A FIELD PROJECT REPORT

on

**“Cervical Cancer Prediction”**

**Submitted**

by

|  |  |
| --- | --- |
| 221FA04233  Padma Sai | 221FA04234  Naga Surya Charan |

|  |
| --- |
| 221FA04452  Yaga Sahithi  221FA04482  Siva Harshitha  221FA04523  Navya |

**Under the guidance of**

*Dr. S. Deva Kumar*

*Assistant Professor*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed to be UNIVERSITY**

**Vadlamudi, Guntur.**

**ANDHRA PRADESH, INDIA, PIN-522213.**



**CERTIFICATE**

This is to certify that the Field Project entitled **“Cervical Cancer Prediction”** that is being submitted by 221FA04233 (Padma Sai), 221FA04234 (Naga Surya Charan), 221FA04452 (Yaga Sahithi), 221FA04482 (Siva Harshitha), and 221FA04523 (Navya)for partial fulfilment of Field Project is a bonafide work carried out under the supervision of Dr.S.Deva Kumar, Assistant Professor, Department of CSE.

|  |  |  |
| --- | --- | --- |
| Guide name& Signature |  | Dr.K.V. Krishna Kishore |
| Assistant/Associate/Professor, CSE | HOD,CSE | Dean, SoCI |



**DECLARATION**

We hereby declare that the Field Project entitled **“Cervical cancer prediction ”** is being submitted by 221FA04233 (Padma Sai), 221FA04234 (Naga Surya Charan), 221FA04452 (Yaga Sahithi), 221FA04482 (Siva Harshitha), and 221FA04523 (Navya) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Dr.S.Deva Kumar, Assistant Professor, Department of CSE.

By

**221FA04233 (Padma Sai),**

**221FA04234 (Naga Surya Charan),**

**221FA04452 (Yaga Sahithi),**

**221FA04482 (Siva Harshitha),**

**221FA04523 (Navya)**

Date:

## ABSTRACT

Cervical cancer is a significant global health issue, particularly in developing countries where screening and early detection programs are less accessible. Machine learning models have emerged as powerful tools to assist in the early detection and prediction of cervical cancer, leveraging data from medical records, cytology tests, and various biomarkers. By analyzing large datasets, these models aim to identify patterns that can predict the onset of cervical cancer, enabling early intervention and potentially saving lives.

The predictive process typically involves the use of classification algorithms like Logistic Regression, Decision Trees, Support Vector Machines (SVM), and Neural Networks. These models are trained on historical patient data, including features such as age, sexual history, smoking habits, and human papillomavirus (HPV) infection status. Once trained, the models can predict whether a patient is at high risk for developing cervical cancer based on new input data, improving accuracy compared to traditional diagnostic methods.

Feature selection and preprocessing are critical aspects of building effective cervical cancer prediction models. By carefully selecting the most relevant features from the data, machine learning algorithms can enhance their predictive accuracy. Common techniques such as Principal Component Analysis (PCA) are used to reduce dimensionality and remove noise from the dataset, making the model more efficient and interpretable. Cross-validation and performance metrics like accuracy, precision, recall, and F1 score are used to evaluate the model's effectiveness.

Overall, machine learning-based cervical cancer prediction provides a non-invasive, data-driven approach to improving cancer screening and early detection. By integrating such models into healthcare systems, medical professionals can be better equipped to make informed decisions, allocate resources efficiently, and ultimately improve patient outcomes

**TABLE OF CONTENTS**

1. [Introduction 1](#_bookmark0)
   1. [What is cervical cancer and what causes it? 2](#_bookmark1)
   2. [The consequences of cervical cancer 3](#_bookmark2)
   3. [The economic and environmental effects of cervical cancer 4](#_bookmark3)
   4. [Current methodologies for cervical cancer detection 5](#_bookmark4)
   5. Applications of machine learning to combat cervical cancer 6
2. [Literature Survey 7](#_bookmark5)
   1. [Literature review 8](#_bookmark6)
   2. [Motivation 10](#_bookmark7)
3. [Proposed System 11](#_bookmark8)
   1. [Input dataset 13](#_bookmark9)
      1. Detailed features of dataset 13
   2. [Data Pre-processing 14](#_bookmark10)
      1. [Missing values 14](#_bookmark11)
         1. [Parameters of the fillna method 15](#_bookmark12)
      2. [Data Encoding 16](#_bookmark13)
   3. [Model Building 17](#_bookmark14)
      1. [MLP Algorithm 18](#_bookmark15)
   4. [Methodology of the system 20](#_bookmark16)
   5. [Model Evaluation 21](#_bookmark17)
   6. [Constraints 23](#_bookmark18)
   7. Cost and Sustainability Impact 24
4. [Implementation 26](#_bookmark19)
   1. Environment Setup 30
   2. Sample code for preprocessing and MLP operations 31
5. Experimentation and Result Analysis 33
6. Conclusion 41
7. References 43

**LIST OF FIGURES**

|  |  |
| --- | --- |
| Figure 1. Confusion matrix and ROC curve for Gaussian Naive Bayes decision tree | 17 |
| Figure 2. Confusion matrix and ROC curve for XGB Classifier | 18 |
| Figure 3. Confusion matrix and ROC curve for Random Forest | 18 |
| Figure 4. Confusion matrix and ROC curve for Naïve Bias & XGB Boost | 19 |
| Figure 5. Confusion matrix and ROC curve for XGB Boost & Random Forest | 18 |
| Figure 5. Confusion matrix and ROC curve for XGB Boost & Random Forest | 19 |

**LIST OF TABLES**

Table 1. Detailed features of dataset 36

Table 2. Accuracy evaluation 37

# CHAPTER-1 INTRODUCTION

### INTRODUCTION

**1.1** **What is cervical cancer and what causes it?**

Cervical cancer is a type of cancer that begins in the cells lining the cervix, the lower part of the uterus that connects to the vagina. The primary cause is long-term infection with certain high-risk strains of the human papillomavirus (HPV), a sexually transmitted virus. Though most HPV infections resolve on their own, persistent infections can lead to cellular changes, potentially resulting in cervical cancer. There are over 100 types of HPV, but only a few strains are considered high-risk for causing cancer.

Aside from HPV, other risk factors contribute to cervical cancer development, such as smoking, which weakens the immune system and damages DNA in cervical cells. Women with compromised immune systems, such as those with HIV, are also at increased risk. Additionally, long-term use of oral contraceptives and giving birth to many children are known risk factors.

Cervical cancer often progresses slowly, beginning as pre-cancerous changes in the cervix's cells (dysplasia). These changes can be detected early through screening tests like Pap smears. If caught early, dysplasia is easily treatable, preventing progression to invasive cancer. This slow progression provides an opportunity for early intervention and treatment if appropriate screening is in place.

The role of HPV vaccination has been crucial in preventing cervical cancer. Vaccines that protect against the most common high-risk HPV strains have significantly reduced the incidence of cervical cancer in populations with high vaccination rates. However, global disparities in vaccination and screening availability contribute to higher cervical cancer rates in low-resource settings.

