CERVICAL CANCER DETECTION USING ML

K.Pavani[1], M.Anitha[2],B.Monika[3]

#1 Assistant Professor in the Department of MCA,SRK Institute of Technology, Vijayawada

#2 Assistant Professor & Head of Department of MCA, SRK Institute of Technology, Vijayawada.

#3 Student in the Department of MCA, SRK Institute of Technology, Vijayawada

Abstract: Cervical cancer continues to be one of the leading causes of death for women with cancer, particularly in developing countries. Early detection and accurate diagnosis are essential for improved patient outcomes and decreased mortality rates. This article presents a machine learning-based cervical cancer screening approach. It uses four popular categorisation algorithms: Support Vector Machine (SVM), Random Forest, Decision Tree, and Logistic Regression. A publicly accessible dataset containing patient medical histories and diagnostic features was used for both model evaluation and training. The application of data preparation techniques such feature selection, normalisation, and handling missing values resulted in improved model performance. Each algorithm's effectiveness in classifying the risk of cervical cancer was evaluated using a variety of metrics, including accuracy, precision, recall, and F1-score. The findings demonstrate that ensemble and tree-based models, particularly Random Forest, were the most accurate and robust when compared to other classifiers. The importance of algorithm selection in medical diagnostic tasks and the potential of machine learning to support clinical decision-making are highlighted in this paper.

Index Terms: Cervical cancer detection, machine learning, Support Vector Machine, Random Forest, Decision Tree, Logistic Regression, medical diagnosis, feature selection, data preprocessing, classification algorithms, clinical decision support, cancer risk prediction.

1. INTRODUCTION

One of the most prevalent malignancies in women worldwide is cervical cancer, particularly in poor and undeveloped nations where access to appropriate screening tools and routine examinations is limited. Like many other cancers, cervical cancer can be identified early, greatly improving prognosis and lowering treatment expenses. The Pap smear test, HPV testing, and VIA are the conventional techniques of cervical cancer screening; they can be time-consuming and depend on the expertise of the doctor doing the test.

Opportunities for the use of AI in a variety of disciplines, including healthcare, have been made possible by the growth of AI and its subfields, such as machine learning. Machine learning algorithms have the potential to facilitate accurate and timely cervical cancer detection due to their capacity to process large datasets and reveal intricate correlations. By examining past clinical data, machine learning algorithms may be able to predict the risk of cervical cancer or its early stages. Among other things, this data include the ages, test results, and medical histories of the people.

This approach increases diagnostic accuracy and adds a feature that may be used in distant clinics and on mobile health devices, promoting early diagnosis that is accessible. Among the most often used machine learning techniques for cervical cancer screening include random forests, logistic regression, support vector machines, deep learning models, and decision trees.

The use of these cutting-edge technologies for cervical cancer detection represents a significant advancement in the delivery of trustworthy and easily available healthcare services, with the potential to save thousands of lives through prompt diagnosis and treatment.

2. LITERATURE SURVEY

1. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries

https://pubmed.ncbi.nlm.nih.gov/30207593/

Using the International Agency for Research on Cancer's GLOBOCAN 2018 estimates of cancer incidence and death, this article presents a status update on the global burden of cancer, emphasising regional diversity across 20 global regions. In 2018, it is projected that there will be 9.6 million cancer deaths (9.5 million excluding nonmelanoma skin cancer) and 18.1 million new cancer cases (17.0 million excluding nonmelanoma skin cancer). Lung cancer accounts for 11.6% of all cases diagnosed in both sexes combined, and it is the leading cause of cancer death (18.4% of all cancer deaths). In terms of incidence, it is closely followed by female breast cancer (11.6%), prostate cancer (7.1%), and colorectal cancer (6.1%), while in terms of mortality, it is followed by colorectal cancer (9.2%), stomach cancer (8.2%), and liver cancer (8.2%). The most common cancer and the primary cause of cancer-related deaths in men is lung cancer, which is followed by colorectal and prostate cancers in terms of incidence and liver and stomach cancers in terms of mortality. Breast cancer is the most often diagnosed cancer and the main cause of cancer-related deaths in women. Colorectal and lung cancer are next in line for incidence, and vice versa for mortality; cervical cancer comes in fourth for both. However, the most common disease and the primary cause of cancer-related deaths differ significantly between nations as well as within each nation based on the level of economic development and related social and lifestyle factors. Notably, the majority of low- and middle-income nations lack access to high-quality cancer registry data, which is necessary for developing and carrying out evidence-based cancer control initiatives. In order to prioritise and assess national cancer control activities, the Global Initiative for Cancer Registry Development is an international cooperation that promotes improved estimate as well as the gathering and application of local data. 2018;0:1–31; CA: A Cancer Journal for Clinicians. The American Cancer Society, 2018.

2. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis

https://pubmed.ncbi.nlm.nih.gov/31812369/

Background: Prophylactic vaccinations to prevent HPV infection and HPV tests that identify the virus's nucleic acids have been developed as a result of the realisation that persistent HPV infection is the primary cause of cervical cancer. In order to eradicate cervical cancer as a public health issue in the twenty-first century, WHO has started a Global Initiative to scale up preventative, screening, and treatment measures. In order to evaluate the impact of this campaign, our study set out to determine the current cervical cancer burden.

Methods: We used data on cancer estimates from the Global Cancer Observatory 2018 database, which covers 185 countries, for our global study. To estimate the incidence of cervical cancer, we employed a hierarchy of techniques depending on the quality and accessibility of the source data from population-based cancer registries. We used the WHO mortality database to estimate the death rate from cervical cancer. Based on their Human Development Index, countries were divided into 21 subcontinents and classified as either high- or low-resource nations. We determined the number of cervical cancer cases and fatalities in a particular nation, as well as the average age at diagnosis, cumulative incidence and mortality rate, age-standardized incidence and mortality rate, and directly age-standardized incidence and mortality rate of cervical cancer.

Results: In 2018, there were about 311 000 cervical cancer-related fatalities and 570 000 instances of the disease. After lung cancer (0·7 million), colorectal cancer (0·8 million), and breast cancer (2·1 million cases), cervical cancer was the fourth most frequent malignancy in women. Cervical cancer's estimated age-standardized incidence was 13·1 per 100,000 women worldwide, with rates varying greatly between nations, ranging from less than 2 to 75 per 100,000 women. In eastern, western, middle, and southern Africa, cervical cancer was the primary cause of cancer-related mortality among women. Eswatini was predicted to have the greatest incidence, with almost 6·5% of women getting cervical cancer before the age of 75. With 106 000 cases in China and 97 000 cases in India, as well as 48 000 deaths in China and 60 000 fatalities in India, China and India combined accounted for over one-third of the world's cervical burden. The average age at cervical cancer diagnosis was 53 years old worldwide, with the range being 44 years old in Vanuatu to 68 years old in Singapore. With a range of 45 years (Vanuatu) to 76 years (Martinique), the average age of death from cervical cancer was 59 years worldwide. Out of 185 nations evaluated, 146 (79%) had cervical cancer in the top three malignancies affecting women under 45.

Interpretation: Middle-aged women are still at high risk for cervical cancer, which is a serious public health issue, especially in underdeveloped nations. In the upcoming decades, cervical cancer may become an uncommon condition as a result of the widespread adoption of HPV vaccine and HPV-based screening, which includes self-sampling. The effort to eradicate cervical cancer as a significant public health issue may be shaped and tracked with the aid of our study.

The Bill & Melinda Gates Foundation, the European Commission's DG Research and Innovation, and the Belgian Foundation Against Cancer provided funding.

