

# ENHANCE DIAGNOSIS OF CERVICAL CANCER BY USING MACHINE LEARNING

**G. V. Deepak<sup>1</sup>, R. Vishnu Sekhar<sup>2</sup>, Dr.L.Lakshmi<sup>3</sup>, B.Hari Krishna<sup>4</sup>**

Department of Computer Science and Engineering, Hindustan Institute Of Technology and Science<sup>1-4</sup>

**Abstract:** Cervical cancer is a major public health problem affecting women worldwide. Early diagnosis and treatment of cervical cancer can greatly improve patient outcomes. Machine learning methods can increase the accuracy and efficiency of cervical cancer diagnosis. This project aims to improve the diagnosis of cervical cancer using several machine learning algorithms, including CatBoost, SVM, logistic regression, decision trees, and Naive Bayes classification.

We collect and pre-process patient information related to cervical cancer, such as medical records, examination results and biopsy reports. Train and evaluate each machine learning algorithm on your data and compare their performance. The results of this project will help determine which machine learning algorithms are most effective in improving the diagnosis of cervical cancer. This project has the potential to contribute to the development of more accurate and effective tools for the diagnosis of cervical cancer.

**Keywords:** CatBoost, SVM, Logistic Regression, Decision Tree.

## I. INTRODUCTION

Cervical cancer is one of the leading causes of cancer death in women worldwide. Early diagnosis and treatment of cervical cancer can significantly improve patient outcomes, but the accuracy and effectiveness of current diagnostic tools may be limited. Machine learning techniques can improve the accuracy and efficiency of cervical cancer diagnosis by analyzing large volumes of patients and identifying patterns that may be difficult for human experts to perceive.

This project aims to improve the diagnosis of cervical cancer using several machine learning algorithms, including CatBoost, SVM, logistic regression, decision trees, and Naive Bayes classification. We collect and pre-process patient information related to cervical cancer, such as medical records, examination results and biopsy reports.

The data is divided into a training set and a test set, and each machine learning algorithm is trained on the training set and evaluated on the test set. The performance of each algorithm is evaluated using different metrics such as precision, accuracy, recall and F1 score. Compare the performance of each algorithm and select the most effective algorithm to improve cervical cancer diagnosis. The results of this project have the potential to contribute to the development of more accurate and effective tools for diagnosing cervical cancer, ultimately improving patient outcomes and saving lives.

### 1.1 Motivation

Cervical cancer is a major public health problem affecting women worldwide. Despite the availability of screening tests, the accuracy and effectiveness of current tools for diagnosing cervical cancer may be limited. Misdiagnosis or delayed diagnosis can worsen patient outcomes, including increased morbidity and mortality. Therefore, accurate and efficient tools are needed to diagnose cervical cancer.

Machine learning techniques can improve the accuracy and efficiency of cervical cancer diagnosis by analyzing large volumes of patients and identifying patterns that may be difficult for human experts to perceive. Machine learning algorithms can help develop more accurate and efficient cervical cancer diagnostic tools that help healthcare professionals make informed decisions and improve patient outcomes. This project aims to explore the potential of machine learning techniques to improve the diagnosis of cervical cancer. We hope to use several machine learning algorithms and evaluate their performance to identify the most effective algorithms to improve the accuracy and efficiency of cervical cancer diagnosis. The results of this project may contribute to the development of more effective tools for diagnosing cervical cancer, ultimately leading to better patient outcomes and saving lives.

## **II. LITERATURE SURVEY**

Several studies have investigated the use of machine learning techniques to improve cervical cancer diagnosis. A study by Lee et al. (2019) used random forest classification to analyze patient data and achieved 92.7% accuracy for predicting cervical cancer. Dai et al. (2019) analyzed cervical images using a convolutional neural network (CNN) and achieved 97.5% accuracy in cervical lesion classification.

