IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Detection of Cervical Cancer using Ensembling

G. Bharathi[1*], D. Aasritha[1], A. Ashok Kumar Reddy[2], A. Praveen Kumar[3], Abbas Hussain[4]

Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru, India

aasritha036@gmail.com

Abstract According to statistics, the third most fatal disease affecting women worldwide is cervical cancer. Nearly 99.7% of occurrences of cervical cancer that result in tumors in the infected area are caused by HPV infection. To treat the condition, doctors advise early diagnosis and treatment. Due to high expense of treatment, the absence of adequate healthcare services and the symptoms' delayed onset, systematic screening for illness detection is not performed in both developed and underdeveloped nation. Machine intelligence makes early identification of a variety of illnesses, including cervical cancer. Also, it is economical and cost-effective to compute. Patients do not need to undergo time-consuming, modern medical procedures, and artificial intelligence can aid in the early identification of cervical cancer. The drawback of the current machine classification methods for illness detection is their reliance on the prediction accuracy of a single classifier. The adoption of a single classification technique does not provide the best prediction due to bias and over-fitting. This research is done using ensemble classification approach based on majority voting to deliver a proper diagnosis that addresses the patient's symptoms or concerns. As a result, the suggested paradigm grants health professionals a second opinion to aid in the early diagnosis and prompt treatment of diseases.

Keywords: Cervical cancer, Ensemble classification, Support vector machine, Random forest, Extra tree classifier, CatBoost classifier, Logistic regression.

1Introduction

A woman's cervix gets cervical cancer. 99% of cervical cancer cases are linked to human papillomaviruses (HPV), an extraordinarily common virus disseminated through sexual contact. Several analyses have demonstrated how cervical cancer therapy and patient recovery can be dramatically impacted by early diagnosis. In starting-stage this cancer can be treated with either surgery or radiation treatment along with chemotherapy. The main therapies for later stages are typically radiation and chemotherapy. Chemotherapy is routinely used to treat advanced cervical cancer (by itself).

The approaches that are currently being used to identify cervical cancer generally focus on a single classifier. Due to the fact that ensemble algorithms frequently produce better results when faced with particular problems, this article suggests using them for the detection of cervical cancer. Support Vector Machine, Random Forest Tresses, Extra Tree Classifier, CatBoost Classifier, and Logistic Regression are just a few of the classifiers used in this paper's methodology.

Papanicolaou developed the Pap smear method, which includes scraping cells from the cervix and sticking them to a glass slide to check for precancerous or malignant changes in the cervix. In order to manage or treat CIN and early cervical cancer and stop disease development brought on by invasive cancer, the Pap smear is a cytologic screening test that is used to identify these disorders. Results from cervical cytology cannot be used to make a diagnosis of CIN or cancer because a biopsy and histologic confirmation are needed.
As a result, a significant fraction of these cancer cases is diagnosed at an advanced stage when there is little chance of recovery. Although signs of cervical cancer sometimes do not manifest until the illness has advanced, cervical cancer screening is essential. Preinvasive lesions almost always lead to five years of survival in women.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

2 Literature Survey

The third most fatal disease affecting women worldwide is cervical cancer. The main cause of this illness is HPV, which infects a location and causes a tumor there. Health professionals advise early detection and vaccination as treatments for the illness. Nearly 99.7% of cases of cervical cancer are brought on by HPV infection. With an approximate 528,000 instances reported in 2012, it is a prevalent malignancy in women. Accurate and speedy disease detection and diagnosis are essential for current cancer research.

It's feasible that in the future, cervical cancer may be detected by a wide range of machine intelligence. Machine intelligence research into cervical cancer prediction has rapidly risen over the previous five years, in part due to the relevance of this ailment and its quick identification[3]. Thus, machine learning technologies are exciting and crucial for the timely identification of cervical cancer in the medical area. These clever machine learning applications greatly aid in giving medical professionals a second view.

Most cervical cancer detection methods concentrate on pap smear results, cervical imaging analysis, and illness signs. Several different categorization approaches are used in various research domains. For particular datasets of people with cervical cancer, the bulk of recent research compares the performance of one classifier against that of other classifiers.

