**Cervical Cancer Screening**

**Minor Project I**

Submitted By:

**Astha Dwivedi (9917103186)**

Under the supervision of:

**Dr. Devpriya Soni**



**Department of CSE/IT**

**Jaypee Institute of Information Technology, Noida**

**November 2019**

**ACKNOWLEDGEMENT**

I would like to place on record my deep sense of gratitude to Dr. Devpriya Soni, Associate Professor of Computer Science and Engineering, Jaypee Institute of Information Technology, India for her generous guidance, help and useful suggestions.

I express my sincere gratitude to Dr. Surbhi Gupta, Dept. of Cancer, Sarojini Naidu Medical College, Agra, for her stimulating guidance and supervision throughout the course of present work.

I also wish to extend my thanks to Dr. Ajay Gupta for his constructive suggestions to improve the quality of this project work.

**Signature(s) of Students**

Astha Dwivedi (9917013186)

**i**

**Abstract**

**Objective:** Cervical cancer is the leading gynaecological malignancy worldwide. Cervical cancer the fourth most frequent cancer in women with an estimated 570,000 new cases in 2018 representing 6.6% of all female cancers. Though HPV vaccine is available but these are expensive still not available in many countries. The high mortality rate from cervical cancer globally could be reduced through a comprehensive approach that includes prevention, early diagnosis, effective screening and treatment programmes.

**Method:** The demographic information, habits, and historic medical records of 858 patients collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela consisting thirty-two attribute is obtained from UCI Repository. Diverse classification techniques (like SVM, Decision Tree Algorithm, Logistic Regression and Gaussian Naive Bayes) are applied that show the advantage of feature selection approaches to the best predicting of cervical cancer disease. Furthermore, dimensionality reduction techniques, feature correlation and feature extraction have been used for improving the accuracy of the classifier. A user interface is also developed using flask for API so that a woman can test whether she needs a biopsy test or not.

By above research and study, we found out the most consequential risk factors that can possibly lead to cervical cancer in a woman and thus they should be made aware about various screening tests such as biopsy etc so that this disease can be diagnosed in its precancerous stage and can be easily cured.

**ii**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
|  | **Name**  **Acknowledgement** | **Page No.**  **i** |
|  | **Abstract** | **ii** |
|  | **Table Of Contents** | **iii** |
|  | **List of Figures** | **iv** |
| **Chapter 1** | **Introduction** | **1** |
| **Chapter 2** | **Background Study** | **2** |
| **Chapter 3**  **3.1.**  **3.2.**  **3.3.**  **3.4.** | **Requirement Analysis**  **Hardware**  **Software**  **Functional Requirements**  **Non- Functional Requirements** | **3** |
| **Chapter 4** | **Detailed Design** | **4** |
| **Chapter 5** | **Implementation** | **5** |
| **Chapter 6**  **Chapter 7** | **Experimental results and analysis**  **Conclusion of the report and future scope**  **References**  **Power point presentation** | **6**  **7**  **8**  **9** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**iii**

**List of Figures**

|  |  |
| --- | --- |
| **Fig No.** | **Labels** |
| Fig 4.1 | Steps in Model Building |
| Fig 6.1 | Feature Selection using Feature Importance (WITHOUT FEATURE CORRELATION AND EXTRACTION) |
| Fig 6.2 | Heatmap showing feature correlation |
| Fig 6.3 | Feature Selection using Feature Importance (WITH FEATURE CORRELATION AND EXTRACTION) |
| Fig 6.4 | ROC Curves for Decision Tree for measuring performance of Classifier |
| Fig 6.5 | User interface using Flask (backend) and  HTML and CSS (frontend) |

**List of Tables**

|  |  |
| --- | --- |
| **Table No.** | **Label** |
| Table 6.1 | List of extracted features with their abbreviations |
| Table 6.2 | Depicting Accuracy of Various Models with Oversampling and Under sampling Data |

**iv**

**Introduction**

Cervical cancer is a cancer arising from the cervix. It is due to the abnormal growth of cells that have the ability to invade or spread to other parts of the body. Although it is the most preventable type of cancer, each year cervical cancer kills about 4,000 women in the U.S. and about 300,000 women worldwide. In most cases, cervical cancer symptoms are not noticeable at its early stages. Early on, typically no symptoms are seen. Later symptoms may include abnormal vaginal bleeding, pelvic pain or pain during sexual intercourse. While bleeding after sex may not be serious, it may also indicate the presence of cervical cancer.