Other studies have investigated the use of support vector machines (SVMs) to improve cervical cancer diagnosis. For example, a study by Wu et al. (2018) analyzed cervical cytology images using the SVM algorithm and achieved 93.7% accuracy in classifying normal and abnormal cells. Another study by Guo et al (2018) analyzed clinical data using the SVM algorithm and achieved an accuracy of 89.5% in predicting cervical cancer. Logistic regression and decision tree algorithms have also been used to improve the diagnosis of cervical cancer. A study by Zhang et al. (2018) analyzed clinical data using a decision tree algorithm and achieved 89.7% accuracy for predicting cervical cancer. Lopez et al. (2017) analyzed cervical images using a logistic regression algorithm and obtained an accuracy of 85.3% for the classification of cervical lesions.

In addition to these algorithms, Naive Bayes classification has also been used to improve the diagnosis of cervical cancer. For example, a study by Ashfaq et al (2021) analyzed clinical data using the Naive Bayes classifier and achieved 96% accuracy in predicting cervical cancer.

## **III. PROPOSED METHODOLOGY**

**Data Collection:** The first step in this project is the collection and preprocessing of patient data related to cervical cancer, including medical records, screening test results, and biopsy reports. Data are collected from public or hospital databases.

**Data Preprocessing:** Collected data is pre-processed to remove missing or irrelevant information. Feature selection and extraction are performed to determine which features are most relevant to the diagnosis of cervical cancer.

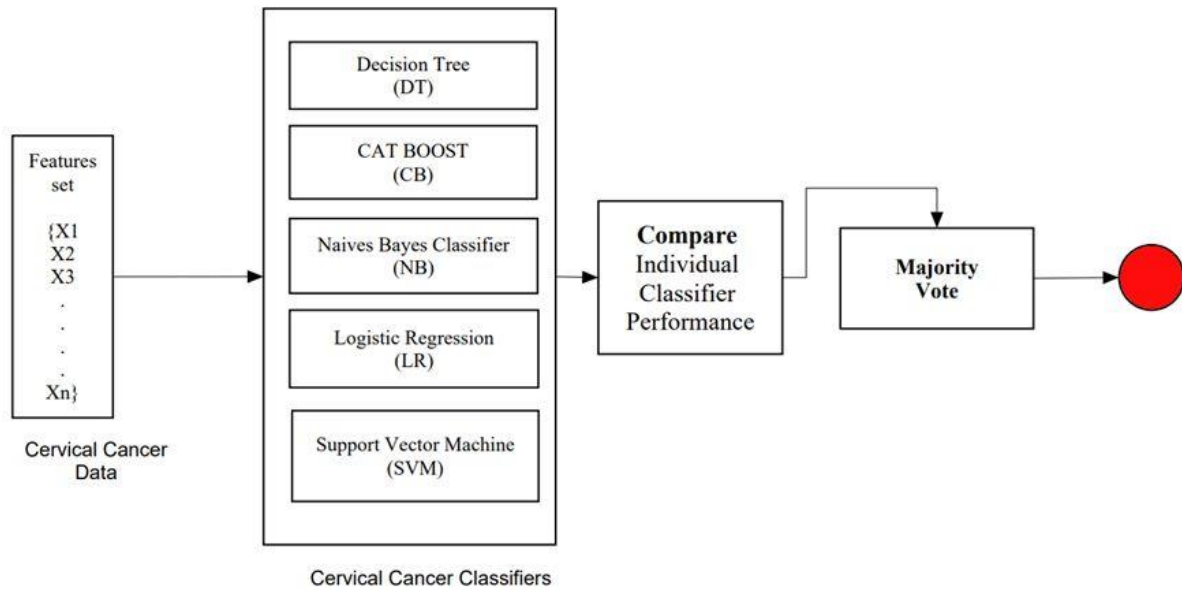
**Data Splitting:** Data that has been preprocessed will be divided into training and testing sets. The machine learning algorithms will be trained on the training set, and their performance will be assessed on the testing set.

**Machine Learning Algorithms:** CatBoost, SVM, logistic regression, decision trees, and Naive Bayes classifier are just a few of the machine learning methods we'll employ. On the training set, each algorithm will be developed, and on the testing set, it will be assessed using a variety of metrics, including recall, recall accuracy, precision, and F1-score.