In summary, cervical cancer's development is multifactorial, with HPV infection as the primary cause. Effective screening and vaccination programs have been instrumental in reducing cases, but access remains a challenge in many regions

**1.2 The consequences of cervical cancer**

If not detected early, cervical cancer can have devastating consequences. The cancer can spread beyond the cervix to nearby tissues such as the vagina, pelvic walls, or even distant organs like the lungs and liver. This process, known as metastasis, significantly reduces the chances of successful treatment and recovery. Early-stage cervical cancer has high cure rates, but late-stage cancer is far more challenging to treat and often fatal.

The disease also imposes a psychological toll on patients and their families. The emotional stress of diagnosis, treatment, and the impact on reproductive health can be overwhelming. Late-stage cancer often leads to high medical costs and may require intensive treatments like chemotherapy or radiation, further straining resources and leading to lost productivity.

Women diagnosed with advanced cervical cancer often undergo aggressive treatments, including surgery, chemotherapy, and radiation. These treatments can be physically taxing, causing side effects like fatigue, pain, and nausea, further reducing the patient's quality of life. In many cases, these treatments can result in infertility, which can have emotional and psychological impacts, especially for younger women who wish to have children.

The burden of cervical cancer is disproportionately high in low-income and middle-income countries. Lack of access to regular screening and healthcare facilities means that many women are diagnosed at later stages when treatment options are limited. In contrast, countries with comprehensive screening programs have seen significant reductions in cervical cancer incidence and mortality rates.

Survivors of cervical cancer may face long-term health issues, including chronic pain, emotional distress, and difficulties related to reproductive health. The emotional impact of cancer diagnosis and treatment can lead to anxiety, depression, and feelings of isolation. For many, dealing with the uncertainty of recurrence or long-term effects from treatment poses additional mental health challenges.

Thus, the consequences of cervical cancer extend beyond the physical. The disease affects the emotional, psychological, and social well-being of individuals, their families, and communities. Addressing cervical cancer requires not only medical treatment but also comprehensive psychosocial support.

**1.3** **The economic and environmental effects of cervical cancer**

Cervical cancer places a significant economic burden on healthcare systems and individuals. For patients, costs associated with diagnosis, treatment, and post-treatment care can be overwhelming. In low-resource settings, where healthcare costs are often paid out-of-pocket, these financial burdens can push families into poverty. Even in high-income countries, where healthcare coverage is more comprehensive, indirect costs such as loss of income due to illness or caregiving responsibilities can be substantial.

At a societal level, cervical cancer leads to loss of productivity as women in their prime working years face illness or die prematurely. The ripple effects of this loss are felt in households, communities, and national economies. Women play a critical role in their families and workplaces, and their loss due to preventable conditions like cervical cancer can have lasting socio-economic impacts.

From an environmental perspective, the treatment of cervical cancer contributes to healthcare-associated waste, particularly in resource-intensive therapies like chemotherapy and radiation. Medical waste disposal is a growing environmental concern, and the long-term sustainability of healthcare systems must balance the need for effective cancer treatments with minimizing environmental harm.

Preventive measures, such as vaccination and early detection, provide a cost-effective solution to cervical cancer, reducing both the economic and environmental burdens of the disease. Vaccination campaigns, while costly upfront, significantly lower healthcare expenditures over time by preventing the disease. Similarly, regular screening programs can detect pre-cancerous changes early, reducing the need for expensive treatments for advanced cancer.

In summary, the economic and environmental costs of cervical cancer are substantial but can be mitigated through investment in prevention and early detection strategies. Reducing the incidence of cervical cancer not only saves lives but also lowers the long-term financial and environmental impacts on healthcare systems.

**1.4** **Current methodologies for cervical cancer detection**

Cervical cancer is primarily detected through screening programs using two key methods: the Pap smear and HPV testing. The Pap smear involves collecting cells from the cervix and examining them for abnormal changes that could indicate pre-cancerous or cancerous conditions. This test has been the cornerstone of cervical cancer screening for decades, significantly reducing the incidence of cervical cancer in populations where it is routinely performed.

HPV testing, a more recent advancement, directly detects the presence of high-risk HPV strains that are most likely to cause cervical cancer. HPV tests are often used in conjunction with or as a follow-up to abnormal Pap smear results. They are highly effective in identifying women at risk for cervical cancer, even before cellular changes occur. This allows for earlier intervention and closer monitoring.

In low-resource settings, visual inspection with acetic acid (VIA) is used as an alternative to Pap smears and HPV tests. VIA is a cost-effective method where the cervix is visually examined after applying acetic acid, which highlights abnormal areas. However, it is less accurate than laboratory-based methods and relies heavily on the skill of the healthcare provider.

While these screening methods have been successful in reducing cervical cancer rates, they are not without limitations. False positives and false negatives can occur, leading to unnecessary treatments or missed diagnoses. Additionally, access to regular screening remains a challenge in many parts of the world due to financial, logistical, and cultural barriers.

While screening methods have reduced cervical cancer incidence in many parts of the world, they still have limitations. False positives, limited accessibility in rural areas, and cost are barriers that need to be addressed. Researchers are continuously working to improve diagnostic accuracy and accessibility by incorporating new technologies.

As a result, there is a growing interest in improving cervical cancer detection methods using modern technology, particularly in the application of machine learning models that can analyze complex datasets and provide more accurate, scalable, and accessible diagnostic solutions.

**1.5 Applications of machine learning to combat cervical cancer**

Machine learning (ML) has emerged as a powerful tool in the fight against cervical cancer, offering new possibilities for improving detection, diagnosis, and treatment. ML models can analyse vast amounts of data from medical records, imaging, and genetic testing to identify patterns and predict a patient's risk of developing cervical cancer. By learning from historical data, these models can provide insights that go beyond traditional diagnostic methods.

One of the key applications of ML in cervical cancer prediction is improving the accuracy of screening results. For instance, ML algorithms can be trained to analyze Pap smear images or HPV test results, reducing the likelihood of human error and increasing detection accuracy. These algorithms can also prioritize cases based on the level of risk, enabling more efficient resource allocation in healthcare systems.

In addition to enhancing existing screening methods, ML models can incorporate various data sources such as a patient’s medical history, lifestyle factors, and genetic markers to provide personalized risk assessments. This allows for more tailored screening schedules and interventions, focusing resources on individuals at the highest risk while reducing unnecessary procedures for others.

Another promising area is the development of ML-driven diagnostic tools that can be deployed in low-resource settings. By using AI models that run on mobile devices or cloud platforms, healthcare providers can offer cervical cancer screening and risk assessments in areas without access to specialized laboratories. These tools can democratize access to life-saving screening, particularly in underserved regions.

Furthermore, machine learning has potential applications beyond early detection, such as predicting treatment outcomes and personalizing therapies. ML models can analyze patient data to determine the most effective treatment strategies, potentially improving survival rates and reducing treatment-related side effects.

# CHAPTER-2 LITERATURE SURVEY

## LITERATURE SURVEY

#### Literature review

Several efforts have been made to evaluate the performance of Machine Learning (ML) algorithms for pre - screening for cervical cancer [10]. Cervical cancer is the predominant source of cancer-related deaths in women globally, which behaves epidemiologically like a low-infectious venereal illness. There are several risk factors for cervical cancer that are associated with HPV exposure. The implementation of screening also had an impact on the significant variations in incidence between nations. Liquid-based cytology (LBC), visual inspection with acetic acid and a typical Pap smear, and HPV testing are the main screening techniques used. However, the accuracy of the Pap smear and other techniques are not always accurate. Hence, there is a need to spread more awareness among women and society to undertake preventive measures. Cervical cancer can become the first cancer to be eliminated by humans [16].