In a research article, the author specifies that he had used the CYENET-Colposcopy Ensemble Network a new version of the previous model, is created to automatically classify cervical tumors from colposcopy pictures. For VGG19, the classification accuracy was 73.3%. The CYENET model now has a classification accuracy of 92.3% [2].

In a research article, Machine learning algorithms like multi-layer perceptron, decision trees, random forest, K-Nearest Neighbor and Naïve-Bayes have been used for prediction. The final prediction model had an accuracy of 87.21% [1].

In a research paper, the author used the Ensemble classifier a new version of the previous model consists Decision Tree Classifier and Random Forest. The classification accuracy of Decision Tree Classifier is 85.11% and Random Forest is 87.90% [4].

In a research article, the researcher in a study describes both the CervDetect and a hybrid strategy that includes RF and shallow neural network. CervDetect uses machine learning algorithms to analyse medical data in order to determine the health conditions of aggressive cervical development. The proposed strategy has a 93.6% accuracy rate, according to the results.[8]

3 Related Work

According to statistics, the third most fatal illness affecting women globally is cervical cancer. The major cause of this illness is HPV, which infects a location and causes a tumour there. Health professionals advise early detection and immunisation as treatments for the illness. Nearly 99.7% of cases of cervical cancer are brought on by HPV infection. With an estimated 528,000 cases reported in 2012, it is also one of the most prevalent tumours in women. Modern cancer research demands accurate illness diagnosis and early detection.

Currently, the Thinprep Cytologic Test (TCT) and the identification of the human papillomavirus (HPV) are the most widely utilised screening methods for cervical cancer early detection. The HPV test finds high-risk viral infections that might cause cancer and cervical lesions. TCT evaluates if pathogenic factors are causing aberrant alterations in the cervix's cells that might result in cervical cancer. It is hypothesised that there must be additional synergistic elements that may raise the risk of cervical cancer because the path from HPV to cancerization is still quite long.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

Studies have revealed significant variations in the likelihood of HPV cancerization at various ages. Cervical cancer is a highly preventable illness, but due to socioeconomic variables and educational attainment, few women are familiar with its origins, risk factors, preventative strategies, and treatment options. Developed nations continue to have a lower incidence of cervical cancer than less developed nations. In actuality, low-income nations account for 95% of fatalities from cervical cancer. HPV infection, which is spread through intercourse, is another important cause.

The bulk of cervical cancer diagnosis techniques rely on the study of cervical imaging, pap smear findings, and general disease symptoms. Many different categorization methods are used in various research contexts. The majority of current research uses a single classifier and compares its performance to that of other classifiers on specific datasets of individuals with cervical cancer. A few papers with references used ensemble classification using various strategies to improve the accuracy rate of the classification method. The prediction accuracy of classification systems was significantly improved by pre-processing, extraction of features, and quantization. This paper introduces a new ensemble method that combines the majority vote and the prediction results of multiple weak classifiers to create a reliable ensemble classification model.

4 Proposed System

The success of each classifier is essential for the most accurate cervical cancer diagnosis. We have seen in the literature that the orientation of dataset in deviations and incomplete data points is strongly connected with the performance of the classifier. It is additionally clear that certain classifications dominate some moderate or minor data points. Particularly for larger data points, they perform poorer when the standard deviations are higher. The timely and correct detection of the disease calls for a potent and hard remedy since a precise detection of cancer is necessary for medical professionals to provide the finest treatments. To get the best classification results for cervical cancer prediction, this work uses an ensemble classification approach on a significant voting mechanism.

The dataset used in the paper contains the risk indicators of cervical cancer.