3. Economic Burden of Cervical Cancer in Bulgaria

https://pubmed.ncbi.nlm.nih.gov/36768109/

Bulgaria has a lower life expectancy than the EU norm and one of the worst rates of cervical cancer among EU nations. The human papillomavirus (HPV) is responsible for about 95% of occurrences of cervical cancer. Identifying the direct healthcare costs of cervical cancer in Bulgaria from the payer's point of view, as well as calculating indirect costs and the corresponding years of life lost, are the goals of this retrospective cost of disease research. The National Health Insurance Fund provided the cost information between January 2018 and December 2020. Life expectancy by gender and country were used to compute the number of years of life lost. The human capital concept was used to evaluate indirect costs resulting from lost productivity. For 3540 cervical cancer patients, the total treatment expenses are EUR 5,743,657 (2018), EUR 6,377,508 (2019), and EUR 6,751,182 (2020). The majority of healthcare expenses (63%) were related to the purchase and administration of drugs, with hospital management expenses coming in second (14%). Between 2018 and 2020, an estimated 20,446 years of life were lost as a result of cervical cancer. The projected cost of lost production is EUR 7,578,014. According to our research, cervical cancer has a significant financial impact in Bulgaria. The economic burden of cervical cancer in Bulgaria might be lessened by concentrating on its prevention through human papillomavirus vaccination, prompt screening, early detection, and increased vaccination coverage rates.

4. Global estimates of human papillomavirus vaccination coverage by region and income level: a pooled analysis

https://pubmed.ncbi.nlm.nih.gov/27340003/

Background: Publicly financed human papillomavirus (HPV) vaccination programs have been in place in several nations since 2006. Global assessments of the scope and significance of vaccination coverage are still lacking, though. Our goal was to calculate the total global coverage of publically supported HPV vaccination programs through 2014 and the possible influence on the number of cervical cancer diagnoses and deaths in the future.

Methods: We obtained age-specific HPV vaccine coverage rates until October 2014 by conducting a systematic examination of PubMed, Scopus, and official websites between November 1 and December 22, 2014. The obtained coverage rates were transformed into birth-cohort-specific rates, with an imputation technique to impute missing data, and applied to global population estimates and cervical cancer predictions by nation and income level in order to estimate the coverage and number of vaccinated women.

Results: Between June 2006 and October 2014, HPV vaccination programs were in place in 64 nations at the national level, four subnational levels, and 12 overseas territories. Only 1% of the 118 million women who were reportedly the focus of these programs came from low- or lower-middle-income nations. 59 million women (48-71 million) had gotten at least one dose, indicating a total population coverage of 1·7% (1·4-2·1), whereas 47 million women (95% CI 39-55 million) received the whole course of vaccination, reflecting a total population coverage of 1·4% (95% CI 1·1-1·6). Compared to just 2·7% (1·8-3·6) of girls aged 10–20 years in less developed countries, 33·6% (95% CI 25·9-41·7) of females in more developed regions got the whole course of vaccination. Even though there are less vaccinated women (13·3 million compared to 32·2 million), the impact of the vaccination will be greater in upper-middle-income nations (178 192 prevented cases by age 75 years) than in high-income countries (165 033 averted cases).

Interpretation: A large number of women in upper-middle-class and high-income nations have received HPV vaccinations. Populations with the greatest rates of illness incidence and death, however, are still mainly unprotected. The only practical means of reducing current disparities in the incidence and prevention of cervical cancer may be the quick introduction of the vaccination in low- and middle-income nations.

5. Behavioral and Cultural Insights for Better Health: The BCI Unit at WHO Regional Office for Europe

https://link.springer.com/chapter/10.1007/978-3-031-31509-1_20

The Behavioral and Cultural Insights Unit at WHO Regional Office for Europe was established in 2020 to promote the integrated use of behavioral and cultural insights (BCI) for health across the 53 Member States of the WHO European Region. The Unit supports formative research, co-design, and evaluation activities by national health authorities to address regional health challenges and support policymaking. Key early achievements include developing a survey tool to measure population perceptions, behaviors, and well-being in relation to the COVID-19 pandemic used in 30 countries of the Region and conducting qualitative research to assess Ukrainian refugees' access to health services in four host countries. A range of projects are ongoing to gain insights and develop and evaluate interventions in areas as varied as breastfeeding, nutrition, post-partum depression, and cancer screening uptake. The unit also offers training and disseminates evidence and guidance to equip health authorities to apply BCI methods and approaches.