Human papillomavirus infection (HPV) causes more than 90% of cases; most people who have had HPV infections, however, do not develop cervical cancer. Other risk factors include smoking, a weak immune system, birth control pills, starting sex at a young age and having many sexual partners, but these are less important. Identifying those factors and building a classification model to classify whether the cases are cervical cancer or not is a challenging research.

Cervical cancer typically develops from precancerous changes over 10 to 20 years. About 90% of cervical cancer cases are squamous cell carcinomas, 10% are adenocarcinoma and a small number are other types. Diagnosis is typically by cervical screening followed by a biopsy. Medical imaging is then done to determine whether or not the cancer has spread.

Worldwide, cervical cancer is both the fourth most common cause of cancer and the fourth most common cause of death from cancer in women. In 2012, an estimated 528,000 cases of cervical cancer occurred, with 266,000 deaths. This is about 8% of the total cases and total deaths from cancer. About 70% of cervical cancers occur in developing countries and 90% of deaths. In low-income countries, it is one of the most common causes of cancer death. In developed countries, the widespread use of cervical screening programs has dramatically reduced rates of cervical cancer.

Studying detection of Cervical Cancer is a need of the hour which not only could save a lot of assets and money but also save the lives of many unfortunate individuals ailed by it.

We are trying to provide a classification technique for better detection of Cervical Cancer and creating way for future projects.

**1**

**Background Study**

Quite an amount of work has been done on the software implementation part of our project. And for our reference we have considered three such research papers:

The first research paper [1] primarily focuses on novel feature screening algorithm by deriving relevance measures from the decision boundary of Support Vector Machines. It alleviates the “independence” assumption of traditional screening methods, e.g. those based on Information Gain and Augmented Variance Ratio, without sacrificing computational efficiency.

The second research paper [2] is a study from different researchers in the last few years. The first contributor showed that transform learning method was the main purpose of the study to show its impact on sharping the accuracy. The second contributor has studied the same data using cost-sensitive classification with Decision Tree classifier. The good results have been reported by the third contributor such that Support vector machine with different approaches were applied to yield almost 94% accuracy. However, cervical cancer disease has been deliberated from different angles. For instance, the fourth contributor pointed out the cervical cancer control in HIV-positive women and expectation of reducing the mortality in 2030.

The third research paper [3] is a study to appraise the perception and utilization of cervical cancer screening services among female nurses in Federal Teaching hospital, Abakaliki, Nigeria i.e. amongst the developing countries. A cross-sectional study was done using semi-structured questionnaires on 408 consenting female nurses. Data was represented using simple percentages, charts, and Chi-square.

**2**

**Requirement Analysis**

## **Hardware**

* **Minimum Specification**
  + 8 GB RAM Memory
  + i3 Intel Processor
  + Graphic Card required
  + An operating system capable of running python environment

**Software**

* Web Browser (Google Chrome, Firefox, Internet Explorer7+)
* Jupyter
* Tensorflow
* Flask
* React
* Heroku
* An operating system capable of running python environment

**Functional Requirements**

* Queries user about the risk factors of cervical cancer and analysing those factors gives result that whether a person should take biopsy test or not.
* A user interface that gives the predicted result whether the user should take biopsy test or not on the basis of the information provided by the user.

**Non-Functional Requirements**

* High Performance: The algorithm should be optimized from time to time to give accurate results.
* Reliable: The system should perform appropriate calculations and generate accurate results.
* Maintenance: The dataset should be maintained and updated as rare cases are identified of cervical cancer so that accurate results are delivered to the user.
* Security: Make sure that the user’s personal information is not obtained without his/her permission and is not leaked to other users.

**3**

**Detailed Design**

The project is implemented in following phases:

**Data Pre-processing:** This involves missing values treatment using standard measurements like the mean for numerical values and median for categorical attributes. There were two features with 92% of missing values which are STDs: Time since first diagnosis and STDs: Time since last diagnosis, so they have been omitted.