Al Mudawi et al. present a perceptive method to predict cervical cancer in patients with machine learning algorithms. A classification score of 100 percent was obtained using the random forest algorithm, and an accuracy of 99% was achieved using SVM. Decision tree, logistic regression, SVM, KNN, AdaBoost, XGBoost, and random forest were also deployed and validated. The study observes that the performance of traditional machine learning methods is relatively low [17].

These studies have used different techniques such as classifier methods, feature selection techniques, decision tree classifiers, and over/under sampling to balance the data. The studies have also employed different algorithms like Logistic regression, SVM, Decision tree, Random forest and Ad boost classifier to predict the outcome of the disease [18].

Abdoh et al. concluded in their research that the following factors pose the highest risk for the development of this disease: sexually transmitted disease (STDs), intra-uterine device (IUD), hormonal contraceptives and the age at which first sexual intercourse happens. Wu and Zhou claimed that the number of sexual partners, the age when first sexual intercourse happens, the number of smoke packs smoked per year and the number of years that the patient uses hormonal contraceptives increase the possibility of developing cervical cancer. Nithya and Ilango identified ten core features as being most important for predicting cancer [2].

One prominent study utilized Support Vector Machines (SVM) to classify cervical cancer risk based on patient demographic and clinical data. The study demonstrated that SVM could achieve high accuracy in identifying high-risk patients, making it a valuable tool for early intervention. Similarly, Random Forest (RF) algorithms have been employed to analyse complex datasets, showing robustness in handling large feature sets and achieving reliable prediction outcomes [12].

One of the author used XG Boost, SVM, and Random Forest (RF) to examine the information on cervical disease. The data was retri

eved from the "UCI machine learning repository," and it contains the clinical records of 858 patients along with 32 risk factors and 4 objective variables. To cope with the dataset's unevenness, they used Borderline-SMOTE. Using a quality arrangement module, they developed a concept and material support pattern to predict cervical disease. In after researching several techniques for cervical cancer growth. KNN got the highest accuracy rate i.e., 83.16% [19].

XGBoost, SVM, and Random Forest were used by Deng et al. to evaluate data on cervical disease. The dataset, which consists of the clinical histories of 858 individuals as well as 32 risk factors and four objective variables, was gathered from the "UCI machine learning repository." To deal with the dataset's unevenness, they employed Borderline-SMOTE [20].

#### Motivation

The motivation behind applying machine learning models to cervical cancer prediction stems from the growing need for more accurate, scalable, and accessible diagnostic tools. Traditional screening methods, while effective, face challenges in terms of accessibility and accuracy, especially in low-resource settings where healthcare infrastructure is limited. In regions with inadequate access to trained cytologists and laboratory facilities, the implementation of ML-based solutions could dramatically improve early detection rates and reduce mortality.

Cervical cancer remains a significant public health issue worldwide, particularly in developing countries where women often lack access to preventive healthcare services like regular Pap smears and HPV vaccinations. The burden of cervical cancer is particularly high in these regions, and the motivation to develop machine learning models lies in their ability to provide cost-effective, accurate, and scalable screening tools that can be deployed even in resource-constrained environments.

Another major driver for integrating ML into cervical cancer prediction is the potential for reducing human error in diagnosis. Manual interpretation of Pap smears and HPV tests can lead to diagnostic variability, with results dependent on the skill and experience of the clinician. By utilizing machine learning algorithms that can consistently analyze data without bias, it becomes possible to standardize the diagnostic process and improve the overall quality of care.

Furthermore, the ability of machine learning models to analyze a wide range of data, including patient demographics, lifestyle factors, and genetic information, motivates researchers to explore personalized screening approaches. Unlike traditional one-size-fits-all screening programs, ML models can assess individual risk factors and provide tailored recommendations for screening and follow-up. This personalized approach can optimize resource allocation by focusing on high-risk individuals, ultimately improving health outcomes.

Lastly, the advancement of technology and increased availability of large medical datasets have further motivated the use of machine learning in healthcare. With the growing availability of electronic health records (EHRs), genomic data, and advanced imaging techniques, ML algorithms can leverage these diverse data sources to provide more comprehensive and accurate predictions of cervical cancer risk. This, in turn, encourages further research and development in the field of ML-driven healthcare diagnostics.

# CHAPTER-3 PROPOSED SYSTEM

### PROPOSED SYSTEM

The proposed system aims to develop an accurate and scalable machine learning model to predict cervical cancer risk. The system is designed to work with a dataset containing patient demographics, medical history, and test results, ensuring that it can process large volumes of data from diverse sources. The system follows a structured workflow: data collection, preprocessing, model training, evaluation, and deployment. Each stage of this process is designed to improve the system’s ability to predict cancer risk with high accuracy.

First, the system will collect relevant datasets, potentially sourced from healthcare institutions or public databases. This data will consist of features like age, number of pregnancies, smoking habits, Pap smear results, and HPV infection status, which are known to correlate with cervical cancer risk. Once collected, the data will be preprocessed to ensure it is ready for machine learning models. This step will involve handling missing data, encoding categorical variables, and normalizing numerical data.

After preprocessing, machine learning algorithms will be applied to build the predictive model. Various models will be tested, including logistic regression, decision trees, random forests, and deep learning models like the Multilayer Perceptron (MLP). Each model will be trained on the preprocessed data and evaluated to determine which one provides the highest predictive accuracy.

One of the key features of the proposed system is its adaptability. The system can be updated and retrained as more data becomes available, allowing it to evolve with new medical knowledge and patient information. This adaptability ensures that the system remains relevant and effective in detecting cervical cancer over time.

Another critical component of the system is model evaluation. The system will utilize performance metrics like accuracy, precision, recall, and F1 score to assess the effectiveness of each model. These metrics provide insight into the model’s ability to correctly classify patients as high or low risk, helping to minimize false positives and false negatives.

The ultimate goal of the system is to provide healthcare professionals with a powerful tool for early cancer detection. By automating the diagnostic process, the system can save time, reduce errors, and improve outcomes for patients. In particular, it has the potential to make cervical cancer screening more accessible in low-resource settings, where access to medical specialists is often limited.

#### Input dataset

The input dataset for this system will comprise various patient records that include both demographic and clinical data. This dataset is the foundation of the system as it provides the raw data needed to train and test the machine learning models. Common sources of these datasets could be hospitals, public health organizations, or publicly available cancer-related data repositories.

The dataset includes variables such as patient age, number of pregnancies, smoking habits, HPV test results, and medical history, all of which are considered risk factors for cervical cancer. To ensure that the system is versatile, the dataset should be large and diverse, representing a wide array of patient profiles. This helps the model generalize better and improve prediction accuracy when dealing with new, unseen data.

#### Detailed Features of the Dataset

The features in the dataset represent the key variables that will be used to predict cervical cancer risk. These features can be categorized into demographic, lifestyle, and medical factors. Each of these categories provides important insights into the patient’s overall health and potential cancer risk, making them essential for the machine learning model.

