HELWAN UNIVERSITY
Faculty of Computers and Artificial Intelligence
Medical Informatics Program

[CERVECTOR (CERVICAL CANCER DETECTOR)]

A graduation project dissertation by:

[Martina Adel Lamei Abdelnoor (20208185)]

[Hanan Natay Fawzy Abdelmalak (20208091)]

[Eriny William Makram (20208024)]

[Yasmine Ali Mohammed Ali (20208294)]

[Mohammed Ayman Mohamed Barghash (20208191)]

Submitted in partial fulfilment of the requirements for the degree of Bachelor of Science in Computers & Artificial Intelligence, at the Medical Informatics Program, the Faculty of Computers & Artificial Intelligence, Helwan University

Supervised by:

[Dr. Marwa Abdelfattah]

June 2023

جامعة حلوان
كلية الحاسبات والذكاء الإصطناعي
برنامج المعلوماتية الطبية

[تشخيص سرطان عنق الرحم بواسطة تعلم الآلة]

رسالة مشروع تخرج مقدمة من:

[مارتينا عادل لمعى عبدالنور (20208185)]

[حنان نتعى فوزى عبدالملاك (20208091)]

[ارينى وليم مكرم (20208024)]

[ياسمين على محمد على (20208294)]

[محمد ايمن محمد برغش (20208191)]

رسالة مقدمة ضمن متطلبات الحصول على درجة البكالوريوس في الحاسبات والذكاء الإصطناعي، ببرنامج المعلوماتية الطبية كلية الحاسبات والذكاء الإصطناعي، جامعة حلوان

تحت إشراف:

(د: مروة عبدالفتاح)

يونيه / تموز 2023

ACKNOWLEDGEMENT

"We would like to express our sincere appreciation to everyone who has contributed to the successful completion of this project. Firstly, we are grateful to our supervisor for providing us with guidance and support throughout the project. We would also like to thank the medical community for providing the datasets used in this study.

We would like to acknowledge the support of our colleagues who provided us with valuable insights and feedback. We would also like to thank our families and friends for their unwavering support and encouragement.

Finally, we would like to acknowledge the power of machine learning, which has enabled us to develop a model for the classification of cervical cancer. This project has been an incredible learning experience, and We are grateful for the opportunity to have worked on it."

TABLE OF CONTENTS

CONTENTS

LIST OF FIGURES

ABSTRACT

Cervical cancer is a major cause of mortality among women worldwide. Early detection and diagnosis of cervical cancer can significantly improve patient outcomes. Classification of a pap smear is a manual and time-consuming task (Martin, 2003). This thesis deals with the classification of features extracted from single-cell pictures of Pap smears. In this project, we propose a machine learning-based approach for classifying cervical cancer. Our objective is to develop an accurate and efficient model that can classify cervical cell images as normal or abnormal and their types we also proposed a machine learning Prediction Model that predicts the result of a Biopsy test and thereby confirms the presence/non-presence of cervical cancer in the patients. We use a dataset pap smear of cervical cell images obtained from the publicly available Herlev University repository.

([21] HErlev (HErlev Pap Smear Dataset) Herlev University Hospital (Denmark), 2005). The dataset contains a total of 917 images, with 242 normal and 675 abnormal images. We preprocess the images to extract features using techniques such as converting images to grayscale, morphological opening, thresholding, labeling regions, and extracting region properties. We then train and evaluate several machine learning models such as Support Vector Machines (SVM), K- Nearest Neighbors (KNN) we also, developed a website to help women calculate their risk of developing cervical cancer using a Logistic regression machine learning model based on the Risk factors dataset obtained from the UCI Irvine machine learning repository ([22] Fernandes, 2017) The dataset comprises demographic information, habits, and historic medical records of 858 patients. Our results show that the SVM model outperforms the other models, achieving an overall accuracy of 95% on the test set for pap smear images and logistic

regression for the risk classifier with an accuracy of 95.2381%. The high accuracy of the model indicates its potential for use in clinical practice as an automated screening tool for cervical cancer. In conclusion, our project demonstrates the feasibility of using machine learning techniques to classify and predict cervical cancer and highlights the potential for developing automated screening tools that can improve early detection and diagnosis of cervical cancer.

Keywords**:** Cervical cancer automated detection, Classification, Data preprocessing, Feature extraction, K- Nearest Neighbors, Machine learning, Performance evaluation, Support vector machines.

LIST OF ABBREVIATIONS

CC: Cervical cancer

ML: Machine learning

SVM: Support vector machine

KNN: K-Nearest Neighbor

TP: True positive

TN: True negative

FP: False positive

FN: False negative

CV: Cross-validation

ACC: Accuracy

FPR: False positive rate

CNNs: convolutional neural networks

GLOSSARY

Cervical cancer: A type of cancer that originates in the cervix, which is the lower part of the uterus that connects to the vagina.

Pap smear: A procedure in which a small brush is used to gently remove cells from the surface of the cervix and the area around it so they can be checked under a microscope for cervical cancer.

Machine learning: A subset of artificial intelligence that involves the use of algorithms and statistical models to enable machines to learn from data and make predictions or decisions without being explicitly programmed.

Classification: A type of supervised learning problem in which the goal is to assign input data to one of several predefined categories based on their features.

Feature extraction: The process of selecting or extracting relevant features from raw data that can be used to train a machine learning model.

Feature selection: The process of selecting a subset of features from a larger set of features that is most relevant to the prediction task.

Model selection: The process of choosing the best model among a set of candidate models based on their performance on a validation set.

Performance evaluation: The process of measuring the accuracy and other performance metrics of a machine learning model on a test dataset.

Normal vs. abnormal classification: A binary classification problem in which the goal is to classify a sample as normal or abnormal based on its features.

Support vector machines (SVMs): A type of machine learning algorithm that constructs a hyperplane or set of hyperplanes in a high-dimensional space to separate different classes of data.

k-nearest neighbors algorithm: a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.

CHAPTER ONE
INTRODUCTION

Cervical cancer is the fourth most common female malignancy worldwide and represents a major global health challenge and the leading cause of cancer death among women in developing countries (Paul A Cohen MD a b, 2019). Early detection and treatment are key to a good outcome, but many women do not get screened for cervical cancer. In recent years, there has been a growing interest in using machine learning to improve the early detection and prediction of cervical cancer. Machine learning algorithms can be used to analyze Pap smear images and identify abnormal cells that may be indicative of cancer. This can help to identify cancer at an earlier stage when it is more treatable. This dissertation explores the use of machine learning to predict cervical cancer from Pap smear images. A variety of machine learning algorithms, including support vector machines (SVMs) and k-nearest neighbors (KNNs), were used. The results showed that these algorithms were able to accurately predict cervical cancer from Pap smear images.

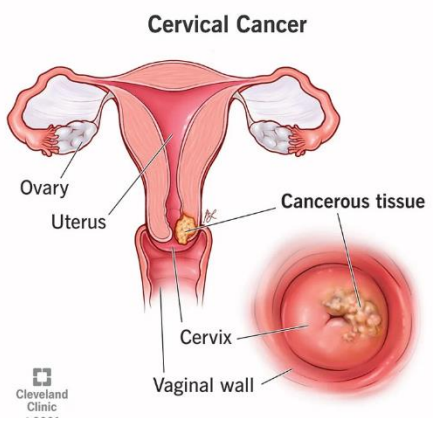


FIGURE 1 CANCEROUS TISSUES FORMING IN THE CERVIX. (CLEVELAND CLINIC)

In addition to developing a machine learning model to predict cervical cancer, this dissertation also developed a website to help women understand the risk of cervical cancer and to get screened. The website provides information about cervical cancer, risk factors, symptoms, and screening options. It also allows women to calculate their risk of developing cervical cancer using a Logistic regression machine learning model based on medical history and lifestyle habits.

Overall, this research has made significant progress in the development of a more accurate and efficient way to predict cervical cancer. By continuing to invest in research in this area, we can help to save lives and improve the quality of life for women around the world.

1.1. OVERVIEW:

This dissertation explores the use of machine learning to predict cervical cancer from Pap smear images. A variety of machine learning algorithms, including support vector machines (SVMs) and k-nearest neighbors (KNNs), were used. A website was also developed that would allow women to use the machine learning model to predict cervical cancer from their medical history and lifestyle habits. The results of this study showed that machine learning can be used to predict cervical cancer from Pap smear images and medical history with a high degree of accuracy

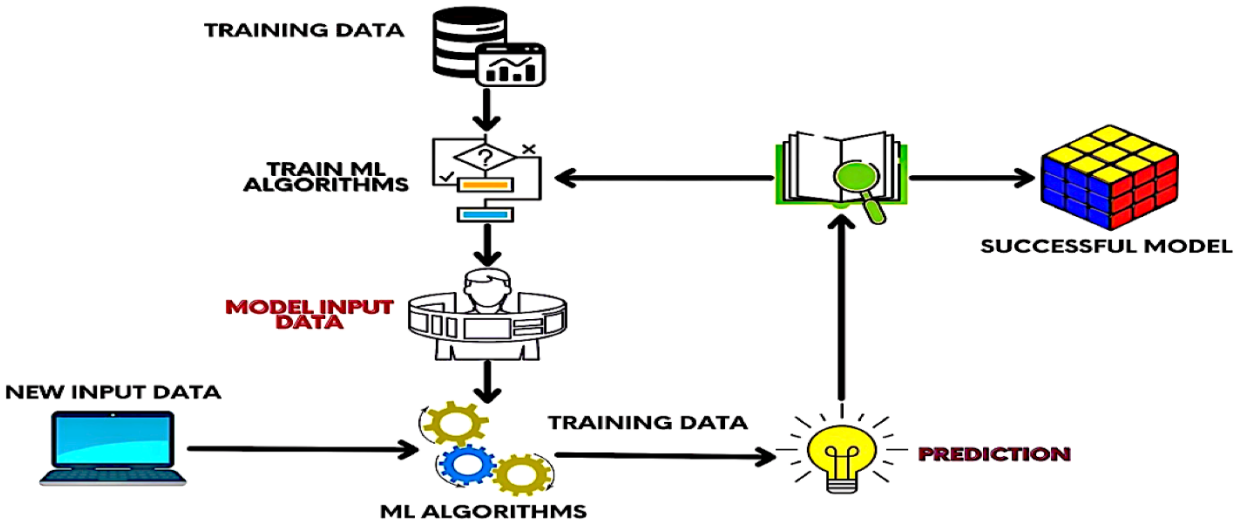


FIGURE 2 PROPOSED RESEARCH MODEL FOR CLASSIFYING CERVICAL CANCER.

1.2. PROBLEM STATEMENTS:

The process of classifying cells from uterine cervix pap smear is a manual and time-consuming task, which relies on the expertise of skilled cyto-technicians (Martin, 2003). The classification of cells as normal or abnormal can be challenging as cytologists faces difficulty in deciding whether the cell are normal or not, leading to false-negative results and delayed diagnoses for women with cancer. Furthermore, the frequent nature of the checkup procedure results in a large amount of data that requires prompt diagnosis. The reliance on skilled cyto-technicians can also be costly and hard to find. Therefore, there is a need for automated classification methods to improve the efficiency and accuracy of the screening process, while also reducing the reliance on skilled technicians. This project aims to develop and compare automated feature extraction and machine learning algorithms for classifying pap smear cells, with the goal of achieving a less than 5% false-positive and false-negative rate, as it is periodically check, so it is repeated frequently which will results in very large data to be diagnosed manually in time.

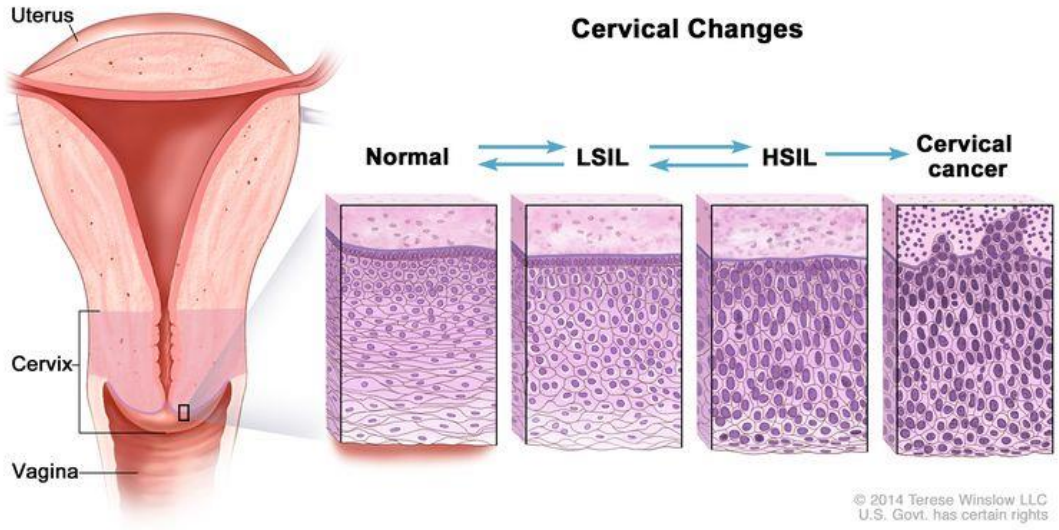


FIGURE 3 CERVICAL CHANGES (NIH CANCER INSTITUTE)

Many women do not get screened for cervical cancer due to concerns about the invasiveness of the procedure. This is a problem because early detection is critical for successful treatment outcomes. To address this issue, a website will be developed to encourage women to get screened for cervical cancer by allowing them to predict their risk of developing the disease. By giving women the tools to assess their risk, the website aims to increase awareness about the importance of screening and encourage women to take action to protect their health.