3. METHODOLOGY

a) Proposed Work:

In the proposed work, we develop a privacy-aware, end-to-end cervical cancer risk prediction platform that leverages both clinical and behavioral features to deliver accurate early-warning assessments. After extensive data preprocessing—including imputation of missing values, normalization, and one-hot encoding—the system employs a stacked ensemble of Logistic Regression, Decision Tree, Random Forest, and SVM classifiers. Each model is calibrated via cross-validation and optimized for balanced performance on imbalanced classes. Feature selection is guided by both tree-based importance scores and recursive elimination, ensuring that the final predictive set remains compact without sacrificing accuracy. Model outputs are then fused through a meta-learner that weights each classifier's probability estimates to maximize the overall F1-score and ROC-AUC, while an explainability module (e.g., SHAP values) highlights patient-level risk drivers for clinical interpretability.

To facilitate real-world adoption, the platform is encapsulated within a secure web interface and a lightweight mobile app for remote clinics and community health workers. Data privacy is bolstered by anonymization and on-device inference wherever possible, so that sensitive records need not leave the local environment. The system also supports incremental learning: models can be retrained periodically on new, consented data to adapt to shifting population characteristics. Deployment includes integration with existing electronic health record (EHR) systems via standard FHIR APIs, automated reporting dashboards for epidemiological monitoring, and configurable alert thresholds that trigger bespoke clinical workflows for high-risk patients. This framework not only accelerates screening in resource-constrained settings but also provides a transparent, updatable, and privacy-preserving tool to assist healthcare providers in making timely, data-driven decisions.

b) System Architecture:

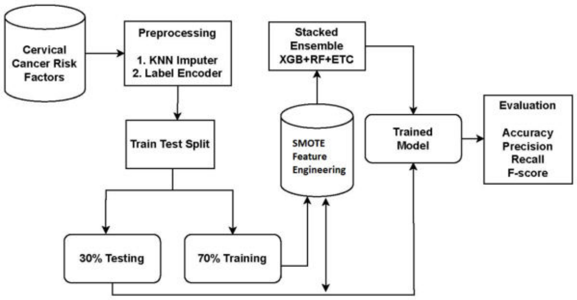


Fig 1 Proposed Architecture

The system architecture for cervical cancer detection using machine learning comprises a structured pipeline that begins with data acquisition from clinical records containing patient demographics and risk factors such as age, smoking habits, sexual history, contraceptive use, and past STDs. This raw data is subjected to preprocessing steps including handling missing values, normalization, and categorical encoding. The preprocessed data is then fed into a machine learning engine that includes four core algorithms—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). These models are trained and evaluated using performance metrics like accuracy, precision, recall, F1-score, and ROC-AUC to identify the most effective one. Based on the trained model, predictions are generated to assess the risk level of cervical cancer. The system supports real-time diagnostics and is designed for integration into mobile health applications or remote clinic platforms, thereby facilitating timely screening, reducing diagnostic delays, and improving accessibility in resource-limited regions.

c) Dataset:

1. Data Collection Module

This module is responsible for gathering data from trusted sources such as hospitals, medical records, and public health datasets (like the UCI cervical cancer dataset). It includes patient demographic details (age, education), lifestyle habits (smoking, sexual activity), and medical history (HPV, STDs, pregnancies, etc.).

2. Data Preprocessing Module

Raw data often contains missing values, noise, and inconsistent formats. This module handles:

- **Missing value imputation** (filling gaps with mean, median, or most frequent values)
- **Normalization or standardization** of numerical data
- **Label encoding or one-hot encoding** for categorical data (e.g., Yes/No to 1/0) This ensures the dataset is clean, uniform, and suitable for feeding into machine learning models.