S/N	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives	Hormonal Contraceptives (years)	IUD	IUD (years)	STDs	STDs (number)	ST
1	18.0	4.0	15.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	15.0	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	34.0	1.0	14.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	52.0	5.0	16.0	4.0	1.0	37.0	37.0	1.0	3.0	0.0	0.0	0.0	0.0	
5	46.0	3.0	21.0	4.0	0.0	0.0	0.0	1.0	15.0	0.0	0.0	0.0	0.0	
6	42.0	3.0	23.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
7	51.0	3.0	17.0	6.0	1.0	34.0	3.4	0.0	0.0	1.0	7.0	0.0	0.0	
8	26.0	1.0	26.0	3.0	0.0	0.0	0.0	1.0	2.0	1.0	7.0	0.0	0.0	
9	45.0	1.0	20.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
10	44.0	3.0	15.0	5.0	1.0	1.266972909	2.8	0.0	0.0	0.0	0.0	0.0	0.0	
11	44.0	3.0	26.0	4.0	0.0	0.0	0.0	1.0	2.0	0.0	0.0	0.0	0.0	
12	27.0	1.0	17.0	3.0	0.0	0.0	0.0	1.0	8.0	0.0	0.0	0.0	0.0	
13	45.0	4.0	14.0	6.0	0.0	0.0	0.0	1.0	10.0	1.0	5.0	0.0	0.0	
14	44.0	2.0	25.0	2.0	0.0	0.0	0.0	1.0	5.0	0.0	0.0	0.0	0.0	
15	43.0	2.0	18.0	5.0	0.0	0.0	0.0	0.0	0.0	1.0	8.0	0.0	0.0	
16	40.0	3.0	18.0	2.0	0.0	0.0	0.0	1.0	15.0	0.0	0.0	0.0	0.0	

Fig. 1. Cervical Cancer Risk Factor Dataset

In this Paper the data set is uploaded and the test data size is given to the model. Depending on test size the model prediction rate is changed. The prediction is performed using the input data entered by the user. The input data is the data taken from the user which contains the data used to predict the cervical cancer symptoms. The prediction data

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

includes age, partners, intercourse, pregnancies, smokes, smokes year, hormonal contraceptives, herpes, diagnosis, Dx_CINe, schiller, citology.
The classifiers used in this research are Support Vector Machine, Random Forest, CatBoost classifier, Extra Tree Classifier, and Logistic Regression classifiers. By merging classification outcomes from many classifiers using a majority vote technique, the best cervical cancer prediction is made.

4.1 Logistic regression: To predict the frequency of a binary dependent variable, logistic regression, an effective classification technique, is utilised. The parameter in logistic regression is a data point with data labelled as 1 or 0. Logistic regression and linear regression seem to be very related, in terms of how they are used. In contrast to logistic regression, which is used to address classification issues, linear regression addresses correlation issues. Logistic regression can be used to quickly pinpoint the variables that will be effective when categorising findings utilising a variety of data sources.

4.2 Support vector classifiers: The Support Vector Machine (SVM) technique, sometimes referred to as the SVM, is a simple yet efficient Supervised Machine Learning method will be utilized to develop both regression and classification models. By utilising the SVM approach, both linearly separable and non-linearly separable datasets can produce great results. Even with little input, the support vector machine method is still able to do miracles.

4.3 Random Forest: Random Forest algorithm is used to address classification and regression problems. It employs supervised techniques, a technique that integrates many classifiers to handle tough issues. A random forest method may employ any number of decision trees. The random forest method establishes a "forest", which is afterwards trained by rebound gathering or tagging. The ensemble situational bagging improves machine learning systems' accuracy. The algorithm chooses the outcome based on the forecasts that the decision trees offer. By averaging the outcomes from various trees, it provides predictions. With more trees, the outcome becomes more accurate.

4.4 CatBoost: CatBoost is a decision tree approach that uses gradient boosting. It was created by engineers and researchers at Yandex, and Yandex and many other businesses utilize it for different purposes, including search, recommendation systems, personal assistants, self-driving cars, weather forecasting, and many more. These businesses include Careem Taxi, Cloudflare, and CERN. Anybody may use it because it is open-source. XGBoost and LightGBM are being challenged by Catboost, the new kid on the block, who has only been around for a little over a year. Excellently, Catboost receives the highest ratings on the benchmark. Yet, this gain becomes notable and important when you examine datasets where categorical characteristics are heavily weighted.