1.3. SCOPE AND OBJECTIVES:

The scope of this project is to develop a machine learning model to classify Pap smear images into normal and abnormal and types of abnormal cells. The model will be trained on a dataset of Pap smear images that have been classified by human experts. The goal is to develop a model that can classify Pap smear images with an accuracy of at least 95%. And risk factors predictor model which will till the women if she is at risk of cervical cancer or not.

- **Objectives**:

The specific objectives of this project are to:

1. Collect a dataset of Pap smear images that have been classified by human experts.
2. Extract features from the images.
3. Train a machine learning model to classify Pap smear images into normal and abnormal.
4. Improve model's accuracy.
5. Train a machine learning model to predict the risk of developing cervical cancer.

6. Develop a web application that allows users to know if they are at risk of cervical cancer.

- **Significance:**

The significance of this project is that it could lead to the development of a more accurate and reliable method of Pap smear screening. This could help to reduce the number of women who are diagnosed with cervical cancer and improve the chances of survival for those who are diagnosed. The website will allow women to determine whether they are at risk of cervical cancer from their homes.

- **Limitations:**

This project is limited by the availability of data. The dataset of Pap smear images that has been collected is relatively small. This means that the model may not be able to generalize well to new data. Additionally, the model will only be as accurate as the human experts who classified the images in the dataset.

1.4 REPORT ORGANIZATION (STRUCTURE):

The report will be organized as follows:

- **Chapter One: (Introduction**) The introduction will provide background information on cervical cancer and Pap smear screening. It will also discuss the limitations of current Pap smear screening methods and the potential of machine learning to improve accuracy.
- **Chapter Two: (Literature Review)** The literature review will discuss the current state of research on machine learning for Pap smear screening. It will also discuss the different machine learning algorithms that have been used for this task.

- **Chapter Three: (Methods)** The methods section will describe the dataset that was used, the machine learning algorithm that was trained, and the evaluation metrics that were used.
- **Chapter Four: (Results)** The results section will present the accuracy of the machine learning model on the training dataset and the test dataset. It will also discuss the performance of the model on different types of Pap smear images.
- **Chapter Five: (Discussion)** The discussion section will discuss the implications of the results and the limitations of the study. It will also discuss future work that could be done to improve the accuracy of the machine learning model.
- **Conclusion**: The conclusion will summarize the main findings of the study and discuss the implications for future research.

1.5. WORK METHODOLOGY:

1.5.1. DATA COLLECTION:

The first step in the research methodology is to collect the data that will be used in the study. In this case, the data that will be collected is Pap smear images. The images are obtained by skilled cyto-technicians using a microscope connected to a frame grabber. The images are taken with a resolution of 0.201μm/pixel. The cyto-technicians have classified all the images into diagnostic categories. Every image is classified by two different cyto-technicians, to ensure the accuracy of the classification. The images on which the cyto-technicians agree are used to create the dataset (Martin, 2003). The dataset was created by Herlev University. and the risk factors dataset from UCI Irvine machine learning repository.

Data Preprocessing:

The next step in the work methodology is to preprocess the data. This involves cleaning the data and removing any errors or inconsistencies, so that it can be used by machine learning algorithms.

1.5.2. FEATURE EXTRACTION:

The next step in the work methodology is to extract features from the data. Features are the characteristics of the data that will be used to train the machine learning models. The features that will be extracted from the Pap smear images using techniques such as converting images to grayscale, morphological opening, thresholding, labeling regions, and extracting region properties.

Machine Learning:

The next step in the work methodology is to train the machine learning models. The machine learning models will be trained on the features that were extracted from the data. The machine learning models that will be used in this study are support vector machines (SVMs) and k-nearest neighbors (KNNs) for classifying the pap smear images.

Logistic regression, RandomForestClassifier, GaussianNB are used for prediction of the risk factors.

1. **Evaluation:**

The next step in the work methodology is to evaluate the machine learning models. The models will be evaluated on a held-out set of data. The held-out set of data is a set of data that was not used to train the models. The models will be evaluated on their accuracy, sensitivity, and specificity. Accuracy is the percentage of cases that were correctly classified.

2. Deployment:

The final step in the work methodology is to deploy the machine learning models. The models will be deployed on a website, it will encourage women to get screened for cervical cancer by allowing them to predict their risk of developing the disease. By giving women the tools to assess their risk, the website aims to increase awareness about the importance of screening and encourage women to take action to protect their health.

The website is built using Flask, a Python web framework. Flask is based on the Werkzeug WSGI toolkit and the Jinja2 template engine. For the front end, we will use HTML and CSS.

1.6 Work Plan (Gantt chart):

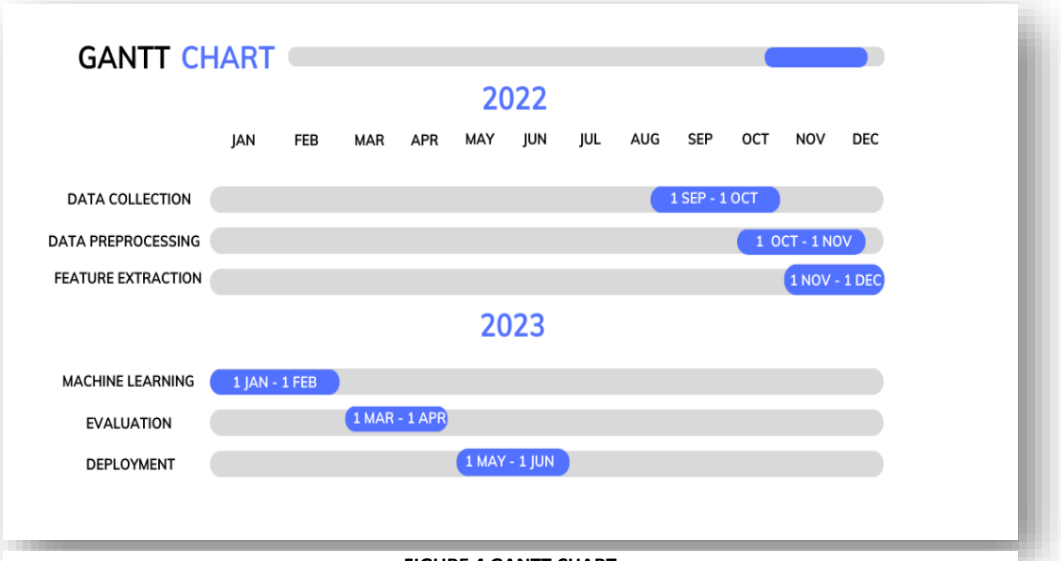


FIGURE 4 GANTT CHART

Start Date: September 2022 **End Date: May 2023**

Tasks:

Data Collection	(September - October 2022)
Data Preprocessing	(October - November 2022)
Feature Extraction	(November - December 2022)
Machine Learning	(January - February 2023)
Evaluation	(March - April 2023)
Deployment	(May 2023)

CHAPTER TWO

RELATED WORK (LITERATURE REVIEW)

2.1 BACKGROUND:

Cervical cancer is one of the most common forms of cancer affecting women worldwide. Early detection is key to successful treatment and improving patient outcomes (Paul A Cohen MD a b, 2019). Currently, the most common method for cervical cancer detection is through Pap smear tests, which involve collecting cells from the cervix and examining them under a microscope. However, this method can be time-consuming, expensive, and requires trained professionals to interpret the results.

Machine learning has the potential to improve the accuracy and efficiency of cervical cancer detection by automating the analysis of cervical smear images. Machine learning algorithms can be trained to identify patterns in images that are indicative of cervical cancer, allowing for faster and more accurate detection.

Recent research has explored the use of machine learning techniques such as deep learning, support vector machines, and KNN for cervical cancer detection. These methods have shown promising results in identifying cancerous cells with high accuracy and efficiency.

However, there are still challenges that need to be addressed in the development of machine learning-based cervical cancer detectors. These include the need for large and diverse datasets, the need for interpretable and explainable models, and the need for real-world validation of the models.

Despite these challenges, the potential benefits of machine learning in cervical cancer detection are significant. By improving the accuracy and efficiency

of cervical cancer detection, machine learning can help to save lives and improve the quality of life for women around the world.

2.2 LITERATURE SURVEY:

Many projects have been proposed to explain how to diagnose and classify different types of cervical cancer Although the literature covers a wide variety of such projects from older to the more recent They used clinical features-based approach, genetic features-based approach and image classification and segmentation to detect and diagnose the presence of cervical cancer. All the above methods use different classification and segmentation techniques to enhance accuracy and to minimize the classification errors of false positive and false negative records and to identify the most related risk factors of cervical cancer.

The following papers focus on clinical feature-based approaches and related works are represented in the next paragraph.

In 1940 Smear Test was first demonstrated by the scientist George Papanicolaou.

Pap test helps in detecting precancerous changes in the cervical cells. Pap's Smear test is of 2 types

1) Conventional 2) Thin preparation.

Both techniques are different in the way the sample is obtained.

In the year 2015 total number of 1,22 500 cases of cervical cancer were detected and out of them, 67,400 lost their life (1). According to the latest census, female population aged 15 years and more is 432 million in India and this is the number which is at risk of acquiring cervical cancer.

(Martin, Pap-Smear Classification at Technical University of Denmark, 2003)

This project between Herlev University Hospital's department of pathology and the commercial company DIMAC. has been formed around a dataset made by Herlev hospital. Herlev hospital has now made a new database, for which classification results are wanted. This thesis is based on single cell pictures of papsmear cells from the uterine cervix. Data collected in the thesis of Byriel (1999) is used, and is called the old data(500 cells). Furthermore, a new single cell papsmear database created by cytotechnicians at Herlev Hospital, and referred to as the new data(917 cells). First step was segmentation of each picture into background, cytoplasm and nucleus to make the later feature extraction possible

Second step was Feature extraction is used to find features that possibly can help in the classification.

Third and last step was the classification they proposed 2 approaches, direct and Hierarchical classification. Two clustering algorithms were chosen for the classification, FCM and GK,

For supervised clustering. All results on the old data was acceptable and below 5% error, except for

The unsupervised GK. The results obtained for HCM were worse than FCM as anticipated, but better than GK on the old data. The supervised clustering showed much better results than unsupervised clustering,

And supervised clustering, respectively. All results with the new data had an overall error above 5%. HCM, FCM and GK could not give acceptable results. The worst results were obtained with HCM. FCM and GK gave about the same error for supervised clustering, but GK gave the best results for unsupervised clustering.

Supervised clustering again showed to give much better results than the unsupervised clustering.

(Pap-Smear Classification using efficient second order neural network training algorithms, 2004) we utilized two highly efficient second order neural network training algorithms, namely LMAM and OLMAM, for the construction of an efficient pap-smear test classifier. Performance comparisons were included in the paper between them, such as Gustafson-Kessel clustering techniques, hard c-means, fuzzy c-means, entropy-based intuctive machine learning, genetic programming and finally, hybrid intelligence methods combining feature selection and clustering techniques.

The proposed algorithms manage to build very efficient pap-smear classifiers under various parameter settings. The best performance of the proposed approach is obtained when OLMAM methodology with 10 hidden nodes is applied, with all (20) features used to build the classifier. For the case of the OLMAM methodology applied with 9 hidden nodes and 20 features used ,the overall classification accuracy for the two class category problem becomes the maximum obtained ever, reaching up to 98.86%.

The most competitive approach in literature for the same problem is the application of a hybrid intelligent approach consisting of feature selection and supervised fuzzy c-means. Best performance for the full-problem classification are (a) the application of standard genetic programming with an overall accuracy of 80.7% and (b) the application of a hybrid intelligent scheme, consisting of feature selection and hierarchical classification as suggested in [20], with an overall correct classification accuracy of 80.5%. Concluding, a clear trade-off seems to exist between the classification accuracy and the comprehensibility of the acquired

output of different computational intelligence methodologies in the pap-smear diagnosis problem.

In 2005 from an existing database of single pap-smear cells with 20 already extracted features, a detailed stand-alone description has been given. Using the database, a simple worksheet

with features and class description along the columns and samples along the rows. The introduction of local transductive methods, applied on the pap-smear data, has improved the overall classification error for some methods compared to previous results. But not all transductive methods improved the results. Simple transductive methods like K Nearest Neighbors(KNN) actually performed worse than the simple inductive Least Square Method(LS). Adding on different weights to KNN method did not improve the results considerably. The optimal classifier tested was Nearest Class gravity Center(NCC) achieving an overall error of 5.1%. An simple method introduced to compensate for pre-dominance of single classes in the data-set. The newly introduced Neuro-Fuzzy Inference Method for Transductive Reasoning(NFI) was applied and showed partly satisfactory results. With an achieved error of 5.67% existing classification results was improved, but on the other hand the results was beaten by the much more simple NCC method. Both NCC and NFI improved earlier optimal classification results of 6.06% for the overall error.