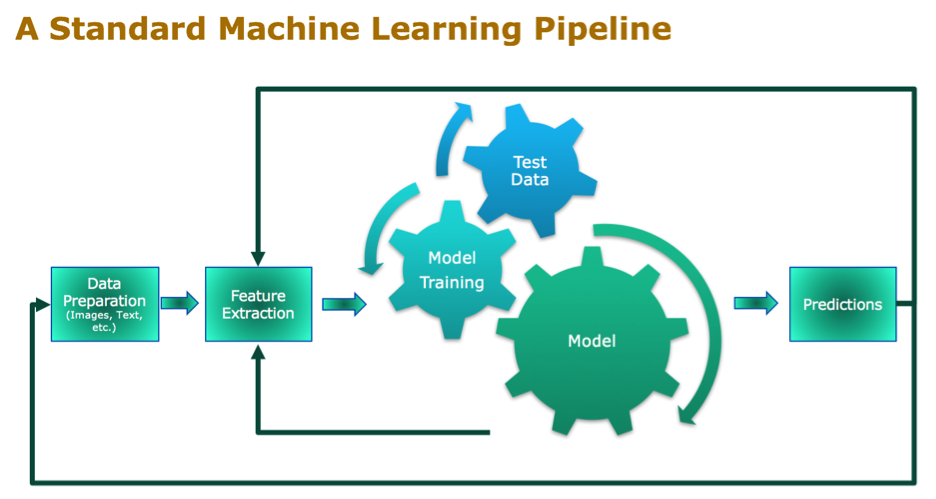
**Handling data imbalance and feature correlation:** Imbalanced data have been studied by applying two techniques, over-sampling and under sampling. Eventually, for sharping the results and looking at the main risk factors of the cervical cancer, feature selection methods are applied like wrapper methods. Wrapper methods such as Sequential Feature Selector (backward version) and Recursive Feature Selector are used. Heatmap is plotted to observe correlation.

**Feature extraction:** By correlating various parameters, some new features are extracted from the data such as: ''YRSS: Years passed since patient had first sexual intercourse NSPP: Number of sexual partners since first time as a percentage. HPA: Hormonal Contraceptives/age TPS: Total packets of cigarettes smoked NPA: Number of pregnancies/age NSA: Number of sexual partners/age NYHC: number of years patient did not take Hormonal Contraceptives APP: number of pregnancy/numbers of sexual partner NHCP: number of years patient took Hormonal Contraceptives after first sexual intercourse as a percentage.

**Model selection:** For finding the suitable classification model, five classifiers including the Logistic regression, K nearest neighbor classifier, Gaussian Naïve Bayes, SVM and Decision Trees are applied and their accuracies are compared.

**Web application:** With the implementation of cervical cancer screening programs, the incidence of cervical cancer and mortality from the disease can be reduced. However, being up to date on the guidelines can be challenging. Thus, we designed a user interface for that can return the predicted result by applying the decision tree classification model on the user provided information. This ML- based web application aims at main user’s requirements, expectations, acceptance and usefulness of a patient support tool.

**4**

  
Fig 4.1: Steps in Model Building(Reference:<https://www.datanami.com/2018/09/05/how-to-build-a-better-machine-learning-pipeline/>)

**4**

**Implementation**

In this project, we will analyzing different Machine Learning models and predict the best one to detect if one is at a risk of having cervical cancer. The models which we will be using are Logistic regression neighbor classifier, Gaussian Naïve Bayes, SVM and Decision Trees.

The project starts with collection of dataset and then applying the different Machine Learning models to predict the accuracy and then determining the best model to diagnose Cervical Cancer. The dataset has been collected from University of California, Irvine repository.

The features we are working on are as follows:

**Data set description:**

The important features used are: Age, First sexual intercourse, number of pregnancies, Number of sexual partners, Smokes, Hormonal Contraceptives and IUD.

AGE Fifty per-cent of cervical cancer diagnoses occur in women ages. Numerous studies report that high poverty levels are linked with low screening rates. HIGH SEXUAL ACTIVITY Human papilloma virus (HPV) is the main risk factor for cervical cancer. In adults, the most important risk factor for HPV is sexual activity with an infected person. Women most at risk for cervical cancer are those with a history of multiple sexual partners, sexual intercourse at age 17 years or younger, or both. Sexual activity with multiple partners increases the likelihood of many other sexually transmitted infections (chlamydia, gonorrhoea, syphilis). USE OF ORAL CONTRACEPTIVES: Studies have reported a strong association between cervical cancer and long-term use of oral contraception (OC). Women who take birth control pills for more than 5 - 10 years appear to have a much higher risk HPV infection (up to four times higher) than those who do not use OCs. HAVING MANY CHILDREN: Studies indicate that having many children increases the risk for developing cervical cancer, particularly in women infected with HPV. SMOKING: Smoking is associated with a higher risk for precancerous changes (dysplasia) in the cervix and for progression to invasive cervical cancer, especially for women infected with HPV.