4.5 Extra Tree Classifier: Extremely Randomized Trees Classifiers, often referred to as Extra Trees Classifiers, are an ensemble learning approach that generates results by combining the classification results of several single decision trees into a "forest". The only conceptual distinction between it and a Random Forest Classifier is how the decision trees in the forest are built. The initial training sample is used to build each decision tree in the Extra Trees Forest. In order to divide the data according to a predetermined mathematical criterion, the next step is for each tree to choose the best feature from a random sample of k features from the feature-set at each test node. This random selection of decision trees leads to a large number of de-correlated decision trees.

The initial training sample is used to build each decision tree in the Extra Trees Forest. The next step is to divide the data into subsets that satisfy specific mathematical criteria by having each tree choose the top feature from a random sampling of k features from the

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

feature-set at each test node. There are several de-correlated decision tees produced as a result of this arbitrary characteristic selection.

The image displays the suggested framework for cervical cancer prediction. By merging the classification outcomes from several classifiers using, a majority vote approach, the best cervical cancer prediction may be achieved.

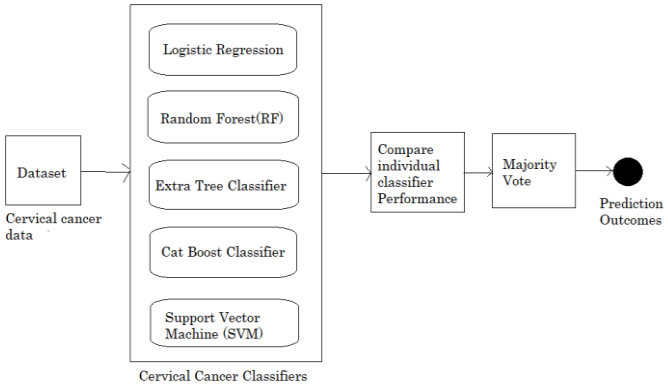

Fig. 2. Architecture of proposed system.

5 Experimental Results

There are 858 subjects and 36 attributes in the dataset. This research performed the required pre-processing to adjust for the missing data since the quality of the data has a significant impact on how well machine intelligence algorithms perform. Modes and medians were employed as a stand-in for the missing nominal and numeric properties in an unsupervised filter that omitted the class property.

Any of many similar model validation techniques, including cross-validation, can be used to test the generalizability of a statistical study's findings to a different data set. Out-of-sample testing or rotation estimation are other names for it. Cross-validation is a resampling method that use various data subsets to assess and train a model across a number of trials. It is frequently used to anticipate the future and assess how effectively a predictive algorithm could function under real-world conditions. A model is frequently evaluated in a prediction issue utilising a dataset of new dataset (or newly observed data) and a training dataset of known data (training dataset).

In order to identify issues like generalising or selection bias and to provide information into the manner in which the system will transfer to an information delivered, cross-validation analyses the model's propensity to predict new information that was not included in its estimate.
In this study we can apply different test sizes for different algorithms. If we give test set size as 20% then the training set size is 80%. When we apply testing sizes 20,30,40 on different models, then the accuracy obtained for each model is recorded below.

Table1. Accuracy of test models with different test sizes.

Model	Accuracy for 20% test size	Accuracy for 30% test size	Accuracy for 40% test size	Accuracy for 50% test size
Logistic Regression	95.03	95.6	94.39	94
Random Forest Classifier	96.3	97.7	97.5	97.5
Extra Tree Classifier	97.8	98.8	96.3	96.8

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

CatBoost Classifier	98.4	98.1	98.1	97
SVC	92.9	94	93	93

On applying different test sizes, we can consider CatBoost classifier as an accurate model for detecting cervical cancer cases at different test sizes because the accuracy of CatBoost classifier get stabled accuracy.