Regrettably, no feature selection was applied in this current project. But previous pap-

Smear projects have showed large improvement using feature selection. Extending the tested methods with feature selection the satisfactory results above could improve.

(Dounias, 2006) initially use a genetic algorithm for feature subset selection. They then use a number of variants of the nearest neighbour classification method (1-Nearest Neighbor, k-Nearest Neighbor, wk-Nearest Neighbor) in the classification phase of the proposed approach.

In another paper, Marinakis and Dounias (2006b) propose a tabu search algorithm for the solution of the feature selection problem. The algorithm is then combined with a number of nearest neighbor based classifiers.

In a third paper, Marinakis and Dounias (2006c) use an ant colony optimization (ACO) methodology, an approach derived from the foraging behaviour of real ants in nature. The method in fact models the problem as the search for a minimum cost path in a graph.

(Reif, 2006) described RF efficacy in various model genetic and proteomic datasets. RF progress fails to classify related traits based on genetic data and proteomics datasets. Experimental findings indicate that using several data sources is useful where the disease definition is uncertain, and the corresponding data basis for the phenotypic outcome is unknown. This study's findings indicate that RF is exceptional for detecting high-dimensional data vector characteristics with minimal main effects and low heritability, but the problem faced by RF chooses only one attribute at each tree split during construction, strictly epistatic.

In 2013, Tseng et al presented three classification models C5.0, support vector machine and extreme machine learning to predict cervical cancer recurrence and to find the most related risk factors by using clinical dataset as, the age, radiation therapy, cell type, tumor grade and tumor size. The dataset was collected from the Medical University Hospital of Chung Shan which contained 168 cases with 12 features for each case. The experiments identified two risk factors related

to recurrence of cervical cancer which were cell type and radiation therapy. Also, the experiment results showed that C5.0 got the highest classification accuracy ratio compared with the other classifiers.

(et, 2014) presented a predictive model using multiple logistics regression analysis and artificial neural network to predict the presence of cervical cancer and to identify the most risk factors related to cervical cancer. The authors used a combination of demographic information and blood samples of 270 cases (68 cervical cancer records and 202 healthy women records). The dataset was collected from the Hospital of Women and Children of Wufeng country. They used features like HPV, 4 genetic factors and educational level. The experiment identified two risk factors which are HLA DRB1∗13-2 and HLA DRB1∗3-17 alleles. These risk factors were the reasons of increasing the risk of developing cervical cancer disease. Paper results showed that the back-substitution fitting of artificial neural network got the highest classification accuracy ratio compared to the other classifier.

(Sobar, 2016) used the theory of behavior in social science to detect the probability of being under the risk of cervical cancer using classification techniques like naïve bayes and logistic regression. The used dataset was a questionnaire from Primary Health Care Hospital in Indonesia which contained 72 cases (22 cervical cancer and 50 normal women). The authors used four main behavior determinant theories like theory of planned behavior and protection motivation theory. Seven questions were answered for each theory. The experiment results showed that naïve bayes outperforms logistics regression in accuracy.

(Sharma, 2016), Sharma presented a classification model to identify the stages of cervical cancer using C5.0 with different options like rule sets, boosting and advanced pruning. Sunny Sharma's paper used a clinical dataset from the International Gynecologic Cancer Society (IGCS). The dataset contained 237 cases with 10 features for each case. The used features are like clinical diameter, uterine body, renal pelvic, and renal primary. The experiment showed that C5.0 with advanced pruning options got the highest accuracy ratio of classifying the stage of cervical cancer.

(Nasira, 2016), Vidya and Nasira worked on predicting the normal cervix. The cancer cervix is evaluated using practical data mining algorithms. Geetha and Thangamani addressed the imbalanced distribution of data and risk factors for cervical cancer diagnosis.

(Zhou, 2017) presented a classification model based on Support Vector Machine (SVM) for cervical cancer diagnose by using cervical cancer risk factors dataset. The authors determined the top ten relevant risk factors for the four target variables which is Hinselmann, Schiller, Cytology and Biopsy. The authors also tried to reduce the processing time by eliminating the unimportant features and using the most important features in classification using RFE and PCA techniques.

(Ghouti, 2020) build a completely integrated cervical cancer identification and cervical cancer screening pipeline from cervical images. The current pipeline comprises two deep neural network-learning models for automated cervical identification and diagnosis of cervical tumors. The first test detects the cervix area 1,000 times faster than the state-of-the-art data-driven simulations, thus obtaining a detection precision of 0.68 in terms of union intersection (IoU) estimation. Self-extracted characteristics are used in the second model to identify cervical tumors.

Such features are trained using two lightweight models focused on co-evolutionary neural networks (CNNs). William et al. performed to reduce the probability of mistake by automating the diagnostic process for cervical cancer from Pap-drug.

(Khamparia, 2021) combined a convolutional network with a variational encoder for data classification. The dimensionality of images data can be reduced by using a variational encoder with a softmax layer with the kernel size of 2x2 and 3x3. Their architecture outperforms the current ML models. Chen et al. developed Cyto Brain that facilitates in subsequent clinical diagnosis, an artificial intelligence (AI) based system. CytoBrain consists of three main modules: (1) to extract only cell images in a whole slide image efficiently, cervical cell segmentation module has been designed. (2) for the cell classification, vgg 16 is used, and a classifier module is designed; Moreover, the last one is the human-aided diagnosis module which can automatically diagnose cervical cancer based on the classification results of cells on a whole slide image.

2.3 ANALYSIS OF THE RELATED WORK:

An analysis of the related work in machine learning projects for cervical cancer detection reveals several common themes and trends. Here are some key findings from the analysis:

1. **Deep learning**: Deep learning techniques, particularly convolutional neural networks (CNNs), have emerged as a popular approach for cervical cancer detection. Several studies have reported high accuracy rates for deep learning-based detectors, suggesting that they have significant potential for clinical implementation.
2. **Feature extraction**: Feature extraction is a critical step in developing a cervical cancer detector. Several studies have explored different feature

extraction techniques, such as texture analysis, shape recognition, and edge detection, to identify the most important characteristics of cervical smear images that are predictive of cancer.

3. **Ensemble models:** Ensemble models, which combine multiple machine learning algorithms to improve prediction accuracy, have shown promising results in cervical cancer detection. Several studies have reported improved performance using ensemble models compared to individual models.
4. **Real-world validation:** Several studies have reported successful validation of machine learning-based cervical cancer detectors in real-world clinical settings. These studies have demonstrated the potential for machine learning to improve early detection rates and ultimately save lives.
5. **Challenges:** Despite the promising results, there are still challenges that need to be addressed in the development of machine learning-based cervical cancer detectors. These include the need for large and diverse datasets, the need for interpretable and explainable models, and the need for real-world validation of the models.

Overall, the analysis of the related work in machine learning projects for cervical cancer detection suggests that there is significant potential for machine learning to improve early detection rates and ultimately save lives. However, further research is needed to address the challenges and limitations of these methods and to bring them closer to clinical implementation.

CHAPTER THREE

THE PROPOSED SOLUTION

3.1. FUNCTIONAL/ NON-FUNCTIONAL REQUIREMENTS:

3.1.1. FUNCTIONAL REQUIREMENTS:

1. **Input Data Requirements**: The project will require a dataset of cervical cancer images with accompanying labels indicating whether each image is normal or abnormal. The input data should be in a format that can be easily read by the machine learning models, such as JPEG, PNG or BMP files, Medical history and lifestyle factors for the risk prediction in the website.
2. **Output Requirements**: The output of the project will be a classification result indicating whether each input image is normal or abnormal, in the website the output will be whether the woman is at risk of cervical cancer or not**.**
3. **Machine Learning Models**: The project will utilize two machine learning models for the pap images: Support Vector Machine (SVM) and K-Nearest Neighbors (KNN). The models will be trained on the input dataset to perform the classification task, and Logistic regression, random forest and GaussianNB for risk classification.
4. **Training Process**: The input dataset will be split into training and testing sets, with a portion of the data used to train the models and the result used to evaluate their performance. The models will be trained using the training set and their performance evaluated using the testing set.

5. **Performance Evaluation Metrics**: The performance of the models will be evaluated using metrics such as accuracy, precision, recall, and F1 score.

3.1.2. NON-FUNCTIONAL REQUIREMENTS:

1. **Performance Requirements**: The project should be able to classify images accurately and efficiently, with minimal delay between input and output. The computational requirements of the models should be reasonable and appropriate for deployment on standard hardware.
2. **Security and Privacy Requirements**: The project should ensure the privacy and security of patient data by following best practices for data handling and storage, and by complying with relevant regulations and laws.
3. **Usability Requirements**: The project should be easy to use for both healthcare professionals and patients, with a clear and intuitive interface that guides users through the classification process.
4. **Maintainability Requirements**: The project should be developed in a modular and extensible way that allows for easy maintenance and updates. Code should be well-documented and follow best practices for software development.
5. **Constraints:** The project should adhere to any constraints that are relevant to the application, such as the availability of data or computational resources.

3.2. SYSTEM ANALYSIS & DESIGN:

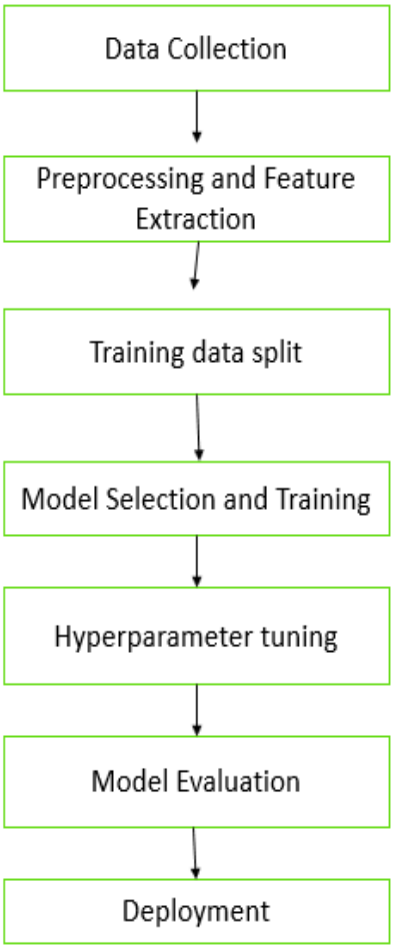


FIGURE 5 BLOCK DIAGRAM

3.2.1. ASSUMPTIONS:

The Pap smear dataset that you are using is reliable and accurate.

The preprocessing and feature extraction methods that you have used are effective in preparing the data for machine learning.

The machine learning algorithms (SVM and KNN) that you have selected are appropriate for the classification task.

The website that you have developed is user-friendly and intuitive for both doctors and women.

The risk factors predictor that you have developed is accurate and reliable in predicting the risk of cervical cancer.

3.2.2. DEPENDENCIES:

The project depends on the availability and quality of the Pap smear dataset. If the dataset is incomplete, inaccurate, or biased, it could affect the accuracy of the machine learning models.

The project depends on the availability and reliability of the machine learning algorithms. If the algorithms are not effective in classifying the pap smear images, the accuracy of the system could be compromised.

The project depends on the availability and reliability of the web development tools and frameworks that you are using to build the website. If these tools are not reliable or if there are compatibility issues, it could affect the performance and functionality of the website.

The project also depends on the availability and reliability of the risk factor data that you are using to develop the risk factor predictor. If the data is incomplete, inaccurate, or biased, it could affect the accuracy of the predictor.

3.2.3. SOFTWARE LIFE CYCLE:

1. **Requirements Gathering:** In this phase, the project requirements are gathered by the stakeholders, including the client's needs, business goals, and technical specifications.
2. **Design:** In this phase, the system architecture is designed, including the algorithms to be used, data flow, and user interface design.
3. **Implementation**: In this phase, the system is developed, and the algorithms are implemented. The data is collected, pre-processed, and features are extracted.
4. **Testing:** In this phase, the system is thoroughly tested to ensure that it meets the project requirements and is functioning as expected**.**
5. **Deployment:** In this phase, the system is deployed in the production environment**.**
6. **Maintenance:** In this phase, the system is maintained and updated as needed to keep it functioning correctly.

The steps for software life cycle for cervical cancer in details:

1. **Data collection and preprocessing:** The system should be able to collect, preprocess, and clean cervical cancer data from various sources to ensure that the data is of high quality and reliable**.**
2. **Data labeling:** The system should provide a mechanism for expert clinicians to label the data as either normal or abnormal.

3. **Model training:** The system should support training of machine learning models using various algorithms and architectures to ensure accurate classification of normal and abnormal cases**.**
4. **Model validation:** The system should be able to validate the trained models using cross-validation and other techniques to ensure that the models are not overfitting to the training data**.**
5. **Model evaluation:** The system should be able to evaluate the performance of the trained models on test data, using metrics such as accuracy, precision, recall, and F1 score.
6. **Model deployment:** The system should be able to deploy the trained models to production for use by clinicians and other stakeholders.
7. **Model monitoring:** The system should continually monitor the performance of deployed models and provide alerts if performance drops below an acceptable threshold.
8. **Model retraining:** The system should support retraining of models on new data to ensure that the classification accuracy remains high over time.
9. **User interface:** The system should have a user-friendly interface that enables clinicians to easily upload new data, access classification results, and provide feedback on the performance of the system.
10. **Security and privacy:** The system should ensure that patient data is stored securely and that access to patient data is limited to authorized personnel only.