We have then preprocessed this data and applied methods to balance the dataset. For analyzing the most important features responsible for cervical cancer we correlated the features using heatmap. This also helped us in extracting new features from the data.

We then analyzed the performance of the classifiers through ROC (**Receiver operating characteristic**) curves. Decision Tree and KNN give highest accuracy of around 93%. For the flask-based user interface decision tree classification model is used for predicting whether a user needs biopsy test or not on the basis of information provided by her.

**5**

**Experimental Results and Analysis**

Fig 6.1: Feature Selection using Feature Importance (WITHOUT FEATURE CORRELATION AND EXTRACTION)

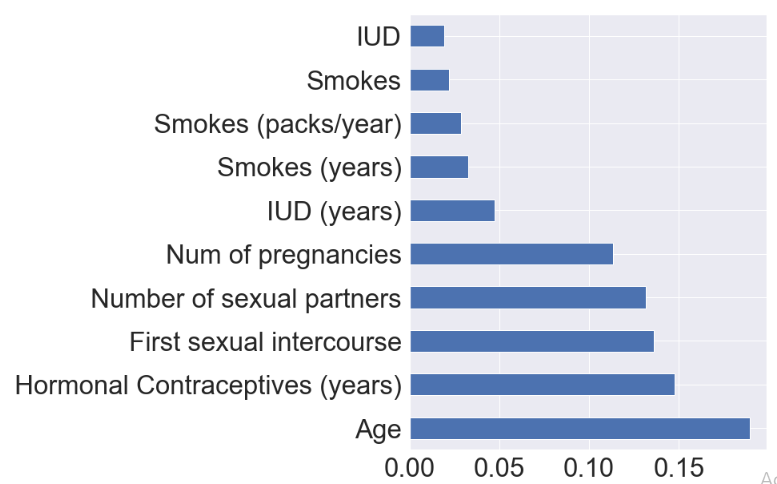
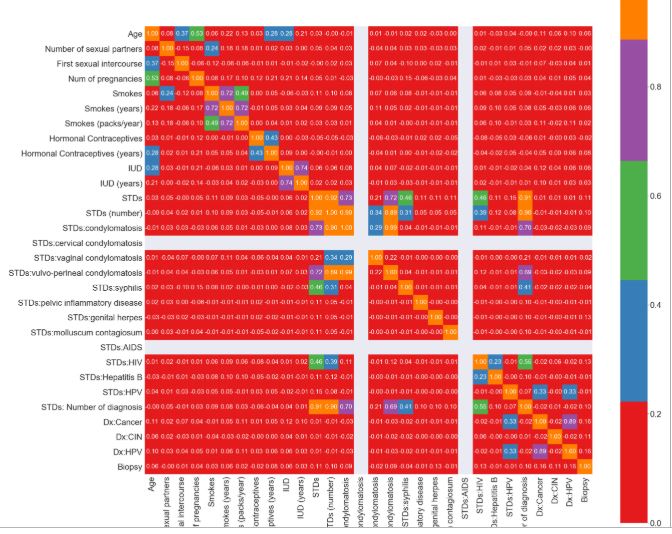