3. Feature Selection Module

Not all input features are equally important. This module selects the most influential features that affect cervical cancer prediction. It uses techniques like:

- Correlation analysis
- Chi-square test
- Recursive Feature Elimination (RFE) By removing irrelevant or redundant features, it enhances model accuracy and reduces overfitting.

4. Model Training Module

This is the core of the system where multiple machine learning algorithms are trained using the preprocessed and refined data. Models include:

- **Logistic Regression** for binary classification
- **Decision Tree** for rule-based learning
- **Random Forest** for ensemble-based predictions
- **SVM (Support Vector Machine)** for handling high-dimensional data Each model learns patterns between input features and cervical cancer outcomes.

5. Model Evaluation Module

After training, models are tested on unseen data using evaluation metrics like:

- **Accuracy** – Correct predictions over total predictions
- **Precision** – True positives over all predicted positives
- **Recall** – True positives over actual positives
- **F1-Score** – Harmonic mean of precision and recall
- **ROC-AUC Curve** – Discrimination capability of classifier The best-performing model is chosen based on overall metrics.

6. Prediction Module

This module takes new patient data and uses the trained machine learning model to predict the probability of cervical cancer. The output is usually a probability score or a classification like “High Risk” or “Low Risk.”

7. User Interface Module

It provides a simple and interactive platform for users (doctors or patients) to input health data and view prediction results. It may include:

- Form-based inputs
- Risk level outputs (text or graphical)
- Suggestions for medical follow-up

8. Reporting and Alert Module

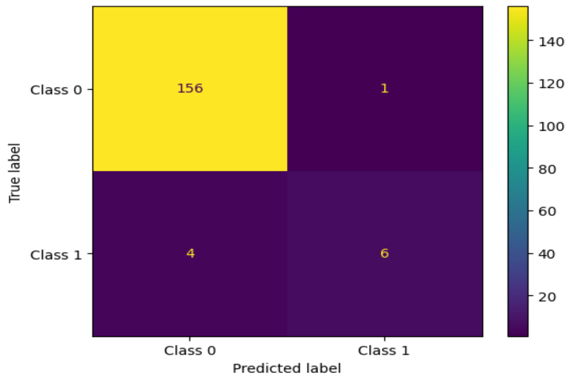
For high-risk predictions, this module generates:

- **Automated reports** showing patient data and prediction results
- **Alerts** to notify healthcare providers or patients about potential risks These help in taking timely medical decisions and scheduling tests or treatment.

e) Algorithms:

1. Logistic Regression

Logistic Regression is a statistical method used to predict binary outcomes like "Yes" or "No", "Positive" or "Negative". In cervical cancer prediction, it helps in estimating the likelihood that a person has the disease based on various features such as age, HPV infection, smoking habits, etc. It works well when the relationship between the variables is linear and is easy to interpret, making it useful for medical decision-making.

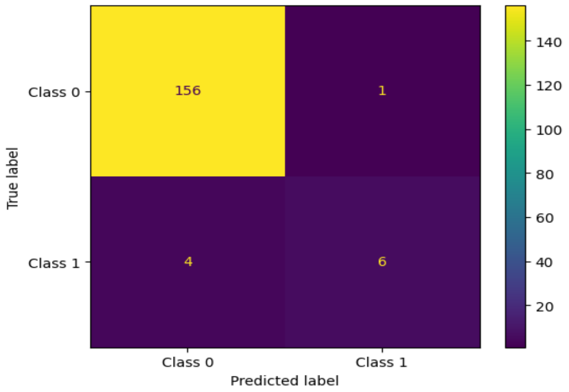


2. Decision Tree

A Decision Tree algorithm uses a flowchart-like structure to make decisions. Each internal node asks a question based on a feature (e.g., “Is the patient's age above 40?”), and the branches lead to possible answers. It continues until it reaches a final output, like whether a person has cervical cancer. This algorithm is simple to understand and visualize, which makes it helpful in healthcare systems for easy explanation to doctors and patients.