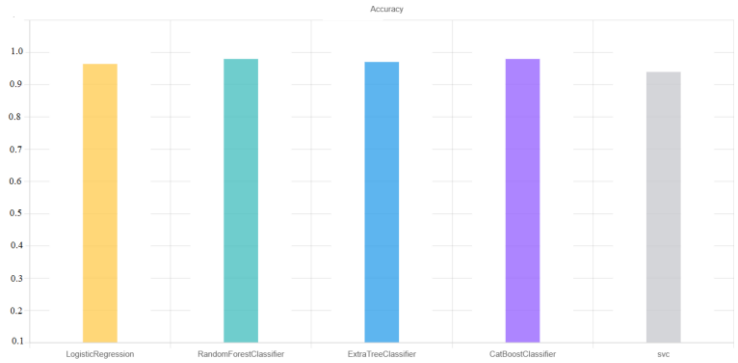

Fig. 3. Accuracy of models with test set size of 20%.

The above graph shows the accuracy of classification models that are used in prediction of cervical cancer symptoms. It contains accuracy of every model when the test set size is 20% and training set size is 80%.

The prediction accuracy of catboost classifier is high when compared to other classifiers when we apply test set size of 20%. And by using the majority vote of classifiers output the result is predicted.

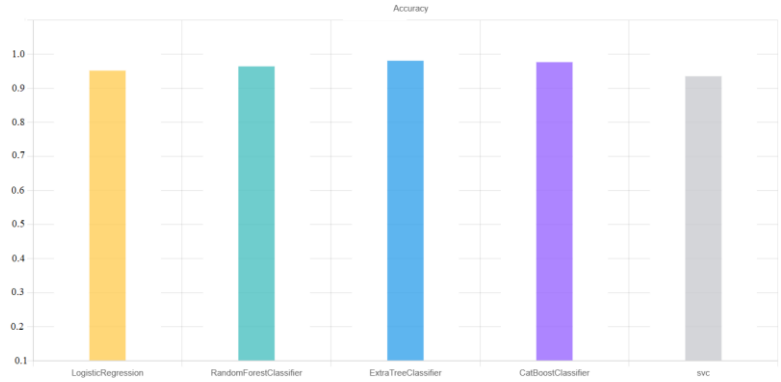

Fig. 4. Accuracy of models with test set size of 30%.

The above graph shows the accuracy of test models when we apply test set size of 30%. When we apply test set size of 30% the accuracy of Extra tree classifier is more when compared to other classifier models.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

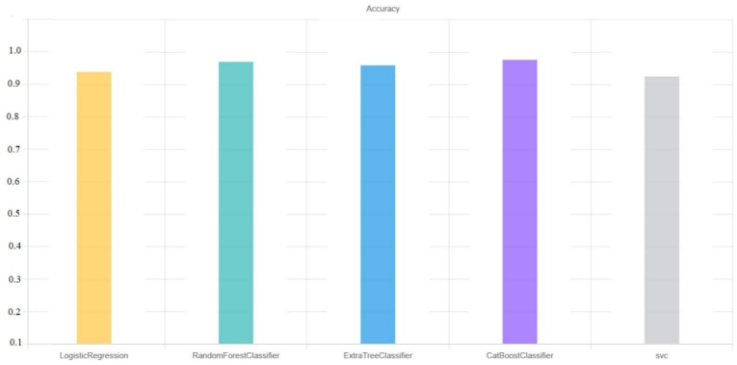

Fig. 5. Accuracy of models with test set size of 40%.

The Fig 5 contains the accuracy of models when we apply test set size of 40%. When we apply test set size of 40% the accuracy of CatBoost classifier is more when compared to other classifier models.