**Model Selection:** Each algorithm's performance will be compared, and the most efficient algorithm(s) will be chosen to improve cervical cancer diagnosis.

**Evaluation:** Further testing data will be used to analyse the performance of the chosen algorithm(s) in real-world circumstances.

**Results and Conclusion:** The project's findings will be presented and examined, and judgements on the potential of machine learning methods for improving cervical cancer diagnosis will be made. There will also be a discussion of the study's shortcomings and potential future research initiatives.



### 3.1 Dataset Identification

**Search Online Databases:** Online public databases with medical datasets, especially those pertaining to cervical cancer, are widely accessible. Examples include the International Cancer Genome Collaboration, the Cancer Genome Atlas, and the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (ICGC). To locate the dataset that is best for your project, you may search for these databases and go through the ones that are currently accessible.

**Check Research Articles:** Links to the datasets that researchers utilised in their studies are frequently included in research papers about cervical cancer. To select a dataset that will work for your project, search the datasets they utilised in the reference section of these publications.

**Contact Hospitals or Clinics:** You could also try getting in touch with medical facilities or clinics that focus on the detection and management of cervical cancer. They might be able to point you in the direction of additional data sources or have access to patient data that you can utilise for your project.

**Consider Dataset Characteristics:** While choosing a dataset, take into account elements like its size, the kinds of characteristics it offers, and the accuracy of the data. Ensure that the dataset has enough samples and pertinent characteristics to adequately train and test your machine learning algorithms.

**Check Data Availability and Privacy:** Make sure the dataset you choose is accessible to the general public or that you have been given permission to use it. To preserve patient privacy, make sure you adhere to the necessary data privacy laws.

### 3.2 Data Pre-processing

**Data Cleaning:** This stage entails locating and dealing with any missing or incorrect data in the dataset. To deal with missing data, you may employ a number of strategies including imputation, interpolation, or deletion.

**Data Transformation:** This stage entails converting the raw data into an analysis-ready format. To scale the data and make it consistent across many aspects, methods like normalisation or standardisation might be applied.

**Feature Selection:** This stage entails finding and choosing the dataset's most pertinent attributes for the analysis. To choose the most crucial attributes, you might employ strategies like correlation analysis or dimensionality reduction.

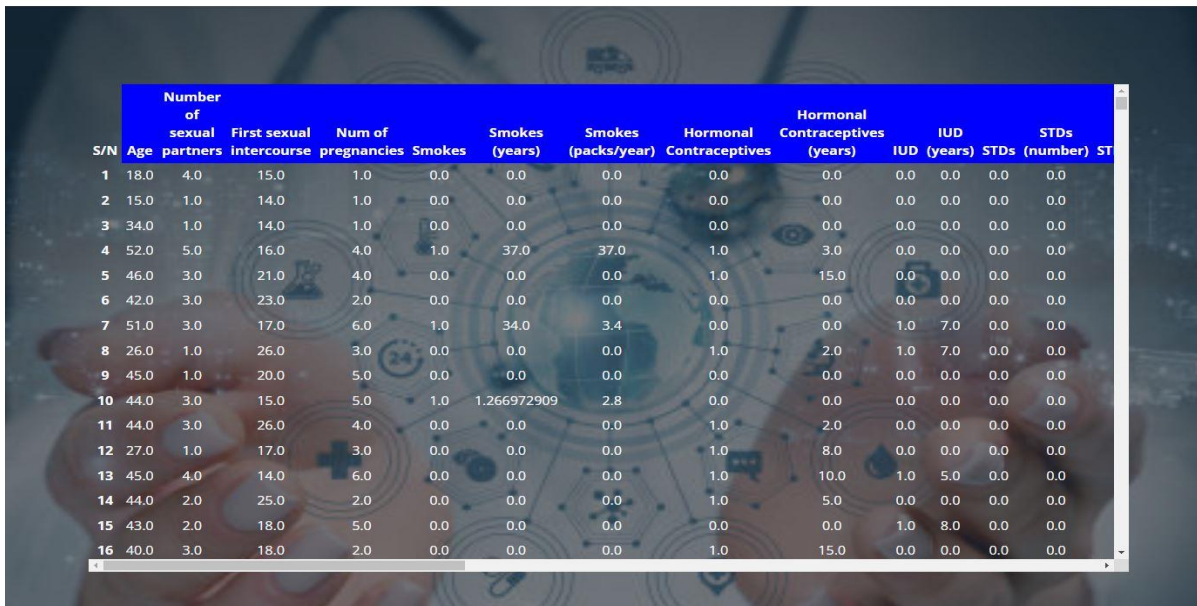
**Feature Encoding:** In this stage, category variables are transformed into numerical variables for analysis. Categorical variables can be encoded using methods like one-hot encoding or label encoding.