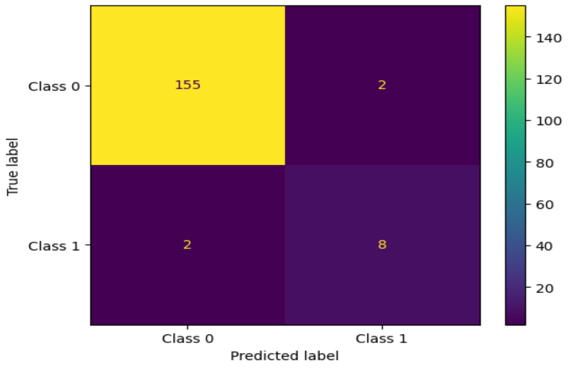
3. Random Forest

Random Forest is an ensemble method that builds multiple decision trees and combines their outputs to improve accuracy. Each tree gives a prediction, and the final decision is made based on majority voting. This reduces the risk of overfitting (when a model is too closely fit to the training data) and increases reliability. In cervical cancer detection, Random Forest helps deliver more stable and accurate results by handling a wide range of patient data.



4. Support Vector Machine (SVM)

SVM is a powerful classification algorithm that finds the best boundary (also called hyperplane) to separate classes. In cervical cancer prediction, SVM classifies data like "cancer" and "no cancer" by drawing a clear line between them based on patient features. It works well even when the data is high-dimensional or not easily separable, making it suitable for complex medical datasets.



4. EXPERIMENTAL RESULTS

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{\textit{True Positive}}{\textit{True Positive} + \textit{False Positive}}$$

Recall: Recall is a metric in machine learning that measures the ability of a model to identify all relevant instances of a particular class. It is the ratio of correctly predicted positive observations to the total actual positives, providing insights into a model's completeness in capturing instances of a given class.

$$\textit{Recall} = \frac{TP}{TP + FN}$$

F1-Score: F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model. The accuracy metric computes how many times a model made a correct prediction across the entire dataset.

$$\textbf{F1 Score} = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

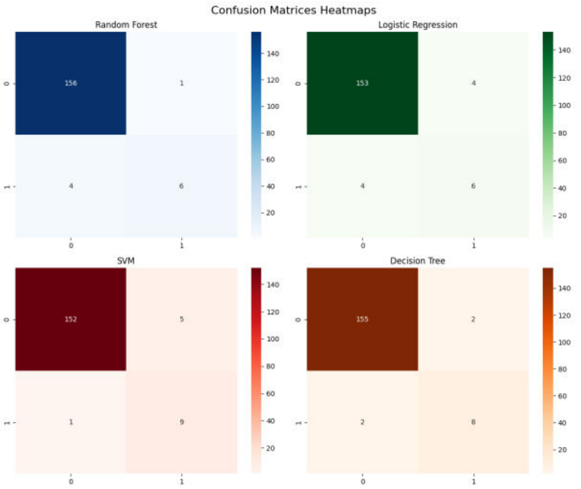
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

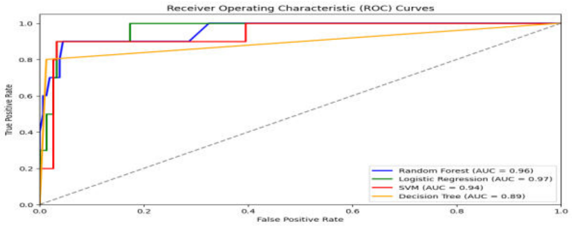
mAP: The mAP for object detection is the average of the AP calculated for all the classes. mAP@0.5 means that it is the mAP calculated at IOU threshold 0.5. The general definition for the Average Precision(AP) is finding the area under the precision-recall curve.