Demographic features include the patient’s age, marital status, and education level. Age is a well-known risk factor for cervical cancer, with the disease being more common in women over the age of 30. Marital status and education level can also indirectly affect cancer risk by influencing lifestyle choices and access to healthcare.

Lifestyle features include smoking status, number of sexual partners, and contraceptive use. Smoking has been linked to a higher risk of cervical cancer due to its impact on the immune system’s ability to fight HPV infections. Similarly, a higher number of sexual partners increases the risk of contracting HPV, the primary cause of cervical cancer.

Medical history features are perhaps the most important in predicting cervical cancer. These include the patient’s HPV test results, Pap smear results, and history of STDs. HPV infection is the leading cause of cervical cancer, so a positive HPV test is a significant risk factor. Pap smear results provide information about abnormal cell changes in the cervix, which can indicate early signs of cancer. A history of STDs can also increase the likelihood of developing cervical cancer.

Other relevant features may include the patient’s number of pregnancies, family history of cancer, and immunization history. Multiple pregnancies have been associated with a higher risk of cervical cancer, possibly due to hormonal changes or the physical stress on the cervix. A family history of cancer may indicate genetic predispositions that could increase the patient’s ris

#### Data Pre-processing

Data pre-processing is the essential process of preparing raw data for analysis and modelling by cleaning, transforming, and structuring it to enhance data quality and utility. It involves tasks like handling missing values, correcting errors, encoding features, and scaling data to ensure it's in an optimal form for further analysis. It encompasses a range of operations and transformations designed to refine raw data, ensuring that it is clean, structured, and amenity subsequent analysis. This process is driven by its manifold significance in data science and analysis.

Through meticulous data cleaning, transformation, feature engineering, dimensionality reduction, outlier handling, scaling, and data splitting, it prepares raw data for more accurate and reliable analysis and modelling. Ultimately, the goal is to obtain more meaningful insights, make informed decisions, and optimize predictive models for a wide range of applications in data science and analysis.

**3.2.1 Missing Values**

Missing values in medical datasets are common and must be handled carefully to avoid introducing bias into the model. The cervical cancer dataset may have missing information for key features like HPV test results, Pap smear outcomes, or lifestyle factors such as smoking habits. Ignoring missing values or handling them incorrectly can lead to inaccurate predictions and reduce the model’s reliability.

There are several techniques for addressing missing values, each with its own advantages and disadvantages. The simplest approach is to remove records with missing values, but this is only feasible if the amount of missing data is small. Removing too many records could result in a loss of valuable information and reduce the dataset’s size, potentially affecting the model’s performance.

Imputation is a more sophisticated method for handling missing values. This involves replacing missing values with estimated ones, such as the mean, median, or mode of the corresponding feature. For example, if a patient’s smoking status is missing, it could be replaced with the most common value for smoking status in the dataset. This approach retains the full dataset and can help prevent data loss.

More advanced imputation methods, such as K-Nearest Neighbors (KNN) imputation, can also be used. KNN imputation works by finding similar records in the dataset and using their values to estimate the missing ones. This method can provide more accurate estimates than simple mean or median imputation, as it takes into account the relationships between features.

It is also important to analyze the pattern of missing values to determine whether they are random or systematic. If certain types of patients (e.g., older patients) are more likely to have missing data, this could introduce bias into the model. Understanding the nature of missing data helps inform the choice of imputation method and ensures that the dataset is handled appropriately.

Overall, handling missing values is a crucial part of the data preprocessing process. By carefully filling in missing data, the proposed system can ensure that it maintains the integrity and accuracy of the dataset, leading to better model performance.

**3.2.1.1 Parameters of the fillna Method**

The fillna method is a commonly used technique in data preprocessing to handle missing values. It allows for the replacement of missing data with a specified value or strategy, such as using the mean, median, or mode of the column. In the context of cervical cancer prediction, this method is useful for ensuring that no important patient data is missing when training the machine learning model.

The fillna method accepts several parameters that control how missing values are replaced. One of the most commonly used parameters is the "value" parameter, which specifies the value to replace missing data. For example, missing values in the "age" column could be replaced with the median age of the patients in the dataset. This is useful when the missing values are numerical and can be meaningfully filled in with statistical estimates.

Another important parameter is the "method" parameter, which allows for forward or backward filling of missing data. Forward filling replaces missing values with the last observed value in the column, while backward filling replaces them with the next observed value. This method is particularly useful for time-series data but can be adapted for other types of data as well.

The "inplace" parameter determines whether the changes should be made directly to the original dataset or if a copy of the dataset should be created. Setting "inplace=True" modifies the original dataset, while "inplace=False" creates a new dataset with the filled values, preserving the original data.

Additionally, the "axis" parameter allows for the specification of whether the fill operation should be applied along rows or columns. By setting axis=0, the missing values are filled along the columns, which is the default behavior. This ensures that missing data is filled for each feature individually, based on the specified strategy.

Using the fillna method with appropriate parameters is essential for ensuring that missing values are handled efficiently and consistently across the dataset. This helps improve the overall quality of the data and ensures that the machine learning model is trained on complete and accurate information.

**3.2.2 Data Encoding**

Data encoding is a vital step in preparing categorical data for machine learning models, which require numerical input. In the cervical cancer dataset, features like smoking status, HPV test results, and Pap smear outcomes are categorical variables that need to be transformed into numerical values to be used in the model. Proper encoding of these variables is crucial for ensuring that the model can interpret the data correctly and make accurate predictions.

One of the most commonly used encoding methods is one-hot encoding. This method transforms categorical variables into a set of binary variables, each representing a category. For example, the smoking status feature, which may have categories like "yes" and "no," would be transformed into two binary variables: "smoker" and "non-smoker." Each patient would receive a 1 or 0 for each of these binary variables, indicating their smoking status. This method is particularly useful for categorical variables with no inherent order.

Label encoding is another technique that assigns numerical values to categories based on their order. For example, if a variable has three categories—"low," "medium," and "high"—label encoding would assign values of 1, 2, and 3, respectively. This method is useful for ordinal variables, where the categories have a meaningful order. However, it may not be appropriate for nominal variables, where there is no inherent ranking.

In some cases, binary encoding can also be used, especially when dealing with high-cardinality categorical variables (i.e., variables with many categories). Binary encoding first converts the categories into numerical values and then encodes these values as binary digits. This reduces the dimensionality of the encoded data compared to one-hot encoding, making it more efficient for large datasets.

The choice of encoding method depends on the nature of the categorical variable and the machine learning algorithm being used. Some algorithms, like decision trees, can handle categorical data without explicit encoding, while others, like neural networks, require numerical input. It is important to select the appropriate encoding method to ensure that the model can learn from the data effectively.

By encoding categorical variables properly, the system ensures that all features are represented numerically, allowing the machine learning model to process the data efficiently. This step is crucial for improving the model’s predictive performance and ensuring that it can handle a wide range of patient data.

#### Model Building

Model building is the core phase of the cervical cancer prediction system. During this phase, various machine learning algorithms are trained on the preprocessed dataset to learn patterns and relationships between the input features and the target variable, which is the cancer risk. The goal is to identify the most effective algorithm for accurately predicting cervical cancer risk based on patient data.