Fig 6.2: Heatmap showing the feature correlation

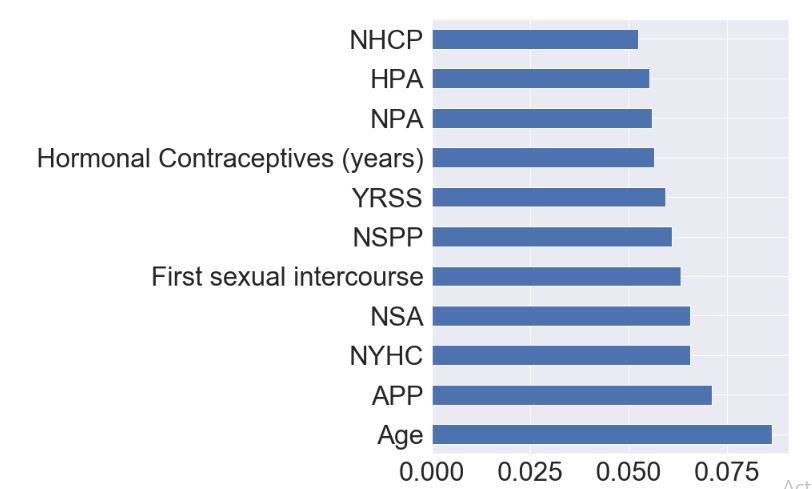


**6**

Table 6.1: List of extracted features with their abbreviations

|  |  |
| --- | --- |
| **ABBREVIATIONS** | **FULL FORM** |
| YRSS | Years passed since patient had first sexual intercourse |
| NSPP | Number of sexual partners since first time as a percentage |
| HPA | Hormonal Contraceptives/age |
| TPS | Total packets of cigarettes smoked |
| NPA | Number of pregnancies/age |
| NSA | Number of sexual partners/age |
| NYHC | number of years patient did not take Hormonal Contraceptives |
| APP | number of pregnancy/numbers of sexual partner |
| NHCP | number of years patient took Hormonal Contraceptives after first sexual intercourse as a percentage''' |

Fig 6.3: Feature Selection using Feature Importance (WITH FEATURE CORRELATION AND EXTRACTION)



**6**

Table 6.2: Depicting Accuracy of Various Models with Oversampling and Under-sampling Data

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MODEL** | **SVM** | **GNB** | **LR** | **DT** | | **KNN** |
|  |  |  |  | **GINI** | **ENTROPY** |  |
| **ACCURACY** | 93.604 | 9.883 | 93.023 | 93.604 | 92.441 | 93.604 |
| **ACCURACY WITH UNDERSAMPLING** | 14.534 | 84.302 | 71.511 | 43.023 | 51.744 | 63.372 |
| **ACCURACY WITH OVERSAMPLING** | 13.372 | 9.8837 | 82.588 | 46.511 | 76.511 | 80.2325 |
| **AUC\_ROC** | 0.5 | 0.4339 | 0.4968 | 0.5 | 0.61 | 0.5 |
| **ROC\_AUC** | 0.536 | 0.547 | 0.5533 | 0.61 | 0.51 | 0.5728 |

Fig 6.4: ROC Curves for Decision Tree Classifier for measuring performance of Classifier

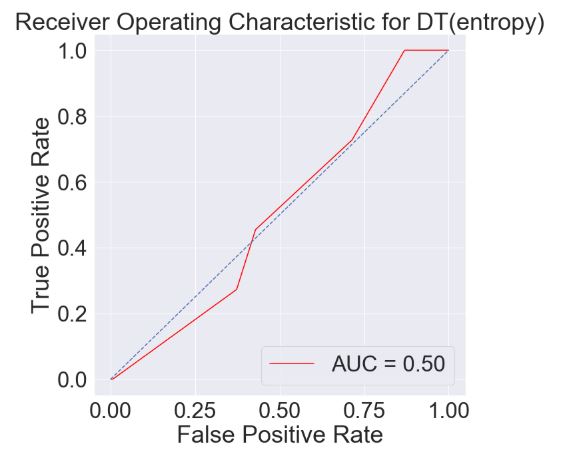


Fig 6.4 (a)

**6**

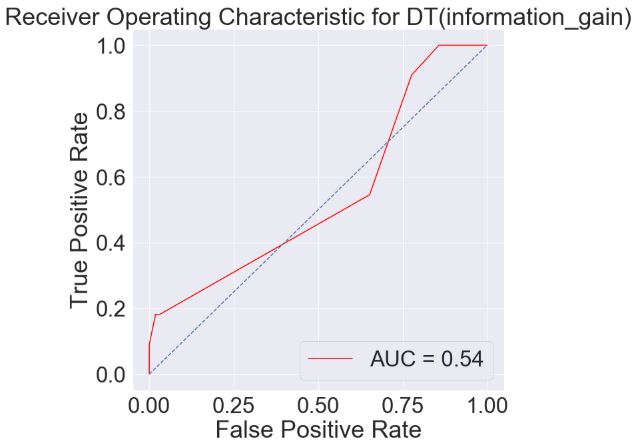


Fig 6.4 (b)

Fig 6.5: User interface using Flask (backend) and HTML and CSS (frontend)



Fig 6.5 (a)