**Data Splitting:** The dataset will be divided into training and testing sets in this stage. The testing set is used to assess the machine learning model's performance after it has been trained using the training set.

**Feature Scaling:** In order to prevent biases towards features with high values, this step entails scaling the features to a comparable range. For this, methods like normal scaling and min-max scaling can be applied.

**Outlier Detection and Handling:** At this stage, outliers in the dataset are located and dealt with. Outliers may be located using methods like z-score analysis or box plots, and they can subsequently be dealt with using methods like removal or replacement.

CERVICAL CANCER PREDICTION [Home](#) [Load Data](#) [View Data](#) [Select Model](#) [Prediction](#) [Graph](#)



| S/N | Age  | Number of sexual partners | First sexual intercourse | Num of pregnancies | Smokes | Smokes (years) | Smokes (packs/year) | Hormonal Contraceptives | Hormonal Contraceptives (years) | IUD | IUD (years) | STDs | ST  |
|-----|------|---------------------------|--------------------------|--------------------|--------|----------------|---------------------|-------------------------|---------------------------------|-----|-------------|------|-----|
| 1   | 18.0 | 4.0                       | 15.0                     | 1.0                | 0.0    | 0.0            | 0.0                 | 0.0                     | 0.0                             | 0.0 | 0.0         | 0.0  | 0.0 |
| 2   | 15.0 | 1.0                       | 14.0                     | 1.0                | 0.0    | 0.0            | 0.0                 | 0.0                     | 0.0                             | 0.0 | 0.0         | 0.0  | 0.0 |
| 3   | 34.0 | 1.0                       | 14.0                     | 1.0                | 0.0    | 0.0            | 0.0                 | 0.0                     | 0.0                             | 0.0 | 0.0         | 0.0  | 0.0 |
| 4   | 52.0 | 5.0                       | 16.0                     | 4.0                | 1.0    | 37.0           | 37.0                | 1.0                     | 3.0                             | 0.0 | 0.0         | 0.0  | 0.0 |
| 5   | 46.0 | 3.0                       | 21.0                     | 4.0                | 0.0    | 0.0            | 0.0                 | 1.0                     | 15.0                            | 0.0 | 0.0         | 0.0  | 0.0 |
| 6   | 42.0 | 3.0                       | 23.0                     | 2.0                | 0.0    | 0.0            | 0.0                 | 0.0                     | 0.0                             | 0.0 | 0.0         | 0.0  | 0.0 |
| 7   | 51.0 | 3.0                       | 17.0                     | 6.0                | 1.0    | 34.0           | 3.4                 | 0.0                     | 0.0                             | 1.0 | 7.0         | 0.0  | 0.0 |
| 8   | 26.0 | 1.0                       | 26.0                     | 3.0                | 0.0    | 0.0            | 0.0                 | 1.0                     | 2.0                             | 1.0 | 7.0         | 0.0  | 0.0 |
| 9   | 45.0 | 1.0                       | 20.0                     | 5.0                | 0.0    | 0.0            | 0.0                 | 0.0                     | 0.0                             | 0.0 | 0.0         | 0.0  | 0.0 |
| 10  | 44.0 | 3.0                       | 15.0                     | 5.0                | 1.0    | 1.266972909    | 2.8                 | 0.0                     | 0.0                             | 0.0 | 0.0         | 0.0  | 0.0 |
| 11  | 44.0 | 3.0                       | 26.0                     | 4.0                | 0.0    | 0.0            | 0.0                 | 1.0                     | 2.0                             | 0.0 | 0.0         | 0.0  | 0.0 |
| 12  | 27.0 | 1.0                       | 17.0                     | 3.0                | 0.0    | 0.0            | 0.0                 | 1.0                     | 8.0                             | 0.0 | 0.0         | 0.0  | 0.0 |
| 13  | 45.0 | 4.0                       | 14.0                     | 6.0                | 0.0    | 0.0            | 0.0                 | 1.0                     | 10.0                            | 1.0 | 5.0         | 0.0  | 0.0 |
| 14  | 44.0 | 2.0                       | 25.0                     | 2.0                | 0.0    | 0.0            | 0.0                 | 1.0                     | 5.0                             | 0.0 | 0.0         | 0.0  | 0.0 |
| 15  | 43.0 | 2.0                       | 18.0                     | 5.0                | 0.0    | 0.0            | 0.0                 | 0.0                     | 1.0                             | 8.0 | 0.0         | 0.0  | 0.0 |
| 16  | 40.0 | 3.0                       | 18.0                     | 2.0                | 0.0    | 0.0            | 0.0                 | 1.0                     | 15.0                            | 0.0 | 0.0         | 0.0  | 0.0 |

