# Detection of Cervical Cancer: A Pursuit of Machine Intelligence towards Sustainable Health

**Mr. Jagadeesh B N** [1]
Assistant Professor
East West Institute of Technology
Bangalore, India
Jagadeesh.nagaraj.nl@gmail.com

**MohimaSinha**[2]
UG Student
East West Institute of Technology
Bangalore, India
mohimasinha1994@gmail.com

**Richa Pandey**[3]
UG Student
East West Institute of Technology
Bangalore, India
imrichapandey02@gmail.com

**Manisha Kumari Sah** [4]
UG Student
East West Institute of Technology
Bangalore, India
manshaasah@gmail.com

**Abstract – Predicting the early stages of significant illnesses using ML-algorithms schemes, inclusive of cancer, kidney failure, and heart attacks, is increasing. Cervical cancer is one of the most frequent illnesses among ladies, and early analysis could be a possible answer for stopping this cancer. Therefore, this have a look at gives an astute manner to expect cervical most cancers with ML algorithms. research dataset, records pre-processing, predictive model choice (PMS), and pseudo-code are the 4 stages of the proposed studies technique. Hence, the proposed version bestows a second opinion to health practitioners for disorder identification and timely remedy.**

***Index Terms – Cervical cancer, machine intelligence, classification algorithms, support vector machine, Naïve Bayes, random forest, decision trees, K-nearest neighbors, ensemble classification.***

## 1. INTRODUCTION

Cervical cancer is a dangerous most cancers, which threatens women's fitness worldwide, and its early signs and symptoms are extraordinarily tough to come across. It is answerable for damaging deep tissues of the cervix and can regularly reach different areas of the human body, such as the lungs, liver, and vagina, that could growth the difficulties concerned.

The pre-invasive stages of cervical most cancers of the uterus stays for a long term. Screening tests can provide successful remedy of precancerous-level lesions, in order that most cancers can be avoided. while the cervix's human papillomavirus (HPV) contamination is left untreated, cervical cancer develops. These days, many research have been conducted on cervical cancer by using modern techniques that provide prediction in the early stage. The maximum vital reasons of this disease among girl populations are lack of understanding, lack of access to resources and clinical centres, and the fee of undergoing everyday exam in a few nations. Machine learning has advanced the overall performance of analyses and the generation of correct patient data.

**The objectives of this study are as follows:**

• To analyse and classify cervical cancer using machine learning algorithms that will

help doctors accurately diagnose the cancer.

• To identify the correlations between the parameters that are likely to be responsible for cervical cancer.

• To conduct a survey that identifies women's concerns about cervical cancer, and

## 2. RELATED WORKS

| S.no | Methods | Dataset | Advantages | Disadvantages |
|---|---|---|---|---|
| 1 | Inception V3 model | Herlev dataset | (i) High accuracy<br>(ii) Good universality<br>Low complexity | (i) The deep network needs further study to investigate cervical cells. |
| 2 | Transfer learning, pretrained DenseNet | Fujian Maternal and child health hospital Kaggle | (i) More feasibility and effective | (i) Limited data |
| 3 | CNN-extreme learning machine- (ELM-) based system | Herlev dataset | (i) Fast learning<br>(ii) Easy convergence<br>(iii) Less randomized | (i) More complexity<br>(ii) Need more investigation |
| 4 | Gene-assistance module, voting strategy | Chinese hospital and Universitario De Caracas, Venezuela | (i) More scalable and practical | (i) Limited datasets |
| 5 | Random forest and Adaboost | Radiotherapy dataset | (i) Better treatment planning | (i) Need to extract features<br>(ii) Painful treatment |
| 6 | ColpoNet | Colposcopy images | (i) Better accuracy<br>(ii) Efficient classification | (i) Need to improve accuracy by extracting relevant information |
| 7 | CNN Model | Papanicolaou-stained cervical smear dataset | (i) Better sensitivity and specificity | (i) Reported 1.8% false-negative images |
| 8 | Fourier transform and machine learning methods. | Microscopic images | (i) Fully automatic system<br>(ii) Saving precious time for the microscopist | (i) The level of complexity is more |
| 9 | CNN-SVM model | Herlev and one private dataset | (i) Good robustness<br>(ii) Highest accuracy | (i) Need improvement to adjust parameter<br>(ii) Need of hand-crafted features |
| 10 | Stacked Autoencoder | UCI database | (i) High accuracy<br>(ii) Reduced data dimension | (i) Training time is very high due to reducing the dimension |
| 11 | PSO with KNN algorithm | Cervical smear images | (i) Better accuracy<br>(ii) Good feature selection | (i) Time-consuming due to two-phase feature selection |
| 12 | Ensemble model | PAP smear image | (i) For 2 class problem achieves the accuracy of 96%<br>(ii) For 7 class problem achieves an accuracy of 78% | (i) Overall of cells are difficult to identify |
| 13 | Multimodal deep network | National Cancer Institute | (i) Good correlation<br>(ii) High accuracy<br>(iii) Learn better complementary features | (i) More complexity in image fusion |