**6**

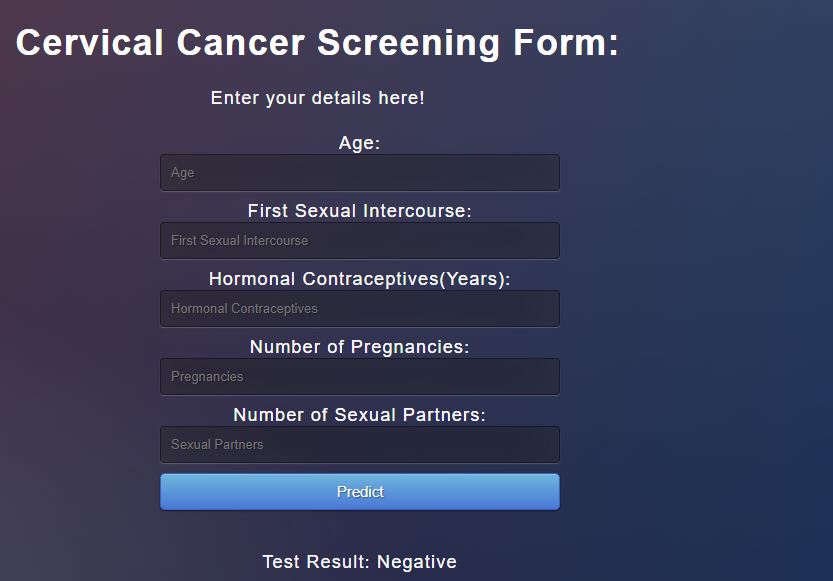


Fig 6.5 (b)

**6**

**Conclusion and future scope**

We have presented a comparison between different machine learning classifiers respect to the best predictive model for Cervical Cancer Dataset. Results show that this data is biased and addressing the imbalanced data is the first step for evaluation. Two techniques have been used to address the imbalanced data; over-sampling and under-sampling. Over-sampling yields better results in case of Gaussian Naïve Bayes than other methods due to higher accuracy obtained by over-sampling while for other classifiers accuracy decreases. Further studies are conducted by using feature selection methods. Filter and wrapper methods have been applied and the selective features are Age, First sexual intercourse, number of pregnancies, Number of sexual partners, Smokes, Hormonal Contraceptives and IUD. Decision Tree and K-Nearest-Neighbour classifier give best results for this dataset with accuracy of around 93%.

Though various screening tests are available for detection of precancerous stage but to keep up with the guideline is a difficult task. Therefore, we have deployed our decision tree classification model using Flask and created a user interface that predicts the result whether patient needs biopsy or not.

As a future work, multi-class classification with the four targets can be done to explore more parameters. Also, most of the studies that have been proposed the last years focus on the development of predictive models using supervised ML methods and classification algorithms aiming to predict valid disease outcomes. Based on the analysis of their results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and classification can provide promising tools for inference in the cancer domain.

There are certain limitations to machine learning model we are using i.e. Decision Tree, the same are mentioned below:

**Limitations of Decision Tree:**

* Instability: They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree.
* Complexity: Among the major decision tree disadvantages are its complexity. Decision trees are easy to use compared to other decision-making models, but preparing decision trees, especially large ones with many branches, are complex and time-consuming affairs.
* Possibility of duplication with the same sub-tree on different paths.

**7**

**References**

**Proceedings paper:**

[1] Zhang J., Liu Y. (2004) “Cervical Cancer Detection Using SVM Based Feature Screening.” In: Barillot C., Haynor D.R., Hellier P.(eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2004. MICCAI 2004. Lecture Notes in Computer Science, vol 3217. Springer, Berlin, Heidelberg.

**Proceedings paper:**

[2] Y. M. S. Al-Wesabi, Avishek Choudhury, Daehan Won (2018) “Classification of Cervical Cancer Dataset” Binghamton University, USA Proceedings of the 2018 IISE Annual Conference K. Barker, D. Berry, C. Rainwater, eds.