3.2.4. TIME PLAN:

1. Requirements Gathering (2 weeks)
2. Design (4 weeks)
3. Implementation (8 weeks)
4. Testing (4 weeks)
5. Deployment (2 weeks)
6. Maintenance (Ongoing)

Total time: 20 weeks

3.2.5. SOFTWARE DESIGN:

Sequence diagram:

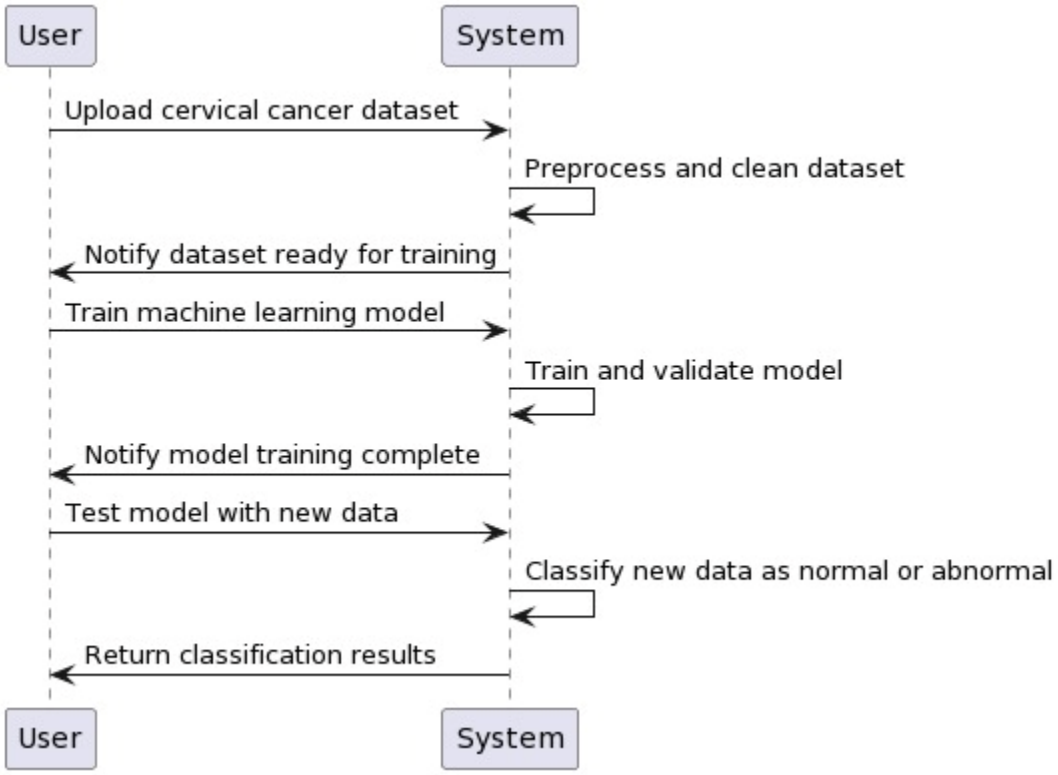


FIGURE 6 SEQUENCE DIAGRAM

State machine diagram:

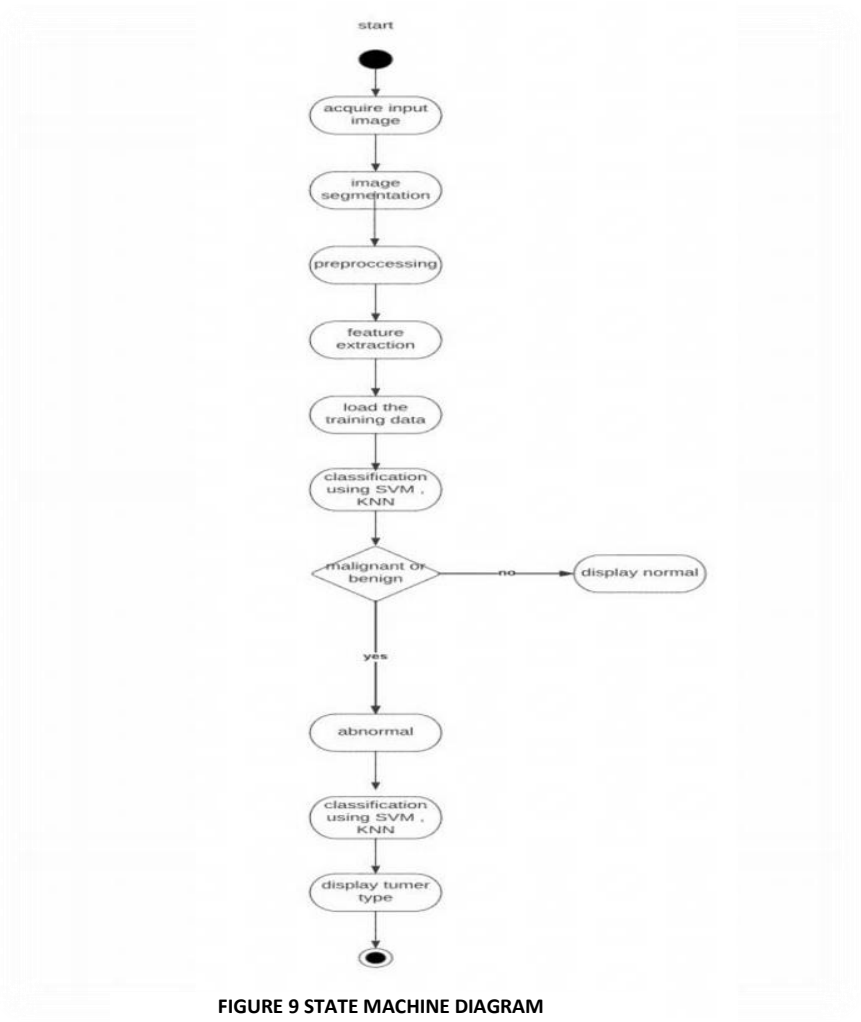


FIGURE 9 STATE MACHINE DIAGRAM

Activity diagram:

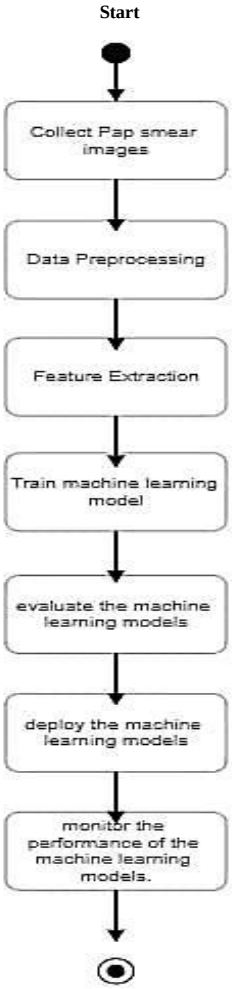


FIGURE 8 ACTIVITY DIAGRAM

Use Case diagram:

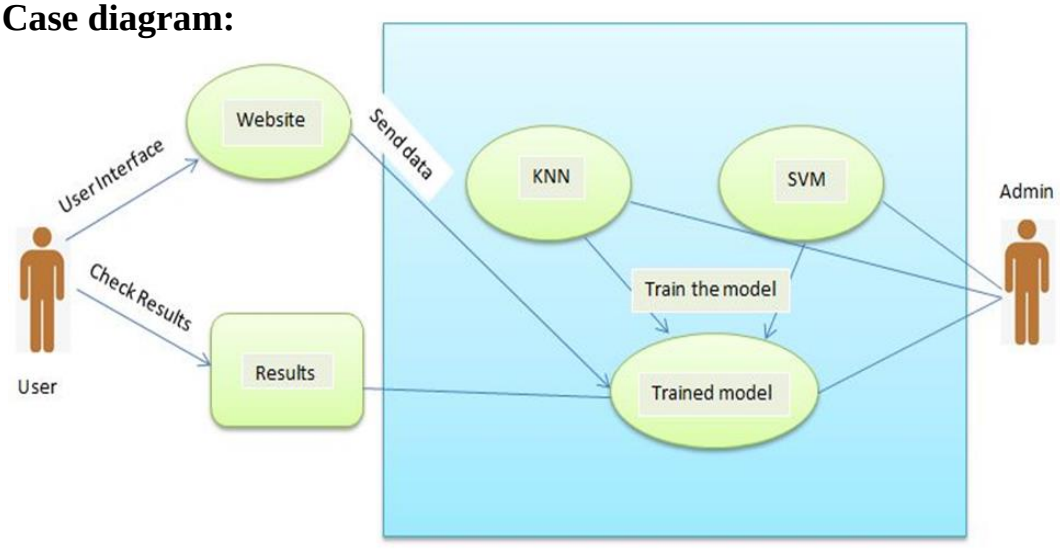


FIGURE 7 USE CASE DIAGRAM

CHAPTER FOUR

IMPLEMENTATION, EXPERIMENTAL SETUP, & RESULTS

4.1. IMPLEMENTATION DETAILS:

Implementing a machine learning project for cervical cancer detection involves several key steps, including data collection, preprocessing, feature extraction, model selection, and evaluation. Here are some more specific implementation details for each of these steps:

4.1.1. DATA COLLECTION:

Collecting a large dataset of cervical smear images is essential for training a machine learning model for cervical cancer detection. The dataset should include a diverse range of images, including both normal and cancerous samples, and should be properly labeled and annotated with relevant patient information, **We used 2 datasets:**

a- Pap smear Data set (HErlev Pap Smear Dataset) Herlev University Hospital (Denmark)

This thesis is based on single cell pictures of papsmear cells from the uterine cervix Furthermore,created by cytotechnicians at Herlev Hospital.

- Contain 7 classes (917 pap smear images , Actual image numbers: 1,834)
- 3 Normal: 242 Normal image *2 Every image has another processed copy
- 4 Abnormal 675 Abnormal image *2

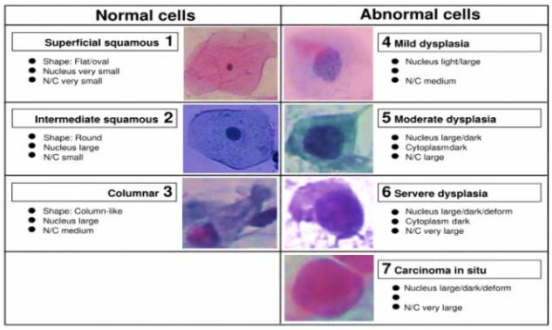


FIGURE 10 CELL TYPE CHARACTERISTICS FOR SINGLE PAP-SMEAR CELLS (CLASSES OF THE DATA SET)

Every image has another processed copy.

Normal - 242 cells

. 1 Superficial squamous epithelial, 74 cells.
. 2 Intermediate squamous epithelial, 70 cells
. 3 Columnar epithelial, 98 cells.

Abnormal - 675 cells

. 4 Mild squamous non-keratinizing dysplasia, 182 cells.
. 5 Moderate squamous non-keratinizing dysplasia, 146 cells.
. 6 Severe squamous non-keratinizing dysplasia, 197 cells.
. 7 Squamous cell carcinoma in situ intermediate, 150 cells.

FIGURE 11 THE 7 DIFFERENT PAP-SMEAR CLASSES

All pictures are segmented into the 3 parts: Background, cytoplasm and nucleus. This segmentation is done one for all and handled by cyto-technicians from Herlev University Hospital using CHAMP. CHAMP is a medical image analysis system based on a colored object recognition algorithm

Feature extraction deals with converting information to a format that is usable for the classifier algorithms. For ex. pictures cannot easily be feed directly into the classifier algorithms. Instead special characteristics are extracted from the pictures. (Martin, 2003)

FIGURE 13 ACTUAL IMAGE

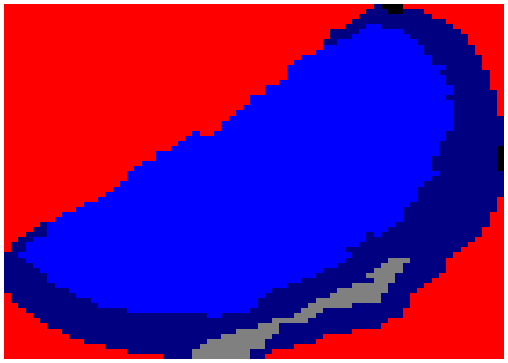


FIGURE 12 SEGMENTED IMAGE

These features extracted and placed inside .csv file to be ready for classification.

This file contains 20 features + ID & class which indicates the type of the cells (columns), 917 rows for each image.

Nucleus area	Cytoplasma area	N/C ratio
Nucl. brightness	Cytopl. brightness	Nucl. shortest diam.
Nucl. longest diam.	Nucl. elongation	Nucl. roundness
Cytopl. shortest diam.	Cytopl. longest diam.	Cytopl. elongation
Cytopl. roundness	Nucl. perimeter	Cytopl. perimeter
Nucl. relative position	Maxima in Nucl.	Minima in Nucl.
Maxima in Cytopl.	Minima in Cytopl.	