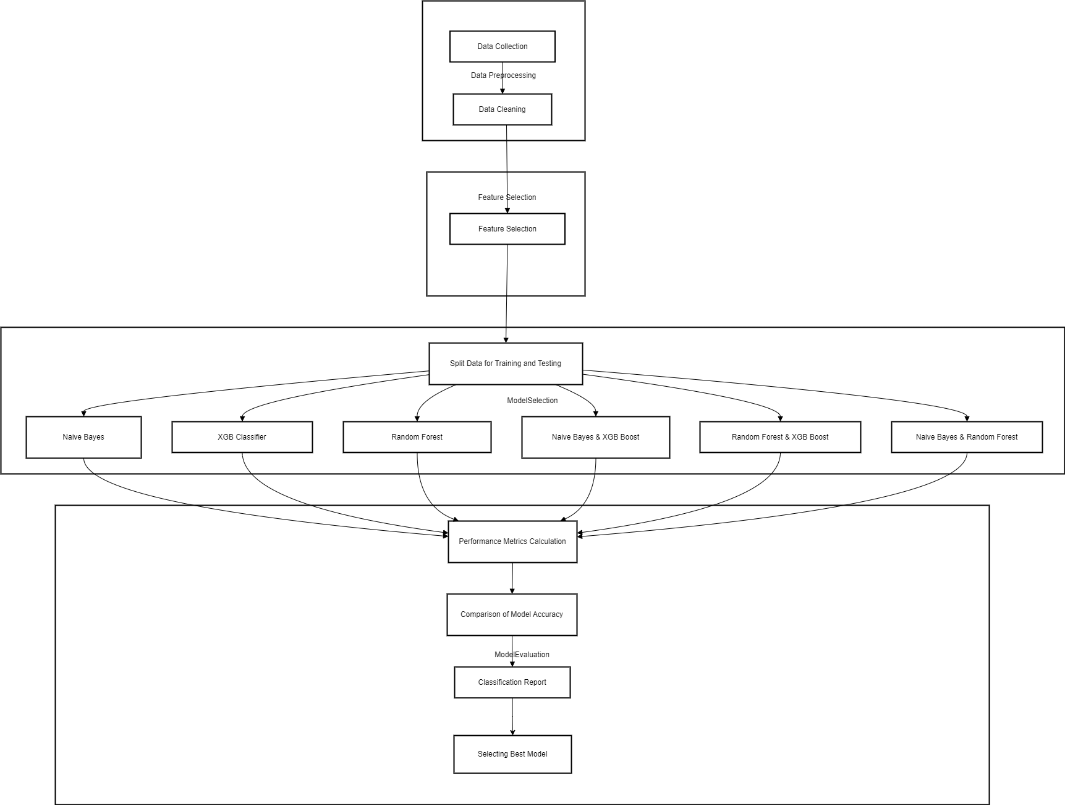
The process begins by splitting the dataset into training and testing sets. The training set is used to build the model, while the testing set is used to evaluate its performance. Different machine learning algorithms, such as logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks, will be applied to the dataset. Each of these algorithms has its own strengths and weaknesses, and their performance will be compared to select the best-performing model.

One of the challenges in model building is ensuring that the model does not overfit or underfit the data. Overfitting occurs when the model learns the training data too well, including noise and outliers, which reduces its ability to generalize to new data. Underfitting, on the other hand, occurs when the model is too simple to capture the underlying patterns in the data. Techniques such as cross-validation, regularization, and hyperparameter tuning will be used to optimize the model and prevent overfitting or underfitting.

The next step is to evaluate the model’s performance using various metrics such as accuracy, precision, recall, and the F1 score. Accuracy measures the proportion of correctly predicted cases, but it may not be sufficient in cases where the dataset is imbalanced. Precision and recall are more informative metrics for evaluating the model’s ability to correctly identify high-risk patients without generating too many false positives. The F1 score combines precision and recall into a single metric that balances both concerns.

Another important consideration in model building is interpretability. While complex models like neural networks may provide high accuracy, they can be difficult to interpret. Simpler models like decision trees may provide more insight into how the model makes its predictions, which can be useful for healthcare professionals. The trade-off between accuracy and interpretability will be carefully considered when selecting the final model.

Once the best model has been identified, it will be fine-tuned and optimized for deployment. This may involve adjusting the model’s hyperparameters, retraining it on a larger dataset, or applying techniques like ensembling, where multiple models are combined to improve predictive performance. The final model will then be ready for integration into the cervical cancer prediction system.



**3.3.2 MLP Algorithm**

The Multilayer Perceptron (MLP) algorithm is a type of artificial neural network that is particularly suited for complex classification problems like cervical cancer prediction. The MLP consists of an input layer, one or more hidden layers, and an output layer. Each layer is composed of neurons that apply a weighted sum of the inputs followed by a non-linear activation function. This structure allows the MLP to learn complex, non-linear relationships in the data.

In the proposed system, the MLP will be used to predict the risk of cervical cancer based on input features such as age, HPV test results, Pap smear results, and lifestyle factors. The input layer will consist of neurons representing each of the features in the dataset. These inputs will be passed through one or more hidden layers, where the MLP will learn to extract important patterns and relationships from the data.

One of the key advantages of MLPs is their ability to handle a wide variety of data types and relationships. Unlike traditional models that assume a linear relationship between the input features and the target variable, MLPs can model non-linear interactions, making them more flexible and powerful for complex prediction tasks. This makes the MLP particularly well-suited for predicting cervical cancer risk, where the relationships between risk factors and outcomes are often complex and non-linear.

Training the MLP involves optimizing the weights of the neurons using an algorithm called backpropagation. During backpropagation, the model’s predictions are compared to the actual outcomes, and the error is propagated back through the network to adjust the weights. This process is repeated iteratively until the model’s performance converges to an optimal level. Techniques like gradient descent and adaptive learning rates are used to speed up the training process and improve convergence.

The MLP’s performance will be evaluated using metrics such as accuracy, precision, recall, and F1 score. One challenge with MLPs is that they can be prone to overfitting, especially when dealing with small datasets or complex models. To prevent overfitting, techniques like dropout (randomly dropping neurons during training) and regularization (penalizing large weights) will be used.

Finally, the MLP will be fine-tuned through hyperparameter optimization. This involves adjusting parameters like the number of hidden layers, the number of neurons in each layer, the learning rate, and the activation functions. The goal is to find the optimal combination of hyperparameters that maximizes the model’s performance on the cervical cancer dataset

#### Methodology of the system

Having discussed the foundational elements in the preceding sections, we now venture into the core of our traffic congestion prediction system. In this section, we embark on a journey through the inner workings of our model, unveiling the methodology that drives our system's ability to forecast traffic congestion. Just as a well-orchestrated symphony requires each instrument to play its part harmoniously, our methodology combines data, pre-processing, modelling, and evaluation to create a seamless and efficient prediction system.

The methodology of the proposed system follows a structured approach to building an accurate and reliable cervical cancer prediction model. The first step in the methodology is data collection, where a comprehensive dataset of patient records is gathered from hospitals, clinics, or publicly available repositories. This dataset includes features like demographic information, medical history, and test results, all of which are relevant to predicting cervical cancer risk.

Once the dataset has been collected, the next step is data preprocessing. This involves handling missing values, encoding categorical variables, and normalizing numerical features. Preprocessing ensures that the dataset is clean and consistent, making it suitable for machine learning models. Handling missing values is particularly important, as incomplete records can introduce bias into the model. Various imputation techniques will be used to fill in missing data, ensuring that the model is trained on complete information.