**Journal Article:**

[3] C. C. Anikwe and C. C. Ifemelumma (2017) “Cervical Cancer Screening: Assessment of

## Perception and Utilization of Services among Health Workers in Low Resource Setting

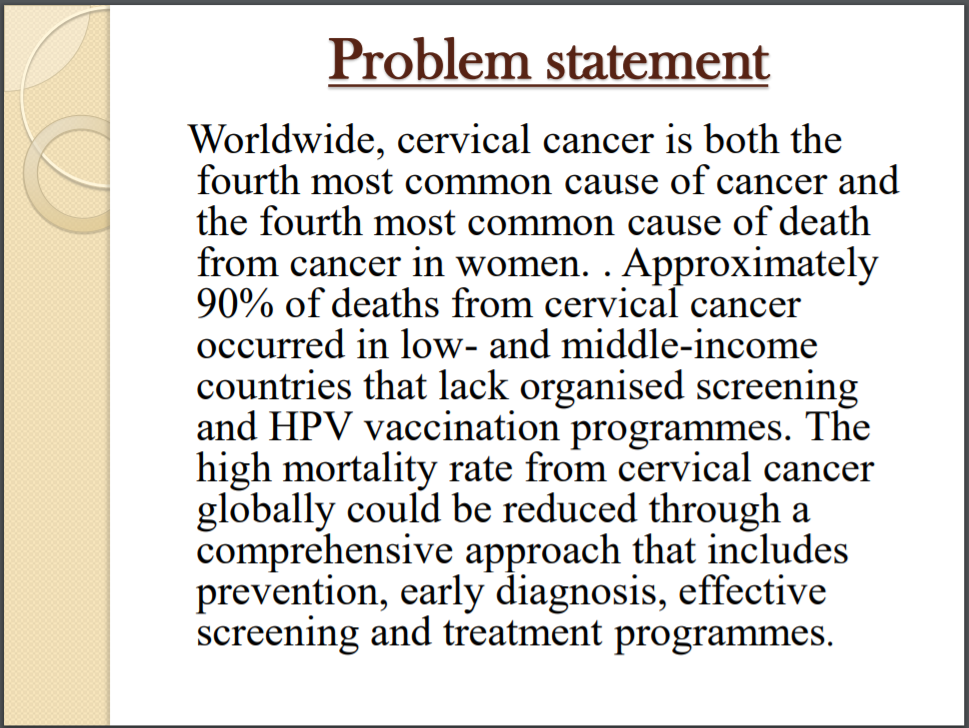
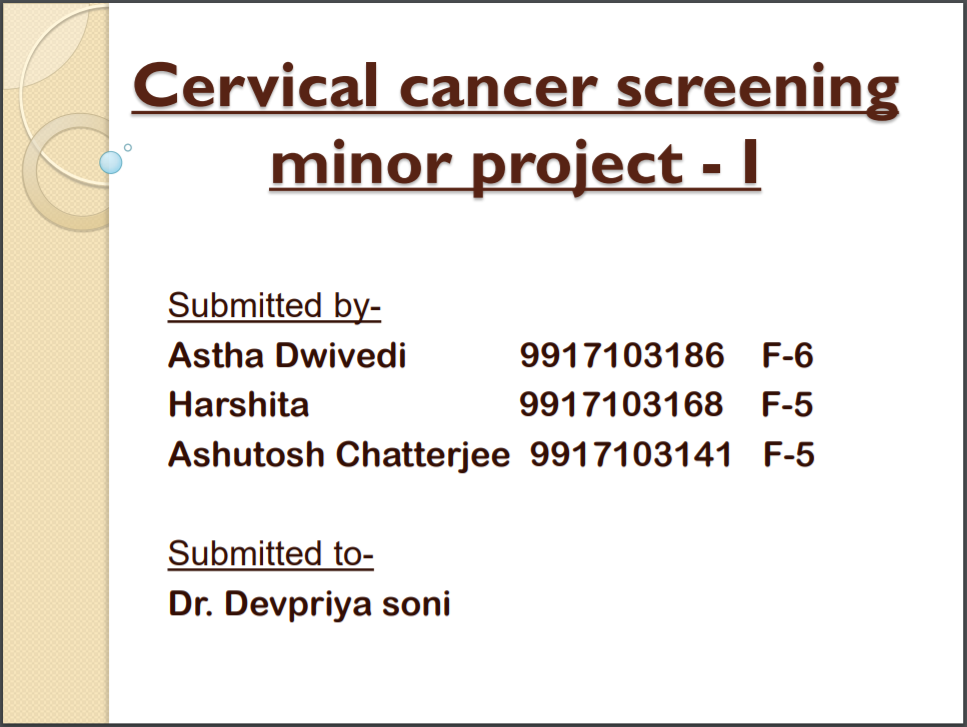
## International Journal of Reproductive Medicine, Volume 2019, Article ID 6505482

**Online:**

[4]DATASET\_SOURCE: [https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29#](https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29)

**8**

**Power Point Presentation**

****

**9**