This research observe proposes an Ensemble type method based totally on majority voting for an accurate diagnosis addressing the patient's clinical conditions or signs. The study records an enhancement in prediction accuracy of ninety-four% that outperforms the prediction accuracies of single classification methods.

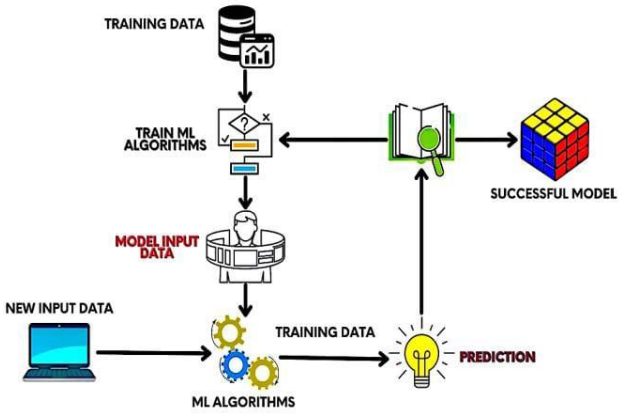
## 3. OVERVIEW OF THE PROJECT

The primary goal of this system is to detect the cervical cancer at early stage. It experiments with a range of classic machine learning methods, including decision tree (DT), logistic regression (LR), support vector machine (SVM), Artificial

Neural Network (ANN), AdaBoost classifier, random forest (RF). We apply the k-fold cross-validation and segment the data points into an aggregate of training, test, and validation segments. Later, we analyse instances with the chosen classifiers. We then define a filter to analyse the individual performance of classifiers. Finally, ensemble classification outcomes are achieved by a Majority Voting Filter.

## ARCHITECTURE:

Figure depicts an architectural diagram of the proposed research. The architectural diagram has been separated into four phases, because the model presented in this study performs some essential tasks in each level.




## 4. EXPERIMENTAL ANALYSIS

The proposed methodology is classified into four segments:

a) research dataset

b) data pre processing

c) predictive model selection (PMS)

d) training method

### *4.1. Research Dataset*

The UCI repository contributed to the dataset "Cervical cancer risk factors for Biopsy". the gathering contains facts approximately 858 people's activities, demographics, and clinical records. a couple of lacking values arise on this dataset for sanatorium patients as a result of several patients declining to reply questions because of privacy concerns.

### *4.2. Data Preprocessing*

Data preprocessing is divided into three sections, that are as follows: data cleaning, data transformation, and data reduction. data preprocessing is critical because it directly affects project success. The data transformation level is stored in place to trade the data into suitable forms for the mining manner. This research combines normalization, attribute selection, discretization, and concept hierarchy generation.

### *4.3. Predictive Model Selection (PMS)*

This segment has highlighted a number of the algorithms that have done a satisfactory level of accuracy at the adopted studies dataset. thus, we have illustrated the theoretical interpretation of these algorithms within the following subsections.