FIGURE 14 FEATURES USED IN THIS PROJECT

b- Cervical cancer (Risk Factors)

This dataset focuses on the prediction of indicators/diagnosis of cervical cancer. The features cover demographic information, habits, and historic medical records.

The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela. The dataset comprises demographic information, habits, and historic medical records of 858 patients. Several patients decided not to answer some of the questions because of privacy concerns (missing values).

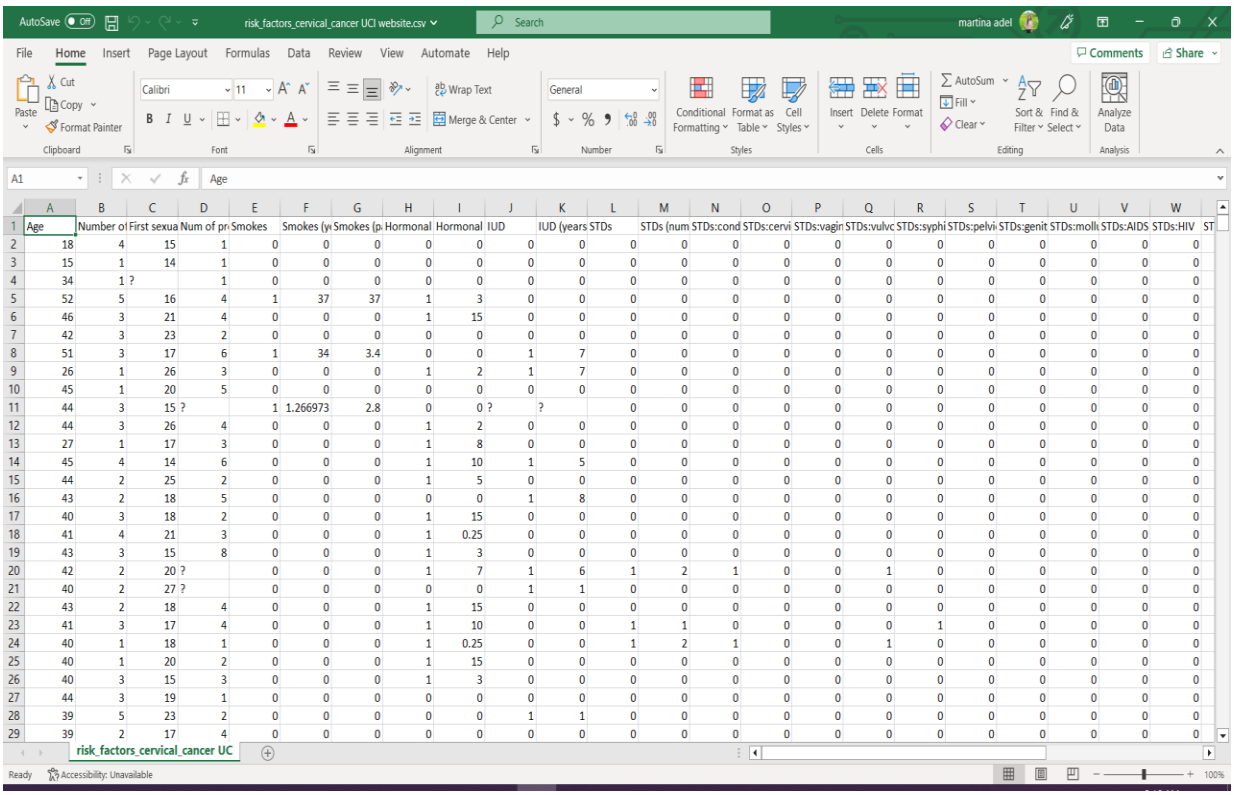
The following are the description of independent and the dependent attributes:

1. Age - It indicates the age of a woman. It is expressed in terms of numerical values
2. Number of sexual partners – It indicates the total number of sexual partners encountered. It is expressed in terms of numerical values.
3. First sexual intercourse- It indicates the age of a woman when she had her first sexual intercourse. It is expressed in terms of the count.

4. Number of pregnancies – It indicates the total number of times the woman got pregnant. It is expressed in terms of the total count.
5. Smokes- It indicates whether the person smokes or not. It is expressed in terms of zeros (does not smoke) and ones(smokes).
6. Smokes (years)- It indicates the total number of years for which the woman is smoking. It is expressed in terms of total count.
7. Smokes (packs/year)- It indicates the total number of packets of cigarettes per year the woman smokes. It is expressed in terms of numbers
8. Hormonal Contraceptives - It indicates whether the patient uses hormonal contraceptives or not.
9. Hormonal Contraceptives (years) – It indicates that for how many years the contraceptive method was used. It was in expressed in terms of total number of years.
10. Intra-Uterine Device- It indicated where the intrauterine contraceptive device was used or not. It was expressed in terms of zeros(did not used IUD) and ones(used IUD).
11. IUD (years) – It indicated that for how many years the IUD was used. It is expressed in terms of the total number of years.
12. STDs - It indicates the presence of Sexually Transmitted Diseases. It is expressed in terms of zeroes and ones.
13. STDs (number) – It indicates the total number of sexually transmitted disease present with the patient. It is expressed in terms of numbers.
14. STDs:condylomatosis – It indicates the presence of Condylomatosis with the patient.

15. STDs:cervical condylomatosis –It indicates the presence of Cervical condylomatosis.
16. STDs:vaginal condylomatosis - It indicates the presence of Vaginal condylomatosis.
17. STDs:vulvo-perineal condylomatosis – It indicates the presence of Vulvo-Perineal condylomatosis.
18. STDs:syphilis – It indicates the presence of Syphilis.
19. STDs:pelvic inflammatory disease- It indicates the presence of pelvic inflammatory disease.
20. STDs:genital herpes – It indicates the presence of Genital Herpes.
21. STDs:molluscum contagiosum – It indicates the presence of Molluscum Contagiosum.
22. STDs:AIDS – It indicates the presence of AIDS in the patient.
23. STDs:HIV – It indicates the presence of HIV in the patient.
24. STDs:Hepatitis B – It indicates the presence of Hepatitis B in the patients.
25. STDs:HPV – It indicates the presence of HPV in the patients.
26. STDs: Number of diagnosis – It indicate the total number of times the STDs have been diagnosed.
27. STDs: Time since first diagnosis – It indicates the total number of years since the first diagnose.
28. STDs: Time since last diagnosis – It indicates the total number of years elapsed since the last diagnose.
29. Dx:Cancer – It indicates the presence of Cancer after the diagnose.
30. Dx:CIN – It indicates the presence of Cervical intraepithelial neoplasia.
31. Dx:HPV- It indicates the presence of Human papillomaviruses.

32. Dx - It indicates the presence any one among cancer, CIN and HPV.
33. Hinselmann – also known as colposcopy, is a medical diagnostic procedure to examine an illuminated, magnified view of the cervix as well as the vagina and vulva.
34. Schiller - Schiller Iodine test is a medical test in which iodine solution is applied to the cervix in order to diagnose cervical cancer.
35. Cytology – also called as PaP smears test, helps detect abnormal cells in the cervix, which can develop into cancer.
36. Biopsy (TARGET) - A cervical biopsy is a surgical procedure in which a small amount of tissue is removed from the cervix. A cervical biopsy is usually done after an abnormality has been found during cytology.



Age	Number of	First sexua	Num of pr	Smokes	Smokes (y	Smokes (p	Hormonal	Hormonal	IUD	IUD (years	STDs	STDs (num	STDs:cond	STDs:cervi	STDs:vagin	STDs:vulvc	STDs:syphi	STDs:pelvi	STDs:genit	STDs:molli	STDs:AIDS	STDs:HIV	ST
18	4	15	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	1	14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
34	1	?	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
52	5	16	4	1	37	37	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
46	3	21	4	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
42	3	23	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
51	3	17	6	1	34	3.4	0	0	1	7	0	0	0	0	0	0	0	0	0	0	0	0	
26	1	26	3	0	0	0	1	2	1	7	0	0	0	0	0	0	0	0	0	0	0	0	
45	1	20	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
44	3	15	?	1	1.266973	2.8	0	0	?	?	0	0	0	0	0	0	0	0	0	0	0	0	
44	3	26	4	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
27	1	17	3	0	0	0	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
45	4	14	6	0	0	0	1	10	1	5	0	0	0	0	0	0	0	0	0	0	0	0	
44	2	25	2	0	0	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
43	2	18	5	0	0	0	0	0	1	8	0	0	0	0	0	0	0	0	0	0	0	0	
40	3	18	2	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
41	4	21	3	0	0	0	1	0.25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
43	3	15	8	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
42	2	20	?	0	0	0	1	7	1	6	1	2	1	0	0	1	0	0	0	0	0	0	
40	2	27	?	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	
43	2	18	4	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
41	3	17	4	0	0	0	1	10	0	0	1	1	0	0	0	0	1	0	0	0	0	0	
40	1	18	1	0	0	0	1	0.25	0	0	1	2	1	0	0	1	0	0	0	0	0	0	
40	1	20	2	0	0	0	1	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
40	3	15	3	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
44	3	19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
39	5	23	2	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	
39	2	17	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

FIGURE 15 RISK FACTORS DATASET

```
In [3]: missing = dataset["Kerne_A"].isna()
        dataset[missing]
```

Out[3]:

ID	Kerne_A	Cyto_A	K/C	Kerne_Ycol	Cyto_Ycol	KerneShort	KerneLong	KerneElong	KerneRund	...	CytoElong	CytoRund	KernePeri	CytoPeri	KernePo

0 rows × 22 columns

```
In [4]: X = dataset[-missing][[ "Kerne_Ycol"]]
        y = dataset[-missing]["K/C"].values
        lin_reg = sm.OLS(y, sm.add_constant(X)).fit()
        print("Adjusted R-squared: {:.3f}%".format(100*lin_reg.rsquared_adj))
        beta = lin_reg.params.values
        print("Estimate:", beta)

        Adjusted R-squared: 2.506%
        Estimate: [0.28033197 0.00162576]

In [5]: dataset["Kerne_Ycol"] = dataset.apply(
            lambda row: (row.K/C - beta[0] - beta[1]*row.KerneShort)/beta[2] if np.isnan(row.Kerne_Ycol) else row.Kerne_Ycol, axis=1
        )
        dataset.isna().sum()

Out[5]: ID           0
        Kerne_A      0
        Cyto_A       0
        K/C          0
        Kerne_Ycol   0
        Cyto_Ycol    0
```

FIGURE 16 EXAMPLE FOR DATA PREPROCESSING IN PAP SMEAR DATA SET

4.1.2. PREPROCESSING:

Preprocessing the data involves several steps; including removing any outliers or errors, addressing missing data, and normalizing or standardizing the data to ensure that it is ready for analysis.

jupyter FinalRiskFactorPredictor Last Checkpoint: 05/26/2023 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
In [15]: #Missing Value Imputation for IUD_years
         x_features_numerical.remove('IUD_years')
         df_impute['IUD_years']=df_impute['IUD_years'].fillna(0)

In [16]: #Missing Value Imputation for Hormonal_Contraceptives
         df_hor=df_impute.drop(['Biopsy'],axis=1)

         x_features_categorical.remove('Hormonal_Contraceptives')
         for i in x_features_categorical:
             df_hor[i]=df_hor[i].fillna(df_hor[i].mode()[0])
         for i in x_features_numerical:
             df_hor[i]=df_hor[i].fillna(df_hor[i].median())

         df_hor=df_hor.astype('float')
         df_hor[x_features_categorical]=df_hor[x_features_categorical].replace(0,'no')
         df_hor[x_features_categorical]=df_hor[x_features_categorical].replace(1,'yes')
         df_hor=pd.get_dummies(df_hor)

         train_hor=df_hor[df_hor.Hormonal_Contraceptives.isnull()==False]
         test_hor=df_hor[df_hor.Hormonal_Contraceptives.isnull()]

         train_hor_x=train_hor.drop('Hormonal_Contraceptives',axis=1)
         train_hor_y=train_hor['Hormonal_Contraceptives']

         test_hor_x=test_hor.drop('Hormonal_Contraceptives',axis=1)
         test_hor_y=test_hor['Hormonal_Contraceptives']
         dt=DecisionTreeClassifier()
         hor_model=dt.fit(train_hor_x,train_hor_y)
         test_hor['Hormonal_Contraceptives']=hor_model.predict(test_hor_x)
```

ocs.io/en/latest/migrate_to_notebook7.html

FIGURE 17 EXAMPLE FOR DATA PREPROCESSING IN RISK FACTORS DATASET

```
In [36]: df_impute.to_csv('df_imputation.csv')

In [37]: df = pd.read_csv('df_imputation.csv', index_col=0) #df_imputation is the new CSV file that doesn't have any null values.