After preprocessing, the dataset will be split into training and testing sets. The training set will be used to build the model, while the testing set will be used to evaluate its performance. Multiple machine learning algorithms will be applied to the training set, including logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks like the Multilayer Perceptron (MLP). Each algorithm will be trained on the data and evaluated based on its ability to predict cervical cancer risk.

The model’s performance will be evaluated using a range of metrics, including accuracy, precision, recall, and F1 score. Accuracy measures the proportion of correct predictions, but it may not be sufficient in cases where the dataset is imbalanced (i.e., where there are more low-risk patients than high-risk patients). Precision and recall are more informative metrics in such cases, as they provide insights into the model’s ability to correctly identify high-risk patients without generating too many false positives. The F1 score combines precision and recall into a single metric, providing a balanced measure of the model’s performance.

Once the best-performing model has been identified, it will be fine-tuned through hyperparameter optimization. This involves adjusting parameters like the learning rate, regularization strength, and the number of neurons in the neural network. Hyperparameter tuning is essential for improving the model’s performance and ensuring that it generalizes well to new data.

The final step in the methodology is model deployment. Once the model has been trained and evaluated, it will be integrated into a user-friendly interface that allows healthcare professionals to input patient data and receive predictions. The system will provide explanations for its predictions, helping doctors understand the factors that contributed to the risk assessment. This makes the system a valuable tool for early detection of cervical cancer, especially in resource-limited settings.

#### Model Evaluation

Model evaluation is critical to ensuring that the cervical cancer prediction model performs well on unseen data and can be trusted for real-world applications. This process involves assessing the model's accuracy, precision, recall, F1 score, and other key performance metrics. These metrics help quantify the model's ability to correctly identify high-risk patients while minimizing false positives and false negatives.

Accuracy is the first metric used in model evaluation. It calculates the percentage of correct predictions made by the model. While accuracy is a basic measure, it may not provide a full picture when dealing with imbalanced datasets, such as those in cervical cancer prediction, where the majority of patients may fall in the low-risk category. A model that predicts most patients as low-risk may still achieve high accuracy, even though it fails to correctly identify high-risk patients.

To provide more insight, precision and recall are used. Precision measures the proportion of true positive predictions out of all patients classified as high-risk. High precision means that most patients flagged as high-risk truly have an elevated risk of cervical cancer. Recall measures the proportion of actual high-risk patients that were correctly identified by the model. High recall ensures that most patients at risk of developing cervical cancer are correctly diagnosed.

The F1 score combines precision and recall into a single metric. It is particularly useful when there is an uneven distribution between the classes (in this case, high-risk and low-risk patients). A high F1 score indicates a good balance between precision and recall, reducing both false positives and false negatives.

In addition to these metrics, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) provide further evaluation of the model’s performance. The ROC curve plots the true positive rate against the false positive rate, and the AUC measures the model's ability to distinguish between patients who are at high risk and those who are not. A higher AUC indicates better model performance in separating the two groups.

Finally, cross-validation techniques will be employed to ensure that the model generalizes well to new data. Cross-validation involves splitting the dataset into multiple subsets and training the model on different combinations of these subsets. This helps prevent overfitting, where the model performs well on the training data but poorly on new, unseen data. By using cross-validation and evaluating multiple performance metrics, the model's reliability and effectiveness can be ensured before deployment.

**Confusion Matrix :**

True Positive (TP): demonstrate accurately predicts the positive class

True Negative (TN): show accurately predicts the negative class

False Positive (FP): demonstrate predicts positive, but it’s negative.

False Negative (FN): show predicts negative, but it’s positive

**Accuracy:**

Accuracy=

**Precision:**

Precision=

**Recall :**

Recall=

**F1 Score:**

F1 Score=2×

#### Constraints

Like any machine learning system, the proposed cervical cancer prediction model faces several constraints that need to be considered during development. These constraints range from data-related challenges to computational limitations and ethical concerns. Understanding these constraints helps in designing a system that is both practical and reliable.

The first constraint is the quality and availability of data. Medical datasets, especially those used for cervical cancer prediction, may have missing or incomplete records. Patient data can vary widely in quality, and gaps in crucial features like HPV test results or medical history can affect the model’s accuracy. To address this, advanced techniques like imputation for missing values must be used, but it still limits the precision of the model.

Another data-related constraint is the potential imbalance in the dataset. Since cervical cancer is relatively rare compared to other medical conditions, most patients in the dataset may fall into the low-risk category. This class imbalance can lead to biased models that predict most patients as low-risk, ignoring the smaller group of high-risk patients. Techniques like oversampling or under sampling and the use of specialized algorithms designed to handle imbalanced data must be implemented to ensure that the model properly identifies high-risk patients.

Computational limitations can also pose challenges, particularly when using complex models like neural networks. Training deep learning models, such as the Multilayer Perceptron (MLP), requires significant computational resources, including powerful hardware like GPUs. For institutions with limited resources, deploying these models may not be feasible. Simplified models or cloud-based solutions could be considered to address this issue.

Ethical concerns are another significant constraint, especially regarding patient privacy and data security. Medical data is highly sensitive, and ensuring compliance with data protection regulations like HIPAA (in the United States) or GDPR (in Europe) is essential. Secure storage, anonymization, and encryption of patient data are necessary steps to safeguard patient privacy.

Finally, the interpretability of the model poses a challenge. Complex models like neural networks, while highly accurate, can be difficult to interpret. Healthcare professionals may be hesitant to rely on a "black-box" model that provides predictions without clear explanations. To address this, explainability techniques such as SHAP (SHapley Additive exPlanations) can be used to provide insights into which features contributed most to each prediction, improving trust in the system.

#### Cost and sustainability Impact

Developing and deploying a machine learning model for cervical cancer prediction involves several cost considerations. These costs include data collection, computational resources, model development, and maintenance. However, the long-term benefits of early detection and prevention of cervical cancer can outweigh these initial costs, especially in terms of saving lives and reducing healthcare expenses.

The first major cost is related to data collection and storage. Gathering a large, high-quality dataset of patient records requires collaboration with healthcare institutions, investment in data acquisition, and secure data storage infrastructure. This can be a significant expense, especially for organizations that need to anonymize and protect sensitive medical data. Cloud-based storage solutions can be used to mitigate some of these costs, but they introduce recurring fees based on usage.

Another cost factor is the computational power required to train and deploy the machine learning models. Training deep learning models like MLPs can be resource-intensive, requiring specialized hardware such as GPUs or access to cloud-based platforms. Institutions with limited resources may face challenges in handling these computational requirements, but cloud platforms like Google Cloud or AWS provide scalable options that allow for efficient model training without large upfront investments.

The sustainability impact of the system is also an important consideration. By focusing on early detection, the system has the potential to reduce the overall burden on healthcare systems. Early diagnosis of cervical cancer leads to less invasive treatments, shorter hospital stays, and reduced healthcare costs. This preventive approach can result in substantial long-term savings for healthcare providers, while also improving patient outcomes.

