Cervical Cancer Prediction Using Machine Learning

Sana G, Mr. Vinay Patel GL

[1] *Student,4th Semester MCA, Department of MCA, BIET, Davangere*

[2]*Assistant Professor, Department of MCA, BIET, Davangere*

ABSTRACT

Cervical cancer is one of the deadliest diseases among women globally, particularly in low-income countries where early detection and treatment are limited. This project proposes a Machine Learning-based Prediction System to detect the risk of cervical cancer at an early stage. By analyzing various factors such as age, sexual health, hormonal usage, genetic history, and prior diagnoses, the system applies classification techniques to estimate the risk category. A Decision Tree classifier is employed to model patterns and identify frequent combinations of high-risk attributes. The system predicts the likelihood and stage of cervical cancer by assigning risk scores based on patient inputs and classifying them into risk levels: low, intermediate, high, and very high. This automated approach supports medical practitioners by improving diagnostic accuracy and enabling timely intervention. The project is built using Python, MySQL, and Anaconda, offering a scalable and interpretable diagnostic tool for early-stage cervical cancer prediction.

Keywords: Sign language, information retrieval, computer vision, natural language processing, accessibility, deaf individuals.

I. INTRODUCTION

With the rise in healthcare data availability, machine learning (ML) has emerged as a valuable tool for disease prediction and personalized medicine. Cervical cancer, primarily caused by the Human Papillomavirus (HPV), often develops without early symptoms and is detected only in advanced stages.

This project uses ML to predict the probability and stage of cervical cancer based on clinical and lifestyle inputs. The system processes datasets containing attributes like sexual activity, number of pregnancies, smoking habits, and medical history. A Decision Tree-based classifier is used to determine cancer risk based on frequent patterns and assign patients to a risk category. This technology-driven approach has the potential to improve detection, reduce manual workload, and guide early medical decisions.

II. LITERATURE SURVEY

Cervical cancer, epidemiologically similar to a low-infectious venereal illness, is the world's largest cause of cancer-related death among women. There are several risk factors for cervical cancer that are associated with HPV exposure. The implementation of screening also had an impact on the considerable variations in incidence among nations. Liquid-based cytology (LBC), visual inspection with acetic acid and a typical Pap smear, and HPV testing are the main screening techniques. Other techniques, like the Pap smear, aren't always precise, though. As a result, women and the general public need to be made more aware of the importance of taking preventative measures. Cervical cancer could be eliminated by humans first [1].

HN Research and a case study have been proposed by

Harsha Kumar et al[2]. The purpose of this case study is to ascertain the degree of cervical cancer awareness among Indian women. A large percentage of cervical cancer cases have poor outcomes because the disease is discovered too late. After being chosen at random, 83 Mangalore City women were questioned about cervical cancer. Only 7–8% of the general public is aware of cervical cancer and how to get screened for it, according to the statistics. Eighty-five percent of the women were unaware of the screening procedure, and eighty-nine percent were just vaguely aware of this cancer. The study comes to the conclusion that women and doctors need to be made aware of the dangers of cervical cancer immediately.

The study by Riham Alsmariy et al. used the cervical ca ncer dataset from UCI. They improved the model's perfo rmance using voting classification, SMOTE, PCA, and s tratified 10fold crossvalidation, and it performed excepti onally well in the Schiller test across several metris[3]. According to Matko Glučina's research, early cervical c ancer detection is much improved by combining multila yer perceptrons and Knearest neighbors with over sampl ing techniques like SMOTEEN and SMOTETOMEK.T he approach produced mean AUC and MCC values of a bove 0.95 acrossall diagnostic techniques, suggesting its efficacy in the early detection of cervical cancer[4]. The early identification of cervical cancer was the main emphasis of Sohely Jahan et al.The study predicted the l ikelihood of cervical cancer using a variety of machine l earning classification techniques based on risk factors.T he study evaluates how well various classification algori thms perform based on the top five criteria.The MultiLa yer Perceptron model consistently outperformed other m ethods when the models were evaluated for accuracy, pr ecision, and recall.Furthermore, the study considers man y datasetsplitting ratios, unlike previous works that focu sed on only one[5].

2.1 EXISTING SYSTEM

Conventional systems rely heavily on physical diagnostic procedures such as cytology, biopsy, and imaging, which are:

- Time-consuming and costly
- Not scalable to large populations
- Limited by lack of early symptoms

Prone to human error and misdiagnosis

2.2 PROBLEM SYSTEM

Cervical cancer contributes significantly to female mortality in low-income regions due to late diagnosis. Existing systems do not efficiently detect cancer in its early stage. There is a need for a smart, ML-based diagnostic tool that can process complex medical inputs and produce early, reliable predictions for cervical cancer detection.