         #Again manually segregating categorical and numerical colmuns

         x_features_categorical = ['Smokes','Hormonal_Contraceptives','IUD','STDs','STDs_condylomatosis','STDs_cervical_condylomatosis',
                                   'STDs_vaginal_condylomatosis','STDs_vulvo_perineal_condylomatosis','STDs_syphilis',
                                   'STDs_pelvic_inflammatory_disease','STDs_genital_herpes','STDs_molluscum_contagiosum','STDs_AIDS',
                                   'STDs_HIV','STDs_Hepatitis_B','STDs_HPV','Dx_Cancer','Dx_CIN','Dx_HPV','Dx', 'Hinselmann', 'Citology',

         x_features_numerical = [x for x in df.columns if x not in x_features_categorical]
```

4.1.3. DATA VISUALIZATION:

1. Pap smear

jupyter GP pap knn,svm-Copy1 knn with crossvalidation Last Checkpoint: 03/08/2023 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

```
In [33]: data.hist(figsize=(20,20))
         plt.show()
```

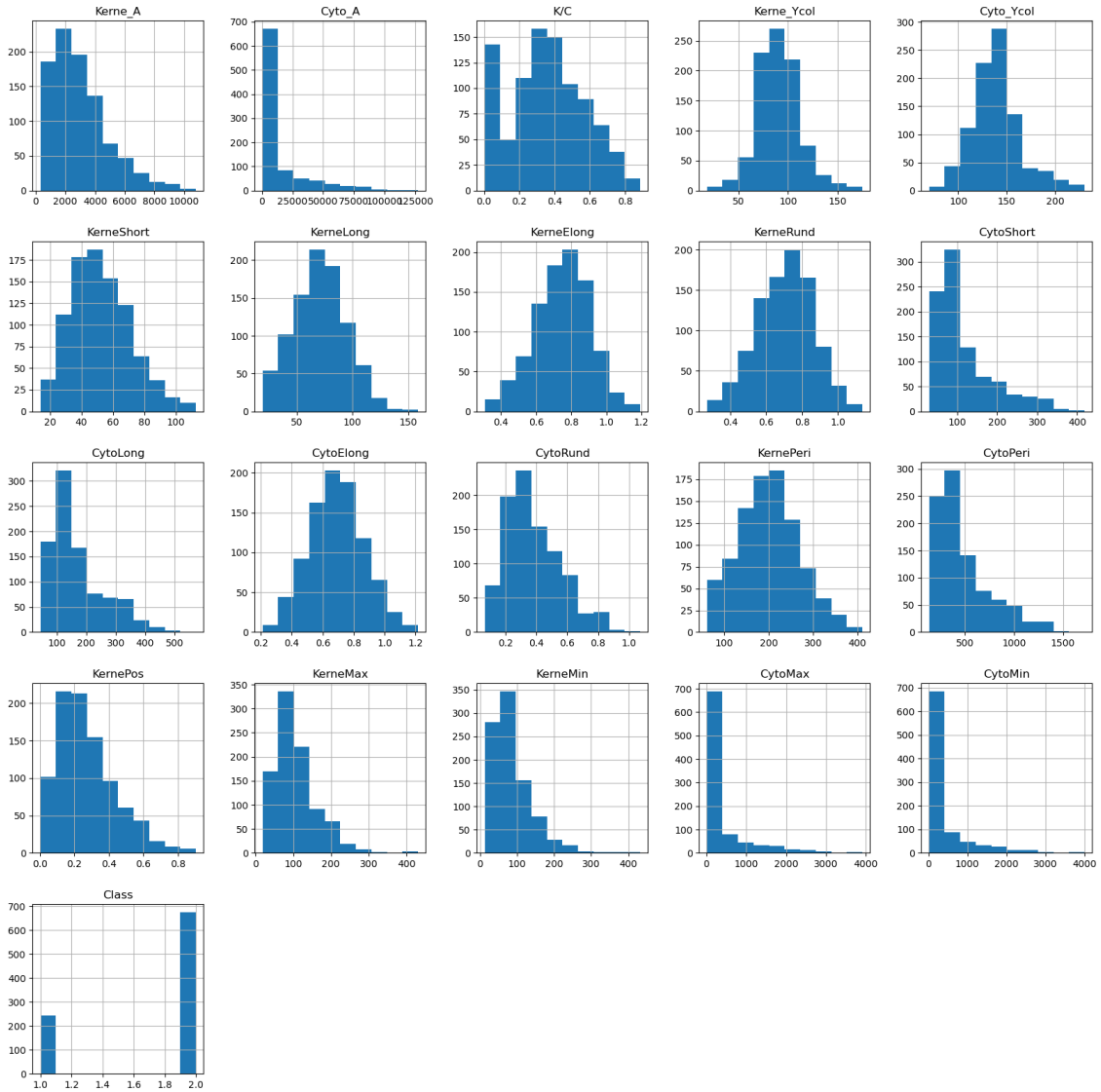


FIGURE 18 HISTOGRAM DATA VISUALIZATION FOR PAP SMEAR DATASET

jupyter GP pap knn,svm-Copy1 knn with crossvalidation Last Checkpoint: 03/08/2023 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)

Run Code

In [32]:
```
#Calculating Confusion Matrix
CM = confusion_matrix(y_test, y_pred)
print('Confusion Matrix is : \n', CM)

# drawing confusion matrix
sns.heatmap(CM, center = True)
plt.show()
```

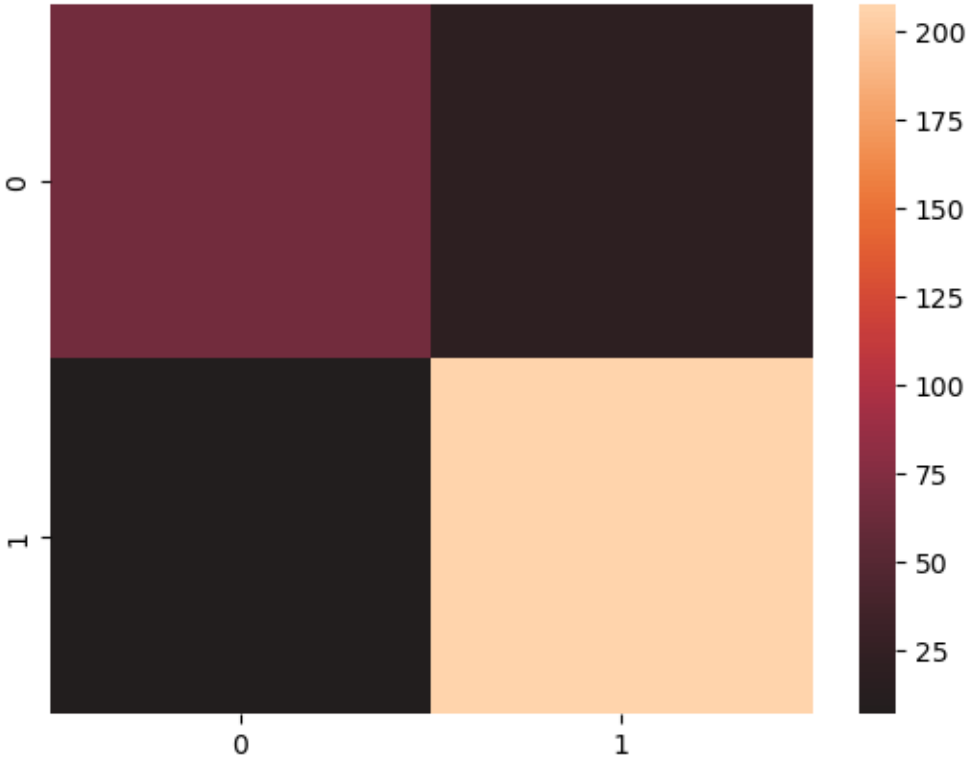


FIGURE 19 CONFUSION MATRIX FOR PAP SMEAR DATASET

2. Risk factors dataset visualization:

jupyter FinalRiskFactorPredictor Last Checkpoint: 05/26/2023 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

In [35]:
```
df_impute[['Age','No_pregnancies', 'No_of_sex_partner',
 'First_sexual_intercourse',
 'Smokes_yrs',
 'Smokes_packs_yr',
 'STDs_No_of_diagnosis', 'Hormonal_Contraceptives_years', 'IUD_years', 'STDs_number']].describe()
```

Out[35]:

	Age	No_pregnancies	No_of_sex_partner	STDs_No_of_diagnosis	Hormonal_Contraceptives_years	STDs_number
count	838.000000	838.000000	838.000000	838.000000	838.000000	838.000000
mean	26.812649	2.271281	2.510143	0.084726	2.392219	0.151551
std	8.529209	1.442143	1.588927	0.295293	3.878695	0.521638
min	13.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25%	20.000000	1.000000	2.000000	0.000000	0.000000	0.000000
50%	25.000000	2.000000	2.000000	0.000000	0.580000	0.000000
75%	32.000000	3.000000	3.000000	0.000000	3.000000	0.000000
max	84.000000	11.000000	28.000000	3.000000	30.000000	4.000000

FIGURE 20 DATA VISUALIZATION FOR RISK FACTORS DATASET

```
In [38]: #Uni Variate Analysis
impute = df.copy()
impute=df.astype('float')

plt.figure(figsize = (12,8))
plt.pie(impute['Biopsy'].value_counts(), labels = ['NO', 'YES'], autopct = '%1.1f%%', labeldistance=1.1, textprops = {'fontsize':
plt.title('Biopsy Percentage', fontsize=20)
plt.show()
```

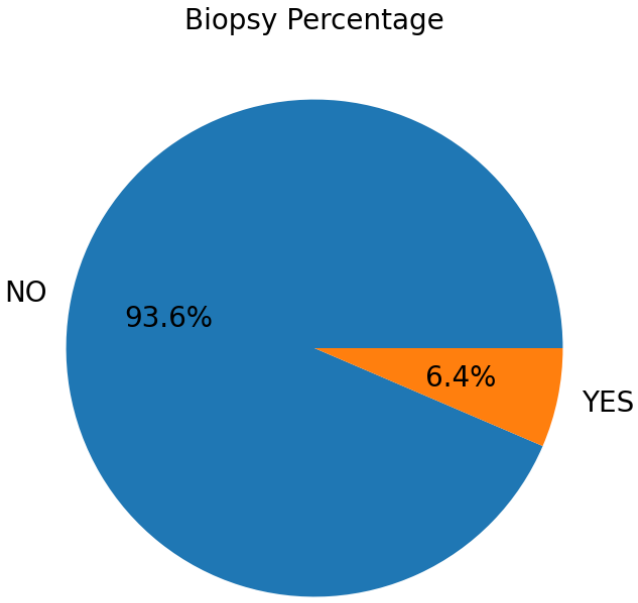


FIGURE 21 UNIVARIATE DATA VISUALIZATION FOR RISK FACTORS DATASET

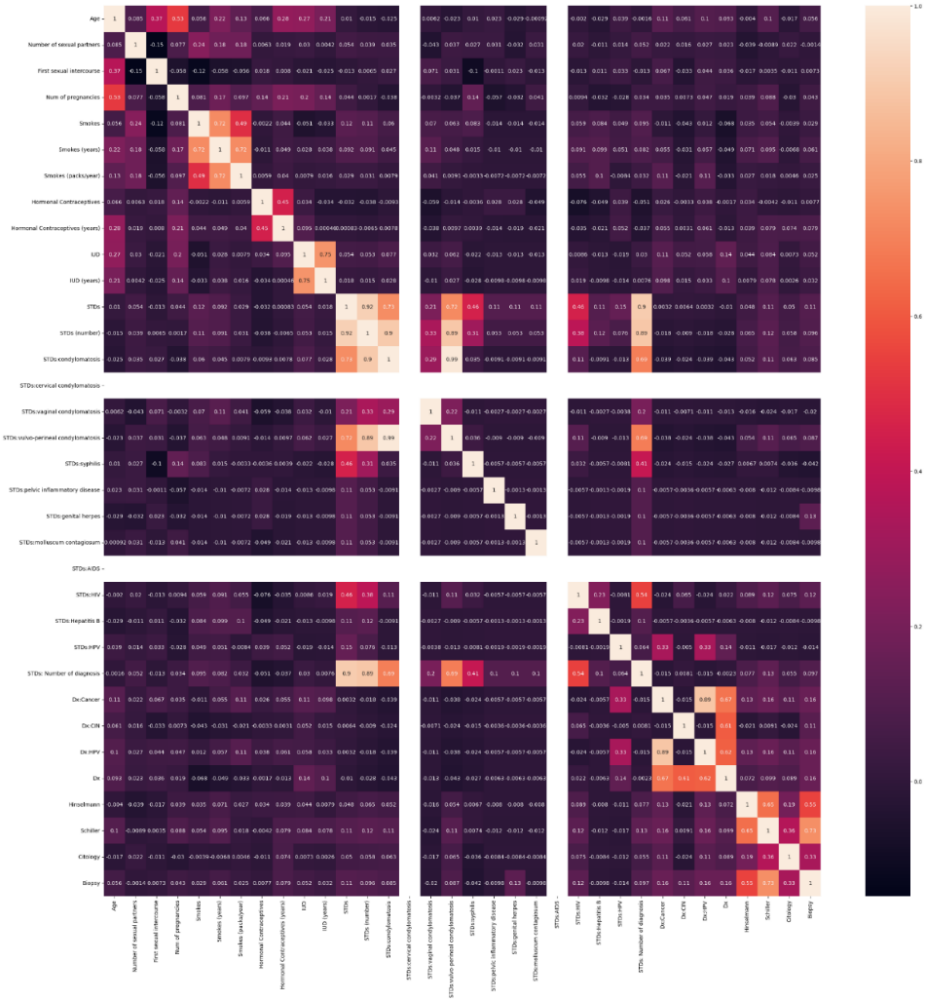


FIGURE 22 CORRELATION MATRIX FOR RISK FACTORS DATASET

3. **Feature extraction:** Feature extraction involves identifying the most important characteristics of the cervical smear images that are predictive of cancer. This may involve using techniques such as converting images to grayscale, morphological opening, thresholding, labeling regions, and extracting region properties.
4. **Model selection**: Selecting an appropriate machine learning model is critical to the success of the cervical cancer detector. The choice of model depends on the specific characteristics of the dataset and the goals of the project. Common models used for classification tasks include KNN , support vector machines, and Logistic regression.

i. For pap smear data set

- Binary classification to classify papsmear images into 2 classes which are normal an abnormal using 2 algorithms

1- Support vector machine with accuracy 95%

2- KNN with accuracy 96%

- Then we classified the types of the abnormal cells (4 classes)
- SVM with accuracy 100% accuracy

3- Risk factors dataset

