlay, Worldwide burden of gynaecological cancer: the size of the problem, *Best Pract. Res. Clin. Obstet. Gynaecol.* 20 (2) (2006) 207–225.
- [11] D. Saslow, D. Solomon, H.W. Lawson, M. Killackey, S.L. Kulasingam, J. Cain, R.A. Smith, American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer, *CA Cancer J. Clin.* 62 (3) (2012) 147–172.
- [12] R.A. Smith, K.S. Andrews, D. Brooks, S.A. Fedewa, D. Manassaram-Baptiste, D. Saslow, O.W. Brawley, Cancer screening in the United States, 2018: a review of current American Cancer Society guidelines and current issues in cancer screening, *CA Cancer J. Clin.* 68 (4) (2018) 297–316.
- [13] M. Batta, Machine learning algorithms—a review, *Int. J. Sci. Res. (IJSR)* 18 (8) (2018) 381–386, doi:10.21275/ART20203995.
- [14] R. Caruana, A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in: *Proceedings of the ACM International Conference Proceeding Series*, 148, 2006, pp. 161–168, doi:10.1145/1143844.1143865.
- [15] B. Charbuty, A. Abdulazeez, Classification based on decision tree algorithm for machine learning, *J. Appl. Sci. Technol. Trends* 2 (01) (2021) 20–28, doi:10.38094/jast20165.
- [16] S. Chen, G.I. Webb, L. Liu, X. Ma, A novel selective naïve Bayes algorithm, *Knowl. Based Syst.* 192 (xxxx) (2020) 105361, doi:10.1016/j.knosys.2019.105361.
- [17] X. Deng, Y. Luo, C. Wang, Analysis of risk factors for cervical cancer based on machine learning methods, in: *Proceedings of 5th IEEE International Conference on Cloud Computing and Intelligence Systems, CCIS*, 2018, 2019, pp. 631–635, doi:10.1109/CCIS.2018.8691126.
- [18] D. Merlin, Improved classification accuracy for identification of cervical cancer, in: *Proceedings of the International Conference on Medical Informatics (ICMI)*, 3307, 2024, pp. 245–258.
- [19] D.M. Reif, A.A. Motsinger, B.A. Mckinney, J.E. Crowe, J.H. Moore, Feature selection using a random forests classifier for the integrated analysis of multiple data types, in: *Proceedings of the IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, 2006 September, doi:10.1109/CIBCB.2006.330987.
- [20] P. Ramos, M.R. Carvalho, M.S. Nogueira, M. Raju, Random forest prognostic factor in colorectal cancer, *J. Phys. Conf. Ser.* (2019), doi:10.1088/1742-6596/1217/1/012098.
- [21] L. Denny, R. Anorlu, Cervical cancer in Africa, *Cancer Epidemiol. Biomark. Prev.* 21 (9) (2012) 1434–1438, doi:10.1158/1055-9965.EPI-12-0334.
- [22] F.Y. Oisanwo, J.E.T. Akinsola, O. Awodele, J.O. Hinmikaiye, O. Olakanmi, J. Akinjobi, Supervised machine learning algorithms: classification and comparison, *Int. J. Comput. Trends Technol.* 48 (3) (2017) 128–138, doi:10.14445/22312803/ijctt-v48p126.
- [23] Y. Freund, L. Mason, The alternating decision tree learning algorithm, in: *Proceedings of the International Conference on Machine Learning*, 99, 1999, pp. 124–133.
- [24] V. Kumar, Feature selection: a literature review, *Smart Comput. Rev.* 4 (3) (2014), doi:10.6029/smartcr.2014.03.007.
- [25] A. Ronacher, Flask: a python microframework, *Python Pap.* 4 (1) (2010) 1–9.
- [26] Grinberg, M. (2018). Flask web development with python tutorial. The flask mega-tutorial. Retrieved from <https://blog.miguelgrinberg.com/post/the-flask-mega-tutorial-part-i-hello-world>
- [27] C. White, D. Sulakhe, Data-driven web applications with Flask and SQLAlchemy, in: *Proceedings of the XSEDE Conference on Diversity, Big Data, and Science at Scale*, 2016, p. 43. Article No.
- [28] Q. Han, W. Jiang, S. Hu, H. Tan, Design and implementation of an efficient web framework based on Flask, in: *Proceedings of the International Conference on Smart Grid and Electrical Automation (ICSGEA)*, 2017, pp. 262–266.
- [29] Z. Zhang, Z. Wu, X. Lin, J. Zhou, Research and application of Python Flask framework in Web development, in: *Proceedings of the International Conference on Logistics, Informatics and Service Sciences (LISS)*, 2016, pp. 1–4.
- [30] J. Zeng, Z. Wang, J. Gu, Development of an intelligent WebGIS platform based on Flask and Python, in: *Proceedings of the 4th International Conference on Frontiers of Image, Video and Signal Processing (ICFIVSP)*, 2020, pp. 1–6.
- [31] J. Choi, K. Choi, Implementation of Web-based diagnostic system using Flask web framework in Python, in: *Proceedings of the International Conference on Information and Communication Technology Convergence (ICTC)*, 2016, pp. 1191–1193.
- [32] J. Evers, Building a RESTful API with Flask and Python, *J. Open Source Softw.* 4 (44) (2019) 1813.