Additionally, the system promotes more efficient use of medical resources by targeting screening and follow-up efforts towards high-risk patients. Instead of subjecting all patients to the same level of scrutiny, the model can identify those who need more intensive monitoring, allowing healthcare professionals to allocate their time and resources more effectively. This can lead to more sustainable healthcare practices by reducing unnecessary tests and procedures.

Finally, the development of this system aligns with global health initiatives aimed at improving access to healthcare in low-resource settings. By providing an affordable, scalable, and reliable tool for cervical cancer prediction, the system can help reduce disparities in cancer detection and treatment. This has the potential to significantly impact public health, particularly in regions with limited access to specialized care.

#### 3.8. Experiment / Product Results (IEEE 1012 & IEEE 1633)

Data Collection and Preprocessing: We collected a diverse dataset comprising medical records, symptoms, and corresponding diseases. Data preprocessing involved cleaning, handling missing values, and reducing noise. The dataset was then split into training and testing sets.

**4. Implementation**

The implementation phase is crucial for transforming theoretical concepts into a functional cervical cancer prediction system. This section involves setting up the appropriate environment, writing the code for data preprocessing, and deploying machine learning models. The system is implemented in a way that ensures scalability, flexibility, and efficiency.

During this stage, the raw data collected from various sources is prepared for use by the machine learning models. This requires preprocessing the data to handle missing values, encode categorical data, and scale the numerical features. After preprocessing, machine learning models, including the Multilayer Perceptron (MLP), are trained and evaluated on the data. The implementation includes the actual code for data handling and model building, ensuring that the system is robust and can be easily updated as new data becomes available.

The system is designed using Python as the programming language, with libraries such as Pandas for data manipulation, Scikit-learn for machine learning models, and TensorFlow/Keras for neural networks like MLP. The modular structure of the implementation allows for easy integration of new datasets and machine learning algorithms, making the system highly adaptable.

The implementation also includes a user interface that allows healthcare professionals to input patient data and receive cancer risk predictions. This interface is designed to be user-friendly and intuitive, enabling non-technical users to interact with the system seamlessly. The predictions made by the system are accompanied by explanations, helping doctors understand the rationale behind the risk assessment.

In summary, the implementation transforms the proposed system into a working tool that can be used in real-world healthcare settings. The code is structured to be efficient, scalable, and easy to maintain, ensuring that the system remains relevant as new data and techniques become available.

**4.1 Environment Setup**

Setting up the environment is the first step in the implementation process. The environment includes both the hardware and software requirements needed to develop and run the cervical cancer prediction system. This system is implemented using Python, making it platform-independent and easily deployable on various systems, including local machines and cloud platforms.

The key software components include Python libraries such as Pandas for data manipulation, Scikit-learn for machine learning models, and TensorFlow or Keras for building deep learning models like the MLP. These libraries provide pre-built functions and tools for data preprocessing, model training, and evaluation, significantly simplifying the implementation process. The system may also require Jupyter Notebooks or Google Colab for an interactive development environment, where code can be written and tested in real-time.

The first step in setting up the environment is to install Python and the required libraries. This can be done using package management tools like pip. For example, to install TensorFlow, the command pip install tensorflow is used. Similarly, other required libraries such as Scikit-learn, Pandas, and Matplotlib can be installed using pip.

In terms of hardware, a system with sufficient computational power is needed, especially for training deep learning models like MLPs. This may require a high-performance CPU or, preferably, a GPU. Cloud platforms like Google Cloud, AWS, or Google Colab provide access to GPU instances, which can significantly speed up model training. These platforms also offer scalable storage solutions for handling large datasets.

Once the environment is set up, the next step is to verify the installation by running simple test scripts. For example, importing the necessary libraries and loading a sample dataset ensures that the environment is correctly configured and ready for model development.

**4.2 Sample Code for Preprocessing and MLP Operations**

The preprocessing stage is crucial for preparing the dataset for the MLP model. The sample code for preprocessing typically includes handling missing values, encoding categorical variables, and normalizing numerical features. Below is an outline of the sample code used for these tasks:

# Import necessary libraries

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler, LabelEncoder

from tensorflow.keras.models import Sequential

from tensorflow.keras.layers import Dense

# Load the dataset

data = pd.read\_csv('cervical\_cancer\_data.csv')

# Handling missing values (filling with the median)

data.fillna(data.median(), inplace=True)

# Encoding categorical variables (e.g., for HPV status)

label\_encoder = LabelEncoder()

data['HPV\_status'] = label\_encoder.fit\_transform(data['HPV\_status'])

# Splitting the dataset into features (X) and target (y)

X = data.drop('cancer\_risk', axis=1) # Features

y = data['cancer\_risk'] # Target

# Normalizing the numerical features

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

# Splitting the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_scaled, y, test\_size=0.2, random\_state=42)

Once preprocessing is complete, the next step is to implement the MLP model. The MLP is built using the TensorFlow/Keras library. Below is sample code to define, compile, and train the MLP:

# Building the MLP model

model = Sequential()

model.add(Dense(64, activation='relu', input\_dim=X\_train.shape[1]))

model.add(Dense(32, activation='relu'))

model.add(Dense(1, activation='sigmoid')) # Binary classification output

# Compiling the model

model.compile(optimizer='adam', loss='binary\_crossentropy', metrics=['accuracy'])

# Training the model

model.fit(X\_train, y\_train, epochs=50, batch\_size=32, validation\_data=(X\_test, y\_test))

In this sample code, the MLP consists of an input layer (which matches the number of features in the dataset), two hidden layers with relu activation functions, and an output layer with a sigmoid activation function (since this is a binary classification problem). The model is compiled using the Adam optimizer and binary cross-entropy loss function, and it is trained on the preprocessed data for 50 epochs.

The trained model can then be evaluated on the test data to determine its performance:

# Evaluating the model

test\_loss, test\_acc = model.evaluate(X\_test, y\_test)

print(f"Test Accuracy: {test\_acc}")

This code implements the preprocessing steps and MLP model for cervical cancer prediction, demonstrating the core operations of the system.

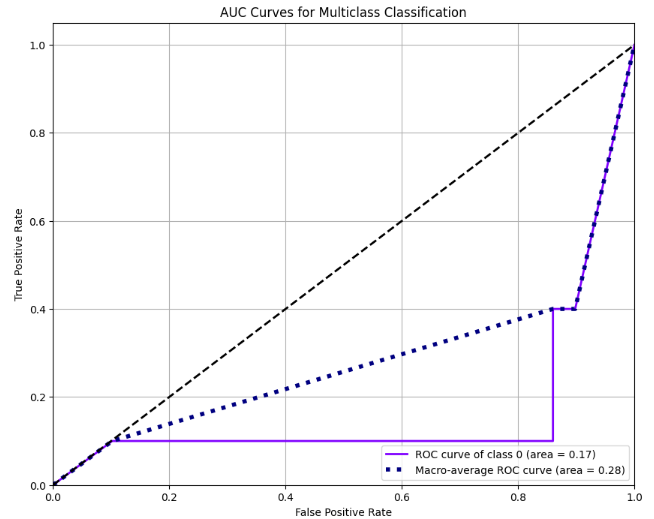
# CHAPTER-4 RESULTS

Naïve Bayes :