```
In [52]: base_df = pd.DataFrame(l)
         base_df
```

Out[52]:

	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
0	LogisticRegression	0.974403	0.952381	0.571429	0.533333	0.615385	0.756118
1	Decision Tree	1.000000	0.928571	0.550000	0.733333	0.440000	0.837131
2	Random Forest	1.000000	0.944444	0.588235	0.666667	0.526316	0.814346
3	GaussianNB	0.146758	0.095238	0.116279	1.000000	0.061728	0.518987
4	KNN	0.950512	0.936508	0.333333	0.266667	0.444444	0.622785

FIGURE 23 COMPARISON BETWEEN ACCURACIES OF RISK FACTORS MACHINE LEARNING MODELS

5. **Evaluation:** Evaluating the model involves testing its accuracy and effectiveness in detecting cervical cancer. This may involve using cross-validation techniques or other methods to ensure that the model is robust and reliable.
6. **Deployment**: We deployed risk factors machine learning model to a website to be used with every woman to know if she is at risk of cervical cancer or not based on her medical history and lifestyle (personally) the website offers every women from her home information and personal advice to avoid cervical cancer and awareness about it.

   Backend: Flask.

   Frontend: HTML & Css

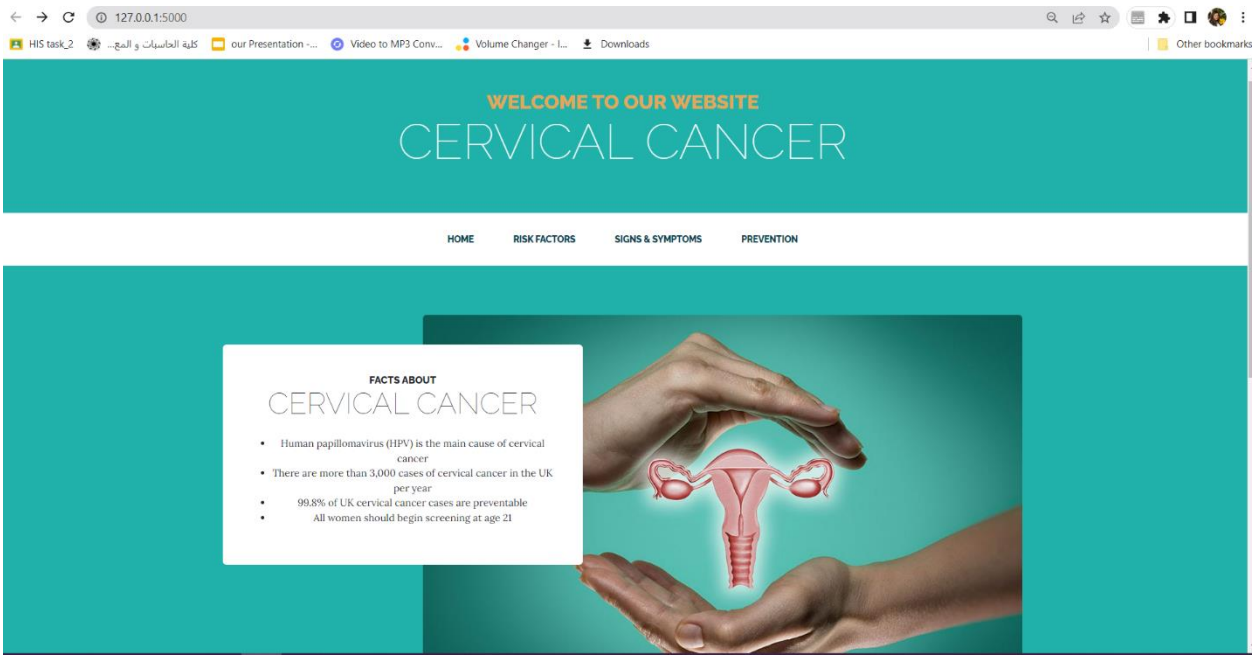


FIGURE 24 OUR WEBSITE HOME PAGE

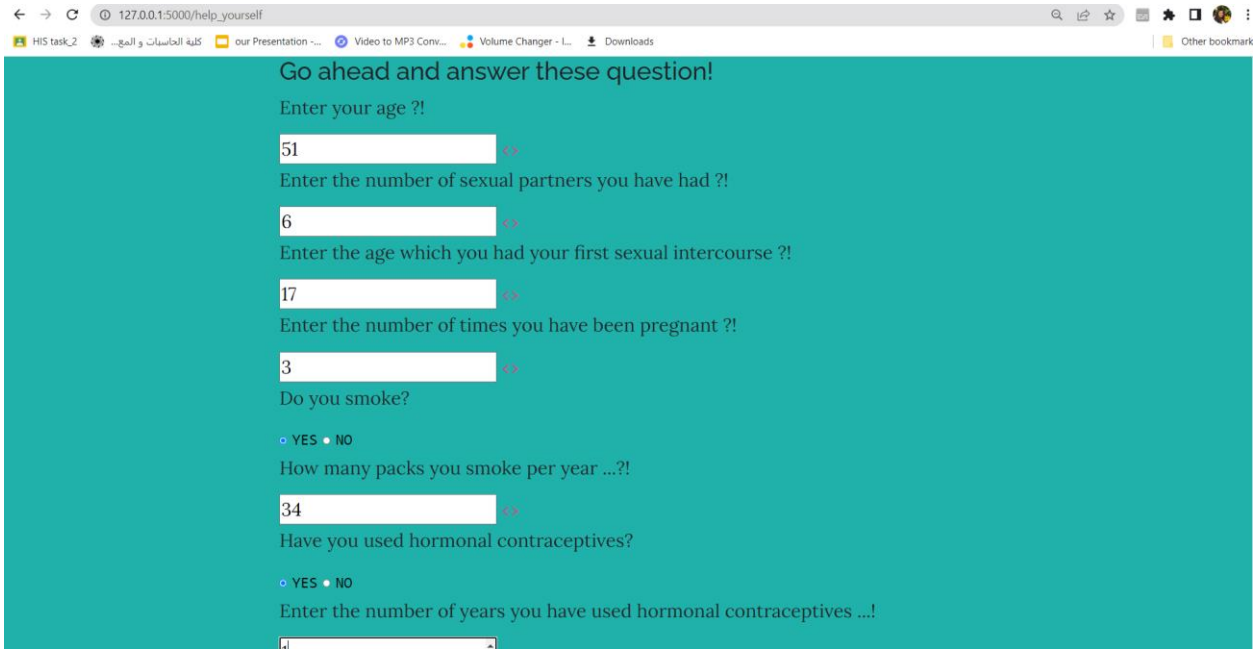


FIGURE 26 RISK FACTORS PREDICTOR PAGE

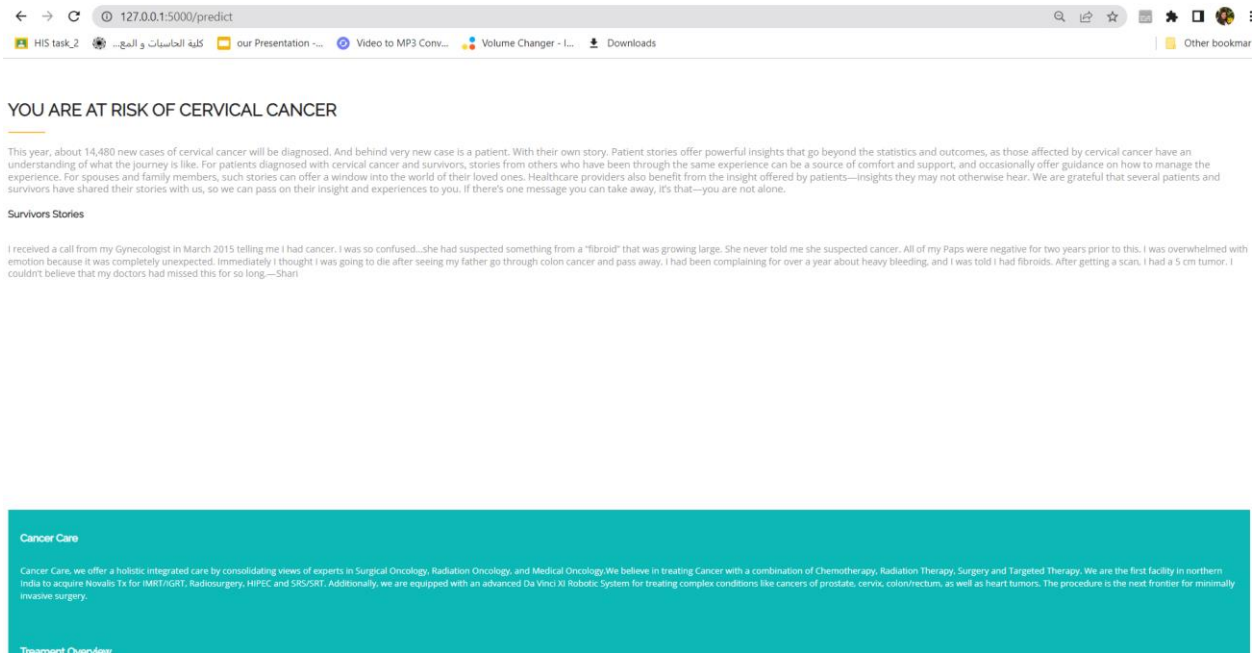


FIGURE 25 PREDICTION PAGE (CASE 1 : AT RISK OF CERVICAL CANCER)

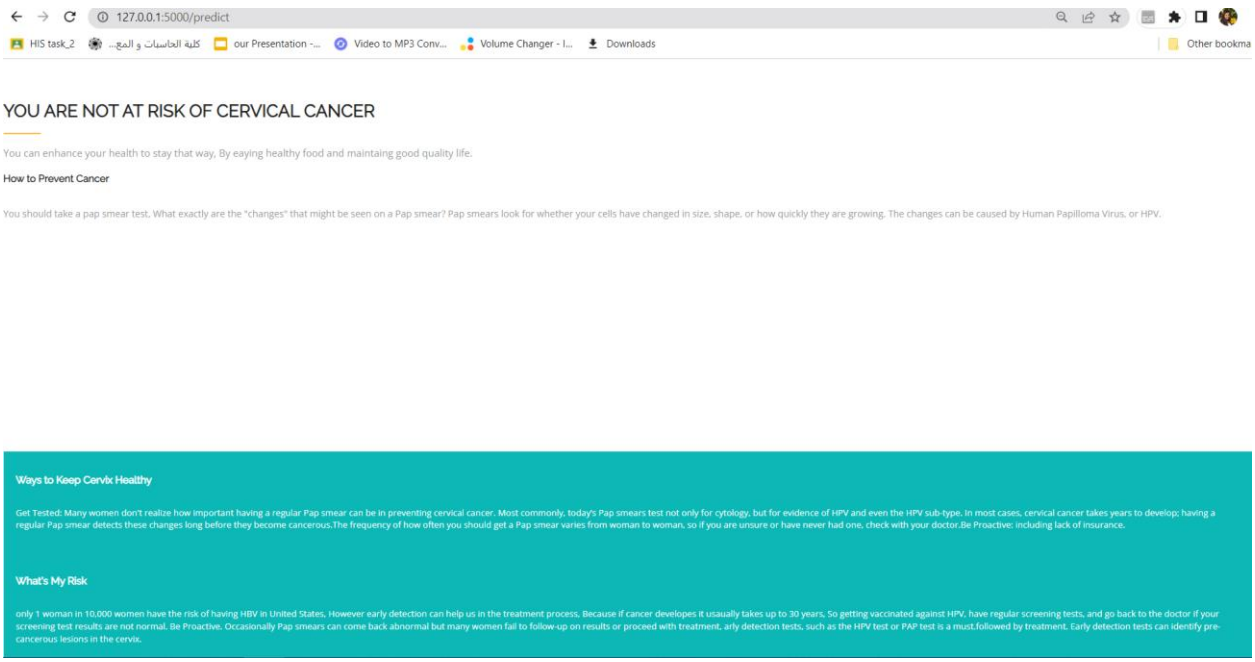


FIGURE 27 PREDICTION PAGE (CASE 2: NOT AT RISK OF CERVICAL CANCER)

In addition to these steps, there are several other important implementation details to consider when developing a machine learning-based cervical cancer detector. **These include:**

- Choosing appropriate hyperparameters for the model, such as learning rate, batch size, and regularization strength.
- Using techniques such as data augmentation to increase the size and diversity of the dataset.
- Ensuring that the model is interpretable and explainable, so that healthcare professionals can understand how the model is making its predictions.
- Deploying the detector in a secure and scalable manner, so that it can be used in real-world healthcare settings.

By carefully considering these implementation details, researchers can develop accurate and reliable machine learning-based cervical cancer detectors that can improve early detection rates and ultimately save lives.

4.2. EXPERIMENTAL / SIMULATIONS SETUP:

An experimental and simulation setup for a machine learning project for cervical cancer detection involves designing an experiment or simulation that can accurately test the performance of the developed detector. Here are some key considerations for setting up such an experiment or simulation:

1. **Data selection**: The data used in the experiment or simulation should be representative of the population for which the detector is intended. The data should include both normal and cancerous samples, and should be properly labeled and annotated.
2. **Data partitioning**: The data should be partitioned into training and testing sets. The training set is used to train the machine learning model, while the testing set is used to evaluate the performance of the developed detector.
3. **Feature extraction**: Feature extraction techniques should be applied to the data to identify the most important characteristics of the cervical smear images that are predictive of cancer.
4. **Model training**: The machine learning model should be trained on the training set using appropriate hyperparameters and optimization algorithms.
5. **Model evaluation**: The performance of the developed detector should be evaluated using metrics such as sensitivity, specificity, and accuracy. These metrics should be calculated on the testing set.
6. **Simulation setup:** In a simulation setup, synthetic data can be generated to simulate different scenarios and test the performance of the developed detector under various conditions.
7. **Performance comparison:** The performance of the developed detector can be compared to other existing detection methods to determine if it outperforms them. We tried to use Intel® Extension for Scikit-learn* is a

free software AI accelerator that brings over 10-100X acceleration across a variety of applications.

8. **Statistical analysis:** Statistical analysis can be performed to determine the significance of the results and ensure that they are not due to chance.

Overall, an experimental and simulation setup for a machine learning project for cervical cancer detection should be carefully designed to ensure that the results are accurate and reliable. By considering these key considerations, researchers can accurately evaluate the performance of their developed detector and compare it to other existing methods.

4.3. CONDUCT RESULTS:

Conducting results in a machine learning project for cervical cancer detection involves analyzing the performance of the developed detector and interpreting the results. Here are some key steps in conducting results:

1. **Analysis of metrics:** The performance of the developed detector should be analyzed using metrics such as sensitivity, specificity, and accuracy. These metrics provide an indication of how well the detector is able to identify cancerous samples.
2. **Performance comparison:** The performance of the developed detector should be compared to other existing detection methods to determine if it outperforms them. This can be done by comparing the metrics of the developed detector to those of other methods.
3. **Statistical analysis**: Statistical analysis should be performed to determine the significance of the results and ensure that they are not due to chance. This can be done using techniques such as hypothesis testing and confidence intervals.

4. **Interpretation of results**: The results of the analysis should be interpreted to provide insights into the performance of the developed detector. This can involve identifying strengths and weaknesses of the detector, as well as identifying areas for improvement.
5. **Visualization of results**: Visualization techniques can be used to present the results in an easy-to-understand format. This can include plots, charts, and other graphical representations of the data.
6. **Reporting of results**: The results of the analysis should be reported in a clear and concise manner. This can involve writing a technical report or creating a presentation to communicate the findings to stakeholders.

Overall, conducting results in a machine learning project for cervical cancer detection is a critical step in evaluating the performance of the developed detector and providing insights into its strengths and weaknesses. By analyzing the metrics, comparing performance to other methods, performing statistical analysis, interpreting the results, visualizing the results, and reporting the findings, researchers can provide valuable insights into the effectiveness of their developed detector and identify areas for future research and improvement.

4.4. TESTING & EVALUATION:

Testing and evaluation are crucial steps in any machine learning project for cervical cancer detection. Here are some key considerations for testing and evaluating the developed detector:

1. **Testing set:** The testing set should be representative of the population for which the detector is intended. It should include both normal and cancerous samples, and should be properly labeled and annotated.

2. **Metrics:** Metrics such as sensitivity, specificity, accuracy, precision, and recall should be used to evaluate the performance of the developed detector. These metrics provide an indication of how well the detector is able to identify cancerous samples.
3. **Cross-validation:** Cross-validation techniques such as k-fold cross-validation can be used to ensure that the results are robust and reliable. In k-fold cross-validation, the data is split into k-folds, and the model is trained and evaluated k times, with each fold serving as the testing set once.
4. **Performance comparison:** The performance of the developed detector should be compared to other existing detection methods to determine if it outperforms them. This can be done by comparing the metrics of the developed detector to those of other methods.
5. **Statistical analysis**: Statistical analysis should be performed to determine the significance of the results and ensure that they are not due to chance. This can be done using techniques such as hypothesis testing and confidence intervals.
6. **Interpretation of results**: The results of the testing and evaluation should be interpreted to provide insights into the performance of the developed detector. This can involve identifying strengths and weaknesses of the detector, as well as identifying areas for improvement.
7. **Deployment**: Finally, the developed detector should be deployed in a real-world setting to determine its effectiveness in practice. This can involve working with healthcare professionals to integrate the detector into their workflow and collecting feedback on its performance.