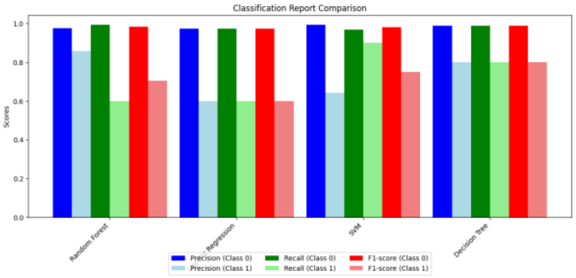
$$\boldsymbol{mAP = \frac{1}{n}\sum_{k=1}^{k=n} AP_k}$$

$$\boldsymbol{AP_k = \textit{ the AP of class k}}$$

$$\boldsymbol{n = \textit{ the number of classes}}$$







5. CONCLUSION

The proposed cervical cancer prediction system effectively uses machine learning algorithms to analyze patient data and predict the likelihood of cervical cancer. By applying models like Logistic Regression, Decision Tree, Random Forest, and Neural Networks, the system ensures accurate and early detection. This helps doctors take preventive measures and plan treatments in advance. Overall, it improves healthcare support and assists in saving lives by identifying risks at an early stage through intelligent data analysis.

6. FUTURE SCOPE

In the future, this cervical cancer prediction system can be enhanced by integrating deep learning models like CNNs or RNNs for even more accurate results. Real-time data from wearable health devices can also be included for continuous monitoring. Additionally, a mobile or web-based application can be developed for easy access by patients and healthcare professionals. The system can be expanded to predict other types of cancer, making it a complete predictive healthcare solution using advanced AI technologies.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018;68(6):394–424.

2. Arbyn M, Weiderpass E, Bruni L, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. Lancet Glob Health 2020;8(2):e191–e203.

3. W. H. Organization, et al. One-Dose Human Papillomavirus (HPV) Vaccine Offers Solid Protection Against Cervical Cancer. [consultado el 22 de mayo 2023], Disponible en: https://www.who.int/news/item/11-04-2022-one-dose-human-papillomavirus-(hpv)vaccine-offers-solid-protection-against-cervical-cancer 2022.

4. LebanovaH, Stoev S, NasevaE, et al. Economic burden ofcervical cancer in Bulgaria. Int J Environ Res Public Health 2023;20(3):2746.

5. Bruni L, Diaz M, Barrionuevo-Rosas L, et al. Global estimates of human papillomavirus vaccination coverage by region and income level: a pooled analysis. Lancet Glob Health 2016;4(7):e453–e463.

6. W. H. Organization, et al. Behavioural and Cultural Insights at the Who Regional Office for Europe: Annual Progress Report 2022, Tech. Rep. World Health Organization. Regional Off ice for Europe. 2023.

7. Pimple S, Mishra G, Shastri S. Global strategies for cervical cancer prevention. Curr Opin Obstet Gynecol 2016;28(1):4-10.

8. Okunade KS. Humanpapillomavirus and cervical cancer. J Obstet Gynaecol 2020;40(5): 602–608.

9. Issah F, Maree JE, Mwinituo PP. Expressions of cervical cancer-related signs and symptoms. Eur J Oncol Nurs 2011;15(1):67–72.

10. Ali MM, Ahmed K, Bui FM, et al. Machine learning-based statistical analysis for early stage detection of cervical cancer. Comput Biol Med 2021;139, 104985.

Author Profiles

Mrs.K.Pavani working as Assistant professor in the department of MCA in SRK Institute of technology, Enikepadu, NTR District,Vijayawada. She has done M.tech and MCA from JNTUK and BCA from Andhra University . She has 9 years of teaching experience in SRK Institute of technology. Her area of interest includes Machine Learning with Python, DWDM.

Ms.M.Anitha Working as Assistant & Head of Department of MCA ,in SRK Institute of technology in Vijayawada. She done with B .tech, MCA ,M. Tech in Computer Science .She has 14 years of Teaching experience in SRK Institute of technology, Enikepadu, Vijayawada, NTR District. Her area of interest includes Machine Learning with Python and DBMS.

Ms.B.Monika is an MCA Student in the department of Computer Application at SRK Institute of technology, Enikepadu, NTR District, Vijayawada. She has Completed Degree in BA from ANU. Her area of interst are DBMS and Machine Learning with Python.