|  |  |
| --- | --- |
|  |  |
|  |  |

XGB Classifier

|  |  |
| --- | --- |
|  |  |



Random Forest

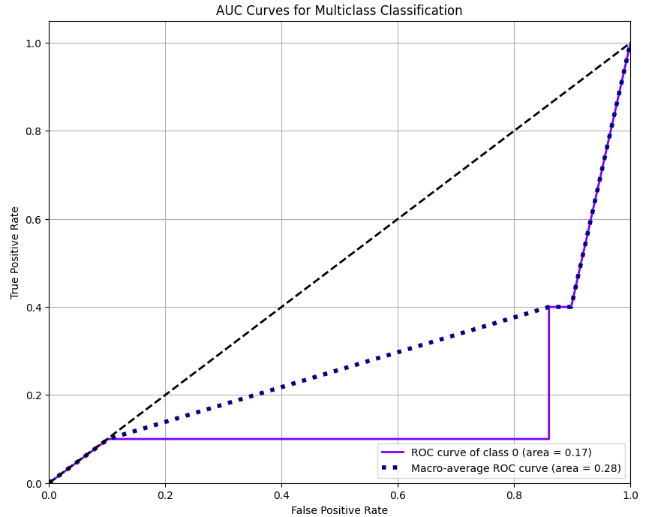
|  |  |
| --- | --- |
|  |  |
|  |  |

Naïve Bayes & XGB Boost

|  |  |
| --- | --- |
|  |  |
|  |  |

Random Forest & XGB Boost

|  |  |
| --- | --- |
|  |  |



Naïve Bayes & Random Forest

|  |  |
| --- | --- |
|  |  |
|  |  |

### REFERENCES

[1] M. Rahimi, A. Akbari, F. Asadi, and H. Emami, “Cervical cancer survival prediction by machine learning algorithms: a systematic review,” BMC Cancer, vol. 23, no. 1, pp. 1–10, 2023, doi: 10.1186/s12885-023-10808-3.

[2] R. Alsmariy, G. Healy, and H. Abdelhafez, “Predicting cervical cancer using machine learning methods,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 7, pp. 173–184, 2020, doi: 10.14569/IJACSA.2020.0110723.

[3] B. Nithya and V. Ilango, “Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction,” SN Appl. Sci., vol. 1, no. 6, pp. 1–16, 2019, doi: 10.1007/s42452-019-0645-7.

[4] R. Kumar, “Machine Learning Techniques for Predicting Cervical Cancer Cases,” no. December, 2022.

[5] J. Singh and S. Sharma, “Prediction of Cervical Cancer Using Machine Learning Techniques,” vol. 14, no. 11, pp. 2570–2577, 2019.

[6] M. Jha, R. Gupta, and R. Saxena, “Cervical Cancer Risk Prediction Using XGboost Classifier,” 2021 7th Int. Conf. Signal Process. Commun. ICSC 2021, pp. 133–136, 2021, doi: 10.1109/ICSC53193.2021.9673474.

[7] N. Al Mudawi and A. Alazeb, “A Model for Predicting Cervical Cancer Using Machine Learning Algorithms,” Sensors, vol. 22, no. 11, 2022, doi: 10.3390/s22114132.

[8] Q. Yin, “The application of machine learning in cervical cancer prediction,” ACM Int. Conf. Proceeding Ser., pp. 12–19, 2021, doi: 10.1145/3468891.3468894.

[9] K. Rayavarapu and K. K. V. Krishna, “Prediction of Cervical Cancer using Voting and DNN Classifiers,” Proc. 2018 Int. Conf. Curr. Trends Towar. Converging Technol. ICCTCT 2018, pp. 1–5, 2018, doi: 10.1109/ICCTCT.2018.8551176.

[10] P. Rawat, M. Bajaj, S. Mehta, V. Sharma, and S. Vats, “A Study on Cervical Cancer Prediction using Various Machine Learning Approaches,” Int. Conf. Innov. Data Commun. Technol. Appl. ICIDCA 2023 - Proc., pp. 1101–1107, 2023, doi: 10.1109/ICIDCA56705.2023.10099493.

[11] F. Edafetanure-ibeh, “EVALUATING MACHINE LEARNING ALGORITHMS FOR CERVICAL CANCER EVALUATING MACHINE LEARNING ALGORITHMS FOR CERVICAL CANCER PREDICTION : A COMPARATIVE ANALYSIS . Author : Faith Tobore Edafetanure-Ibeh . Email : Edafetanureibeh.faith@gmail.com,” no. April, 2024, doi: 10.31219/osf.io/vyuf2.

[12] S. Vaishnodevi, “Predicting Cervical Cancer using Advanced Machine Learning Algorithms,” 2024 2nd Int. Conf. Sustain. Comput. Smart Syst., no. Icscss, pp. 1600–1604, 2024, doi: 10.1109/ICSCSS60660.2024.10625243.

[13] S. Gutierrez-Espinoza and M. Cabanillas-Carbonell, “Machine Learning Analysis for Cervical Cancer Prediction, a Systematic Review of the Literature,” 2021 9th E-Health Bioeng. Conf. EHB 2021, pp. 1–6, 2021, doi: 10.1109/EHB52898.2021.9657567.

[14] A. Shetty and V. Shah, “Survey of Cervical Cancer Prediction Using Machine Learning: A Comparative Approach,” 2018 9th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2018, pp. 1–6, 2018, doi: 10.1109/ICCCNT.2018.8494169.

[15] M. S. Al-Batah, M. Alzyoud, R. Alazaidah, M. Toubat, H. Alzoubi, and A. Olaiyat, “Early Prediction of Cervical Cancer Using Machine Learning Techniques,” Jordanian J. Comput. Inf. Technol., vol. 8, no. 4, pp. 357–369, 2022, doi: 10.5455/jjcit.71-1661691447.

[16] G. Kumawat, S. K. Vishwakarma, and P. Chakrabarti, “Cervical Cancer Prediction Using Machine Learning Techniques,” Lect. Notes Networks Syst., vol. 817, pp. 13–28, 2024, doi: 10.1007/978-981-99-7886-1\_2.

[17] A. Shaikh, P. Verma, and K. Deshmukh, “Efficient Feature Selection Approach for Prediction of Cervical Cancer Using Machine Learning Algorithms,” Proc. IEEE Int. Conf. Model. Simul. Intell. Comput. MoSICom 2023, pp. 446–451, 2023, doi: 10.1109/MoSICom59118.2023.10458795.

[18] S. Devi, C. Komalavalli, and R. Chinnaiyan, “Machine Learning Based Classification Models for Early Analysis and Prediction of Cervical Cancer,” 2023 Int. Conf. New Front. Commun. Autom. Manag. Secur. ICCAMS 2023, vol. 1, pp. 1–5, 2023, doi: 10.1109/ICCAMS60113.2023.10526129.

[19] S. Srivastav, K. Guleria, and S. Sharma, “Predictive Machine Learning Approaches for Cervical Cancer Detection: An Analytical Comparison,” Proc. 17th INDIACom; 2023 10th Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2023, pp. 951–956, 2023.

[20] M. A. Rahman and A. Ghosh, “Risk factor analysis and prediction of Cervical Cancer based on machine learning models,” 2023 26th Int. Conf. Comput. Inf. Technol. ICCIT 2023, pp. 1–4, 2023, doi: 10.1109/ICCIT60459.2023.10441011.