#### 4.3.1. Decision Tree (Dt)

Both classification and regression problems can be solved with the classification and regression tree or CART algorithm, which is also called the DT. The DT seems plenty like the branches of a tree, which is why the word 'tree' is included in its name. The decision tree begins from the 'root node' simply because the tree starts from the root. From the root node, the branches of this tree spread through different decision situations; such nodes are called decision nodes (and referred to as leaf nodes after creating a very last decision).

#### 4.3.2. Random Forest (Rf)

In this algorithm, the decision is made through voting. Such an algorithm is known as ensemble learning. Random forests are made up of many trees or shrubs. simply as there are numerous trees inside the forest, random forests also have many decision trees. The decision that maximum trees make is taken into consideration as the final decision.

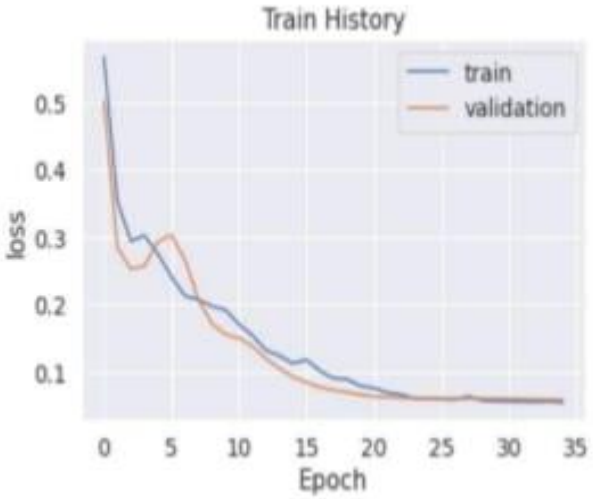
#### 4.3.3. Support Vector Machine (SVM)

In this algorithm, data points are separated with the aid of a hyperplane, and the kernel determines what the hyperplane will look like. If we plot more than one variable in an ordinary

scatter plot, in many cases, that plot cannot separate two or more data classes. The kernel of an SVM is an extensive element, which can convert lower-dimensional data into higher-dimensional area, and consequently differentiate among types.

### *4.4. Training Method*

The support vector machine (SVM) performs nicely on linear and nonlinear data. This method of classifying nonlinear data includes the radial base function. Setting data inside the function area is based closely at the kernel function while plotting many variables in an average scatter plot, it is often not possible to distinguish among various sets of data. An SVM's kernel is a way for transforming lower-dimensional input into higher-dimensional space and figuring out exceptional classes. Similarly, the radial basis characteristic is a nonlinear function. The support vector machine's most famous feature is its capacity to classify items automatically.





## 6. RESULT AND DISCUSSION

This section is categorized into four components:

a) Empirical Consequence Report (ECP)
b) Exploratory Cervical Data Analysis (ECDA)
c) Computational Complexity Analysis (CCA)
d) Comparative Analysis
e) Survey Data Analysis (SDA)

### *6.1. Empirical Consequence Report (ECP)*

The accuracy of predictions from the category algorithms is expected by applying a classification record. The record demonstrates the precision, recall, and f1 score of the key

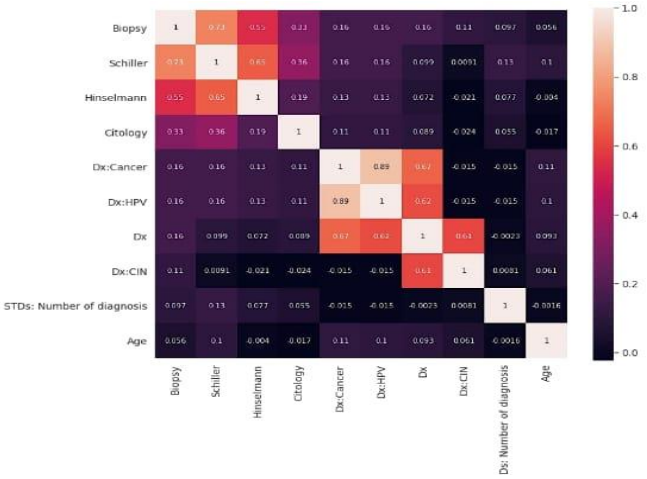
classification metrics on a in step with-class basis.