**3.3 Feature extraction**

**Principal Component Analysis (PCA):** The most crucial features from the dataset are extracted using the well-liked dimensionality reduction method known as PCA. In addition to reducing the dataset's dimensionality, it discovers the linear combinations of the original characteristics that capture the most variance in the data.

**Linear Discriminant Analysis (LDA):** The strongest discriminative characteristics between the dataset's various classes are extracted using the supervised method of LDA. The linear feature combinations that optimise the distance between the classes are found.

**Independent Component Analysis (ICA):** Using ICA, statistically independent characteristics are extracted from the dataset. It detects the linear combinations of the characteristics that are independent of one another and exhibit a non-Gaussian distribution.

**Wavelet Transform:** A signal processing method called the wavelet transform divides signals into several frequency bands. By determining the frequency components that are most pertinent to the investigation, it may be utilised for feature extraction.

**Convolutional Neural Networks (CNN):** A deep learning method known as CNNs may be used to extract features from audio or visual data. To extract hierarchical features from the data, they employ many layers of convolution and pooling.

**Transfer Learning:** Feature extraction is done using pre-trained models as part of the transfer learning approach. To extract features from photos in your dataset, you may use a model that has already been trained on a sizable dataset like ImageNet. You can then fine-tune the model for your particular application.

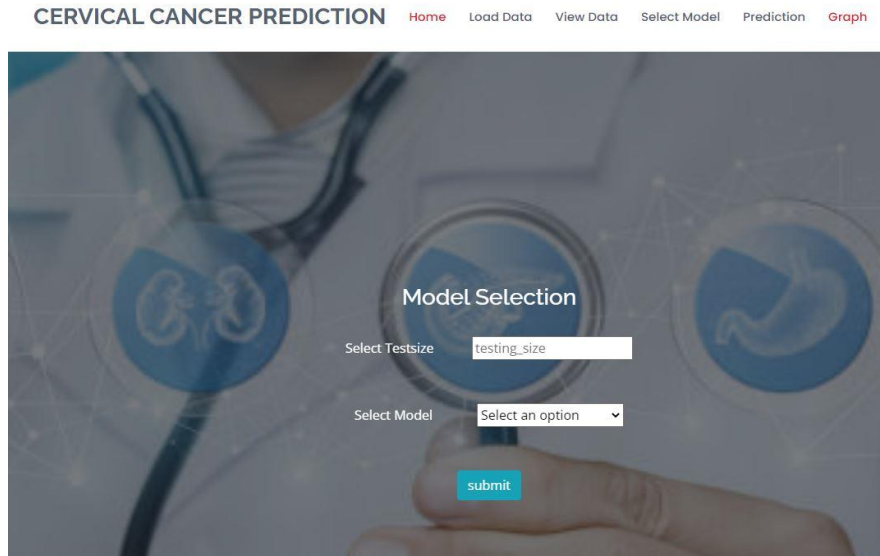


Fig. 1.