Overall, testing and evaluation are critical steps in the development of a machine learning-based cervical cancer detector. By using appropriate metrics, cross-validation techniques, performance comparison, statistical analysis, interpretation of results, and real-world deployment, researchers can evaluate the effectiveness of the developed detector and identify areas for future research and improvement.

CHAPTER FIVE

DISCUSSION, CONCLUSIONS, AND FUTURE WORK

5.1. DISCUSSION:

Cervical cancer is one of the most common forms of cancer affecting women worldwide, and early detection is key to successful treatment (Martin, 2003). Machine learning can play an important role in assisting healthcare professionals in identifying cervical cancer at an early stage, allowing for prompt treatment and better patient outcomes.

A machine learning project aimed at developing a cervical cancer detector would involve several stages, including data collection, preprocessing, feature extraction, model selection, and evaluation.

Data collection is a crucial first step in any machine learning project. In this case, the project would involve collecting a large dataset of cervical smear images, along with information about the patients, such as age, medical history, and other relevant factors.

The next step would be preprocessing the data to ensure that it is clean and ready for analysis. This would involve removing any outliers or errors, as well as addressing any missing data.

Feature extraction is another important step in the process of developing a cervical cancer detector. This involves identifying the most important characteristics of the cervical smear images that are predictive of cancer. These features using techniques such as converting images to grayscale, morphological

opening, thresholding, labeling regions, and extracting region properties, as well as other factors.

Once the features have been extracted, the next step is to select an appropriate model for the detector. There are many different machine learning models that could be used for this purpose, including KNN, support vector machines, and Logistic regression. The choice of model will depend on the specific characteristics of the dataset and the goals of the project.

Finally, the model will need to be evaluated to determine its accuracy and effectiveness in detecting cervical cancer. This might involve using cross-validation techniques or other methods to ensure that the model is robust and reliable.

5.2. SUMMARY & CONCLUSION:

In summary, a machine learning project aimed at developing a cervical cancer detector would involve collecting a large dataset of cervical smear images, preprocessing the data to ensure it is clean and ready for analysis, extracting important features from the images, selecting an appropriate machine learning model, and evaluating the model's accuracy and effectiveness in detecting cervical cancer.

Such a project has the potential to make a significant impact in the field of healthcare by assisting healthcare professionals in identifying cervical cancer at an early stage, allowing for prompt treatment and better patient outcomes. The development of a cervical cancer detector using machine learning techniques can potentially save lives and improve the quality of life for women around the world.

In conclusion, the use of machine learning in healthcare is an exciting area of research with the potential to revolutionize the way diseases are detected, diagnosed, and treated. The development of a cervical cancer detector is just one example of how machine learning can be applied to improve healthcare outcomes and save lives. With continued research and development in this field, we can look forward to even more innovative solutions that will benefit patients and healthcare providers alike.

5.3. FUTURE WORK:

While machine learning has shown promising results in developing a cervical cancer detector, there is still much work to be done to improve the accuracy and reliability of these systems. Here are some potential areas for future work in this field:

1. Developing more sophisticated feature extraction techniques: Feature extraction is a critical step in developing a cervical cancer detector, and researchers are continually exploring new and more sophisticated techniques for identifying the most important characteristics of cervical smear images.
2. Augmenting datasets with additional information: While a large dataset of cervical smear images is essential for training a machine learning model, researchers can also explore incorporating additional patient information, such as medical history and lifestyle factors, to further improve the accuracy of the detector.
3. Developing ensemble models: Ensemble models combine multiple machine learning models to improve prediction accuracy. Researchers can explore developing ensemble models for cervical cancer detection, which can potentially increase the accuracy and reliability of the system.

4. Developing explainable models: Explainable machine learning models can provide insights into how the model makes predictions. This can help healthcare professionals better understand the underlying factors that contribute to cervical cancer and improve patient outcomes.
5. Deploying detectors in low-resource settings: Many low-resource settings lack access to the sophisticated equipment and expertise required for traditional cervical cancer screening methods. Researchers can explore developing machine learning-based detectors that can be deployed in these settings to improve access to early detection and treatment.
6. Integrate colposcopy dataset to make the system more comprehensive.

Overall, there is still much potential for machine learning in developing more accurate and reliable cervical cancer detectors. Continued research and development in this field have the potential to improve early detection rates, save lives, and improve the quality of life for women around the world.

REFERENCES (OR BIBLIOGRAPHY)

[1] Bjerregaard, B. (2002), Computerstyret automatisk udstyr til screening for livmoderhalskræft, Technical report, Amtssygehuset Herlev.

[2] Duda, R. O., Hart, P. E. & Stork, D. G. (2000), Pattern Classication, 2 edn, A Wiley-Interscience Publication.

[3] Gonzalez, R. C. & Woods, R. E. (1993), Digital image processing, AddisonWesley Publishing Company.

[4] Jantzen, J. (1998), Neuro fuzzy modelling, Technical report, Technical University of Denmark, Dept. of Automation, Bldg 326, 280.

[5] Landwehr, D. (2001), Web based pap-smear classication, Master's thesis, Technical University of Denmark(DTU), Dept. of Automation, Bldg 326, 2800 Lyngby, Denmark.

[6] Martin, Erik, et al. Pap-Smear Classication. 2003.

[7] Norup, and onas c960566. Classification of Pap-Smear Data by Transductive Neuro-Fuzzy Methods. 2 May 2005.

[8] g.dounias, and J.jantzen. Automated Identification of Cancerous Smears Using Various Competitive Intelligent Techniques. 28 Sept. 2005.

[9] Nikolaos Ampazis1, et al. *Pap-Smear Classification Using Efficient Second Order Neural Network Training Algorithms*.

[10] Mehmood M, Rizwan M, Gregus ml M and Abbas S (2021) Machine Learning Assisted Cervical Cancer Detection. Front. Public Health 9:788376.

[11] Alias, N.A.; Mustafa, W.A.; Jamlos, M.A.; Alquran, H.; Hanafi, H.F.; Ismail, S.; Rahman, K.S.A. Pap Smear Images Classification Using Machine Learning: A Literature Matrix. Diagnostics 2022, 12, 2900. https://doi.org/10.3390/diagnostics12122900

[12] Mithlesh Arya, and Namita Mittal. "Clustering Techniques on Pap-Smear Images for the Detection of Cervical Cancer." *Research Gate*, Jan. 2018.

[13] SHERIF F. ABDOH, et al. "Cervical Cancer Diagnosis Using Random Forest Classifier with SMOTE and Feature Reduction Techniques." *IEEE*, 26 Sept. 2018.

[14] Sanjay Kumar Singh. "Performance Analysis of Machine Learning Algorithms for Cervical Cancer Detection." *International Journal of Healthcare Information Systems and Informatics*, 2 Apr. 2020.

[15] Yannis Marinakisa. "Pap Smear Diagnosis Using a Hybrid Intelligent Scheme Focusing on Genetic Algorithm Based Feature Selection and Nearest Neighbor Classification." *Elsevier*, 20 Nov. 2008.

[16] Jantzen, J., & Dounias, G. (2006). Analysis of Pap-smear Image Data. In Proceedings of the Nature-Inspired Smart Information Systems 2nd Annual Symposium NiSIS.

[17] Seung Hee Ho, et al. "Analysis on Risk Factors for Cervical Cancer Using Induction Technique." *Elsevier*, July 2004.

[18] Emmanuel Ahishakiye, et al. *Prediction of Cervical Cancer Basing on Risk Factors Using Ensemble Learning*. May 2020.

[19] X. Deng, Y. Luo and C. Wang, "Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods," 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), Nanjing, China, 2018, pp. 631-635, doi: 10.1109/CCIS.2018.8691126.

[20] Paul A Cohen MD, et al. “Cervical Cancer.” *The Lancent*, 18 Jan. 2019.

[21] HErlev (HErlev Pap Smear Dataset) Herlev University Hospital (Denmark), 2005.

[22] Fernandes,Kelwin, Cardoso,Jaime & Fernandes,Jessica. (2017). Cervical cancer (Risk Factors dataset). UCI Machine Learning Repository. https://doi.org/10.24432/C5Z310.