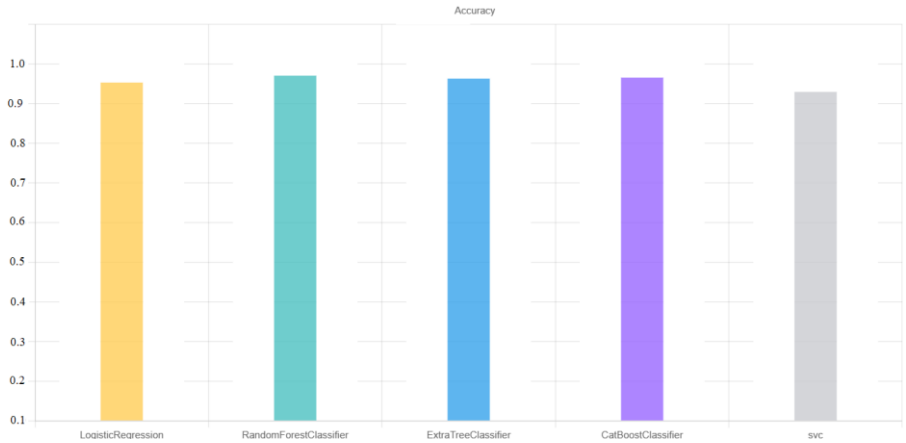

Fig. 6. Accuracy of models with test set size of 50%.

The above graph shows the accuracy of test models when we apply test set size of 50%. When we apply test set size of 50% the accuracy of Random Forest classifier is more when compared to other classifier models.

Whenever test set size changes the accuracy of every model is also changing. When we increase test set size some models accuracy is increasing and also some models accuracy is decreasing. By using ensemble model in this study, we are maintaining minimum of 96% accuracy with different test set sizes and different classification models.

6 Conclusion

The health standards need a precise and trustworthy cervical cancer diagnosis in order to preserve the patients' irreplaceable life. Notwithstanding the challenges and issues, machine intelligence-based solutions now in use are thought to be trustworthy. Yet, there is also a difficulty with how well the various categorization techniques work to find cervical cancer. Different categorization algorithms provide dramatically different outcomes when used on identical datasets because they are sensitive to the nature of the data. Based on a major voting method to accept the highest accuracy findings for cervical cancer detection, this paper provided an ensemble classifier model for diagnosis of cancer. SVM, Random Forest, CatBoost, Logistic Regression and Extra Tree classifiers were among the many classifiers used in the study. In performance evaluation, the proposed ensemble classifier beat individual classifiers and achieved the maximum accuracy of 96% when evaluated

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 12, Dec 2022

towards other predictors. The results of this study can be used by medical practitioners to provide cervical cancer patients with an informed and reliable second opinion in order to treat the disease more successfully.

7 Future Scope

In this paper we have used dataset of size 858 rows. And in future dataset size will be extended. And make the prediction available for different dataset formats.

References

[1]. Emmanuel Ahishakiye, Ruth Wario, Waweru Mwangi, Danison Taremwa 2020 Prediction of Cervical Cancer Basing on Risk Factors using Ensemble Learning.
[2]. Ye Rang Park, Young Jae Kim, Woong Ju, Kyehyun Nam, Soonyung Kim, Kwang Gi Kim Classification of cervical cancer using deep learning and machine learning approach.
[3]. Mehryar Mohri, Afshin Rostami Zadeh, Ameet Talwalkar Foundations of Machine Learning. MIT Press, Cambridge, MA, USA, 2018.
[4]. Jie Su, Xuan Xu, Yongjun He, Jinming Song 2016 Automatic Detection of Cervical Cancer Cells by a Two-Level Cascade Classification System. Anatomical Cellular Pathology.
[5]. Fangqi Li, Shilin Wang, Gongshen Liu 2019 A Bayesian Possibilistic C-Means clustering approach for cervical cancer screening.
[6]. Jiayong Zhang and Yanxi Liu 2004 Cervical Cancer Detection Using SVM Based Feature Screening.
[7]. Chuen-Horng Lin, Y. Chan and Chun-Chien Chen 2009 Detection and segmentation of cervical cell cytoplast and nucleus.
[8]. Mavra Mehmood, Muhammad Rizwad, Michal Gregusml, Sidra Abbas 2021 Machine Learning Assisted Cervical Cancer Detection.
[9]. Rebecka Weegar, M. Kvist, K. Sundstrom, S. Brunak, H. Dalianis 2015 Finding cervical cancer symptoms in swedish clinical text using a machine learning approach and NegEx.
[10]. B. Ashok, P. Aruna 2016 Comparison of feature selection methods for diagnosis of cervical cancer using SVM classifier.