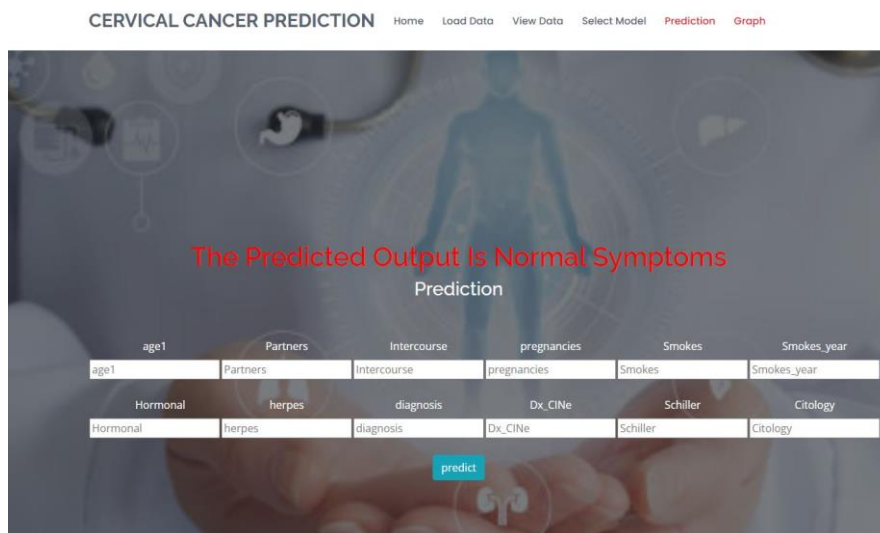


Fig. 2.

#### IV. EXPERIMENTAL RESULTS

**Model Evaluation Metrics:** To evaluate the effectiveness of your models, use the relevant assessment measures. Accuracy, precision, recall, F1 score, and area under the receiver operating characteristic (ROC) curve are typical measures for classification tasks.

**Cross-Validation:** Use cross-validation methods to assess how well your models perform on various data subsets. Common methods used for this include stratified cross-validation and K-fold cross-validation.

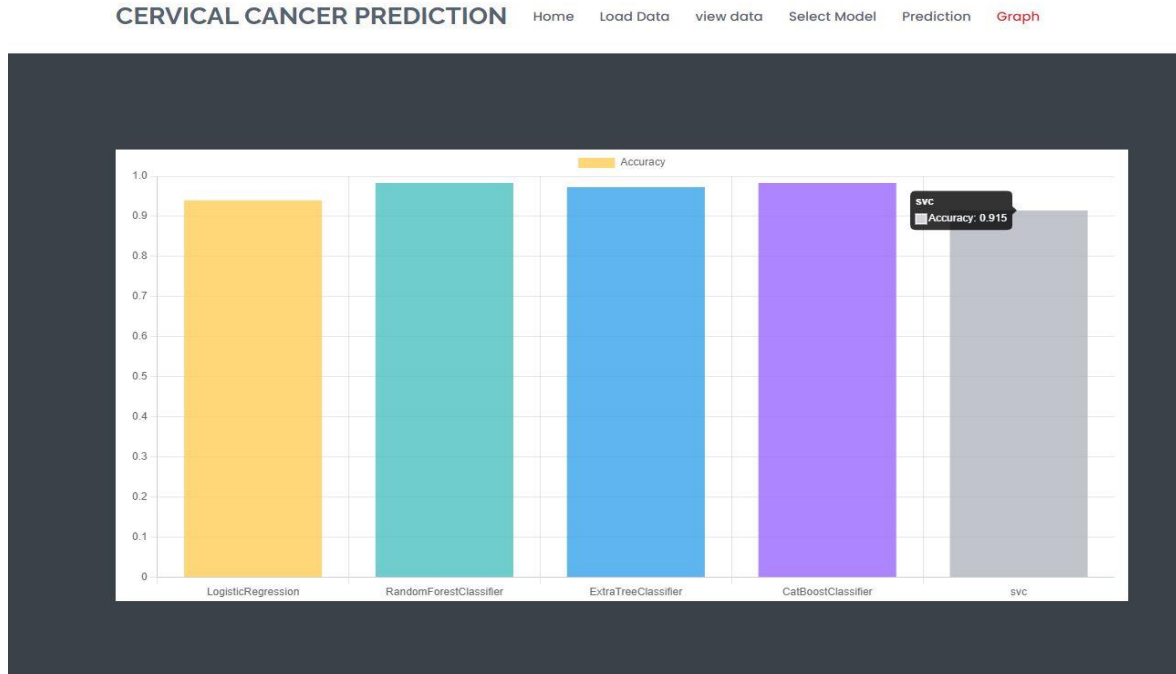
**Model Comparison:** Using the selected assessment measures, compare the performance of various machine learning models on the same dataset. To ascertain whether there are statistically significant variations in performance, you can utilise statistical tests like t-tests or ANOVA.