- [33] D. Hubert, L. Lathuilière, Flask and Python integrated into engineering training, in: *Proceedings of the International Conference on Computing, Networking and Communications (ICNC)*, 2017, pp. 756–760.
- [34] M. Lichman, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Science, 2013 Retrieved from <https://archive.ics.uci.edu/ml/index.php>.
- [35] C. Zhang, H. Jin, Web service development based on Flask and Python, in: *Proceedings of the International Conference on Information Science, Big Data and Internet of Things (ISBDI)*, 2019, pp. 1–4.

- [36] J. Lu, E. Song, A. Ghoneim, M. Alrashoud, Machine learning for assisting cervical cancer diagnosis : an ensemble approach, *Future Gener. Comput. Syst.* 106 (2020) 199–205, doi:[10.1016/j.future.2019.12.033](https://doi.org/10.1016/j.future.2019.12.033).
- [37] C. Marzban, The ROC curve and the area under it as performance measures, *Weather Forecast.* 19 (6) (2004) 1106–1114, doi:[10.1175/825.1](https://doi.org/10.1175/825.1).
- [38] M. Mehmood, M. Rizwan, M. Gregus ml, S. Abbas, Machine learning assisted cervical cancer detection, *Front. Public Health* 9 (December) (2021) 1–14, doi:[10.3389/fpubh.2021.788376](https://doi.org/10.3389/fpubh.2021.788376).
- [39] A. Moldagulova, Documents, in: *Proceedings of the International Conference on Information Systems (ICIS), 2017*, pp. 665–671.
- [40] A. Pandey, A. Jain, Comparative analysis of KNN algorithm using various normalization techniques, *Int. J. Comput. Netw. Inf. Secur.* 9 (11) (2017) 36–42, doi:[10.5815/ijcnis.2017.11.04](https://doi.org/10.5815/ijcnis.2017.11.04).
- [41] D. Parikh, V. Menon, Machine learning applied to cervical cancer data, *Int. J. Math. Sci. Comput.* 5 (1) (2019) 53–64, doi:[10.5815/ijmsc.2019.01.05](https://doi.org/10.5815/ijmsc.2019.01.05).
- [42] A. Rakotomamonjy, Optimizing area under ROC curve with SVMs, in: *Proceedings of the International Conference on Machine Learning (ICML), 2004*, pp. 71–80.
- [43] D.P. Solomatine, D.L. Shrestha, AdaBoost.RT: a boosting algorithm for regression problems, in: *Proceedings of the IEEE International Conference on Neural Networks*, 2, 2004, pp. 1163–1168, doi:[10.1109/ijcnn.2004.1380102](https://doi.org/10.1109/ijcnn.2004.1380102).
- [44] C. Tu, H. Liu, B. Xu, AdaBoost typical Algorithm and its application research, in: *Proceedings of the MATEC Web of Conferences*, 139, 2017, doi:[10.1051/mateconf/201713900222](https://doi.org/10.1051/mateconf/201713900222).
- [45] K. Vembandasamp, R.R. Sasipriyap, E. Deepap, Heart diseases detection using naive bayes algorithm, *IJISSET-Int. J. Innov. Sci. Eng. Technol.* 2 (9) (2015) 1–4 www.ijiset.com.
- [46] S.V.N. Vishwanathan, M.N. Murty, SSVN: a simple SVM algorithm, in: *Proceedings of the International Joint Conference on Neural Networks*, 3, 2002, pp. 2393–2398, doi:[10.1109/ijcnn.2002.1007516](https://doi.org/10.1109/ijcnn.2002.1007516).
- [47] Vogel, P. (2017). A dashboard for automatic monitoring python web services.
- [48] G. Langs, S. Röhricht, J. Hofmanninger, F. Prayer, J. Pan, C. Herold, H. Prosch, Machine learning: from radiomics to discovery and routine, *Radiologe* 58 (2018), doi:[10.1007/s00117-018-0407-3](https://doi.org/10.1007/s00117-018-0407-3).
- [49] W. Li, Y. Yin, X. Quan, H. Zhang, Gene expression value prediction based on XGBoost algorithm, *Front. Genet.* 10 (November) (2019) 1–7, doi:[10.3389/fgene.2019.01077](https://doi.org/10.3389/fgene.2019.01077).
- [50] R. Chauhan, H. Kaur, V. Chang, Advancement and applicability of classifiers for variant exponential model to optimize the accuracy for deep learning, *J. Ambient Intell. Human Comput.*, Springer (2017), doi:[10.1007/s12652-017-0561-x](https://doi.org/10.1007/s12652-017-0561-x).
- [51] M. Vu, J. Yu, O.A. Awolude, L. Chuang, Cervical cancer worldwide, *Curr. Probl. Cancer* 42 (5) (2018) 457–465, doi:[10.1016/j.currproblcancer.2018.06.003](https://doi.org/10.1016/j.currproblcancer.2018.06.003).
- [52] R. Chauhan, H. Kaur, V. Chang, An Optimized integrated Framework of Big Data Analytics Managing Security and Privacy in Healthcare Data, *Wirel. Pers. Commun.* (2020) <https://link.springer.com/article/10.1007/s11277-020-07040-8>.