2.3 PROPOSED SYSTEM

The proposed system:

- Collects input attributes like age, sexual history, contraceptive usage, and HPV status
- Preprocesses data and extracts significant features
- Applies a **Decision Tree classifier** to identify patterns and classify risk levels
- Categorizes risk into four levels (low, intermediate, high, very high)
- Outputs predicted cancer stage and suggests further testing where needed

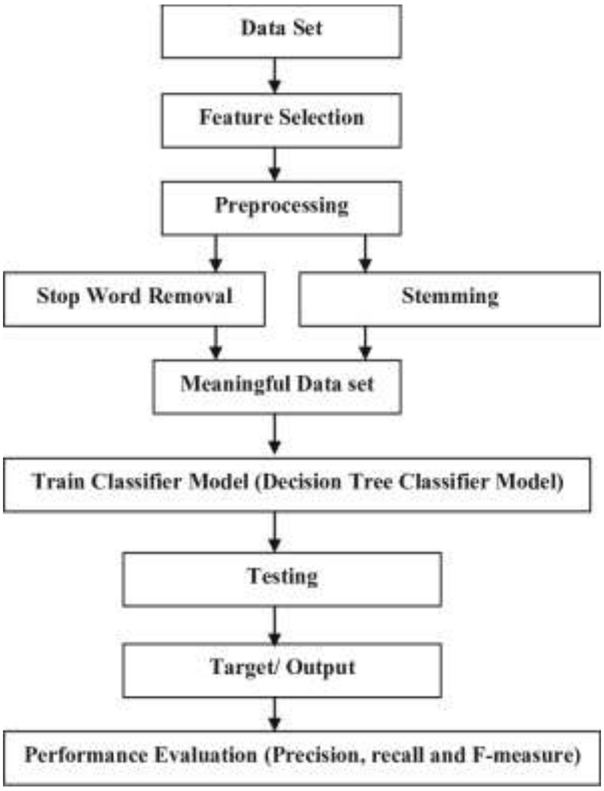
It enhances early detection and supports doctors in prioritizing high-risk cases using data-driven decisions.

III. SYSTEM REQUIREMENT SYSTEM

3.1 FUNCTIONAL REQUIREMENTS

- GUI for patient data input and result display
- Preprocessing and transformation of clinical data
- Decision Tree-based cancer stage prediction
- Risk scoring system for patient classification
- Backend database for test and result storage
- Report suggestion for further testing based on risk level

3.2 ARCHITECTURE DIAGRAM


Data Set
Feature Selection
Preprocessing
Stop Word Removal
Stemming
Meaningful Data set
Train Classifier Model (Decision Tree Classifier Model)
Testing
Target/ Output
Performance Evaluation (Precision, recall and F-measure)

3.3 ARCHITECTURE OVERVIEW

1. **Data Collection**: Structured dataset containing medical and behavioral attributes
2. **Preprocessing**: Cleans missing values, encodes categorical data, normalizes numeric fields
3. **Partitioning**: Splits dataset into training and testing sets
4. **Modeling**: Uses Decision Tree algorithm for classification
5. **Feature Extraction**: Identifies high-impact features (e.g., age, contraceptive use)
6. **Prediction Engine**: Calculates and classifies risk score into one of four levels

Output Interface: Displays prediction with advice on further diagnostic steps

3.4 IMPLEMENTATION

The recommendation system was implemented using Python, leveraging its extensive libraries for data processing and machine learning. Initially, the dataset containing user preferences and item details was collected and thoroughly preprocessed by removing duplicates, handling missing values, and encoding categorical variables where necessary. Once the data was cleaned, feature extraction techniques were applied to identify meaningful patterns. For building the recommendation model, both content-based filtering and collaborative filtering approaches were explored. Content-based filtering analyzed the item attributes to suggest similar products, while collaborative filtering utilized user-item interaction data to predict preferences by identifying similarities between users or items. The model’s performance was assessed using evaluation metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), ensuring its accuracy and reliability. Finally, the system was integrated into a simple user interface, allowing users to receive personalized and accurate recommendations based on their historical interactions and preferences.

IV. CONCLUSION

This ML-based cervical cancer prediction system offers a cost-effective, interpretable, and efficient tool for early detection. It supports healthcare professionals in identifying high-risk patients based on data-driven models. Using decision trees enables both accuracy and explainability, essential for gaining trust in clinical environments.

By combining patient-reported and clinical inputs with predictive analytics, this system can be a powerful addition to cervical cancer screening programs, especially in resource-limited settings.

V. REFERENCES

1. Zhang S, Xu H, Zhang L, Qiao Y. Cervical cancer: Epidemiology, risk factors and screening. Chin J Cancer Res. 2020 Dec 31;32(6):720-728. doi: 10.21147/j.issn.1000-9604.2020.06.05. PMID: 33446995; PMCID: PMC7797226.

2. Harsha Kumar H, Tanya S. A Study on Knowledge and Screening for Cervical Cancer among Women in Mangalore City. Ann Med Health Sci Res. 2014 Sep;4(5):751-6. doi: 10.4103/2141-9248.141547. PMID: 25328788; PMCID: PMC4199169.

3. Riham Alsmariy, Graham Healy and Hoda Abdelhafez, “Predicting Cervical Cancer using Machine Learning Methods” International Journal of Advanced Computer Science and Applications(IJACSA), 11(7), 2020.

4. Glučina, Matko, Ariana Lorencin, Nikola Anđelić, and Ivan Lorencin. 2023. "Cervical Cancer Diagnostics Using Machine Learning Algorithms and Class Balancing Techniques" Applied Sciences 13, no. 2: 1061.

5. J. Electrical Systems 20-1s (2024): 944-955 Jahan, Sohely & Islam, Manowarul & Islam, Linta & Rashme, Tamanna & Prova, Ayesha & Paul, Bikash Kumar & Islam, M. & Mosharof, Mohammed. (2021). Automated invasive cervical cancer disease detection at early stage through suitable machine learning model. SN Applied Sciences. 3. 10.1007/s42452-021-04786-z.