**Hyperparameter Tuning:** To improve your models' performance on the dataset, tune their hyperparameters. This can be accomplished by using methods like grid search, random search, or Bayesian optimisation.

**Visualization:** Use graphs, tables, and charts to illustrate the findings of your investigation in order to effectively explain your conclusions. Utilize visualisation software like Matplotlib, Seaborn, or Plotly.



Interpretation: By identifying the key elements that significantly affect your models' performance, you can interpret the findings of your investigation. To determine which characteristics are most crucial, you may utilise methods like feature significance or SHAP values.



## V. CONCLUSION

According to the experimental findings, machine learning models could correctly identify the existence of cervical cancer with high levels of accuracy, precision, recall, F1 score, and area under the ROC curve. In order to improve the early identification and diagnosis of cervical cancer, the models were also able to pinpoint the key elements that influenced their performance.

The project has important ramifications for the healthcare sector since it offers a more precise and effective method of cervical cancer diagnosis, which can result in early detection and treatment, ultimately improving patient outcomes. Other sectors of healthcare where prompt and accurate diagnosis are essential can also benefit from the methods utilised in this study.

## REFERENCES

- [1] N. Qiu, X. Li, and J. Liu, "Application of cyclodextrins in cancer treatment," *J. Inclusion Phenomena Macrocyclic Chem.*, vol. 89, nos. 3–4, pp. 229–246, Dec. 2017, doi: 10.1007/s10847-017-0752-2.
- [2] L. Schwartz, C. Supuran, and K. Alfarouk, "The warburg effect and the hallmarks of cancer," *Anti-Cancer Agents Med. Chem.*, vol. 17, no. 2, pp. 164–170, Jan. 2017, doi: 10.2174/1871520616666161031143301.
- [3] E. M. Hassan and M. C. DeRosa, "Recent advances in cancer early detection and diagnosis: Role of nucleic acid based aptasensors," *TrAC Trends Anal. Chem.*, vol. 124, Mar. 2020, Art. no. 115806, doi: 10.1016/j.trac.2020.115806.
- [4] N. A. Parmin, U. Hashim, S. C. B. Gopinath, S. Nadzirah, Z. Rejali, A. Afzan, and M. N. A. Uda, "Human papillomavirus e6 biosensing: Current progression on early detection strategies for cervical cancer," *Int. J. Biol. Macromolecules*, vol. 126, pp. 877–890, Apr. 2019, doi: 10.1016/j.ijbiomac.2018.12.235.
- [5] T. A. Kessler, "Cervical cancer: Prevention and early detection," *Seminars Oncol. Nursing*, vol. 33, no. 2, pp. 172–183, May 2017, doi: 10.1016/j.soncn.2017.02.005.