### *6.2. Exploratory Cervical Data Analysis (ECDA)*

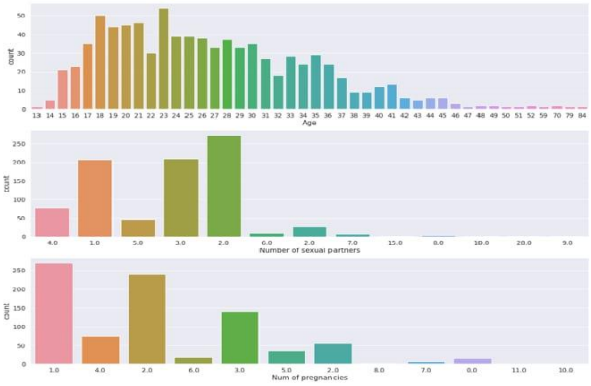
Figure a suggests the correlation graph. Correlation describes how two or more variables are connected. these variables may be input data functions used to forecast our target variable.

Correlation is a mathematical approach used to evaluate how one variable moves or shifts with regards to every other.

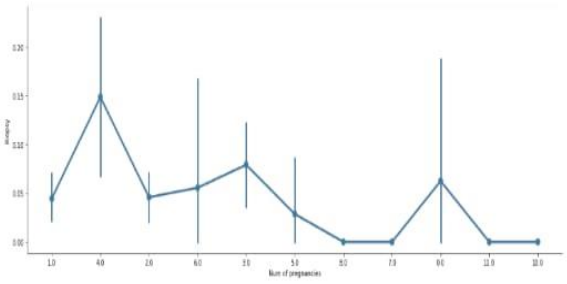
Figures b and c visualize the count measurement concerning the number of pregnancies, the number of sexual partners, and age, and a contrast between biopsy and number of pregnancies.



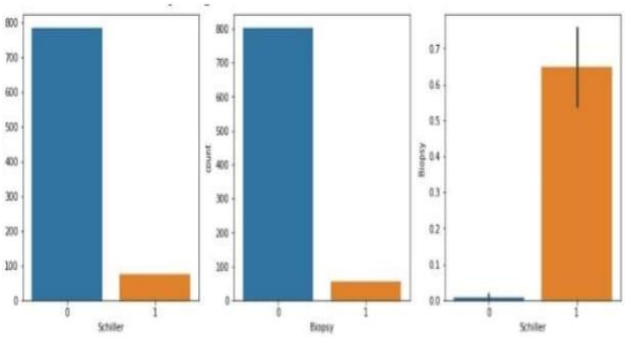

Fig(a): Correlation between different variables




Figure(b): Count measurement



Figure(c): Data visualization of biopsy vs no. of pregnancies




Figure(d): comparison between Schiller and Biopsy

### *6.3. Computational Complexity Analysis (CCA)*

Machine Learning computational complexity is a quantitative examination of the opportunities for effective computer learning. It is targeted on a hit and fashionable gaining knowledge of algorithms and works inside currently deployed machine inference models based totally on computational complexity concept.

### *6.4. Comparative Analysis*

This research has implemented cross-validation, which is a method that examines the research model to gain better residuals. The trouble with validation is that it does not imply how desirable data will be when it is used to make new estimates for a new end result.

*6.5. Survey Data Analysis (SDA)*

Some other part of our studies is conducting survey data analysis. To determine how many people are aware of cervical cancer, we've got finished survey questionnaires primarily based on the intention of this studies. In this studies, a stratified sampling approach has been used; stratified sampling is a similar or homogenous group-primarily based sampling technique.

## 7. CONCLUSION

Early detection increases the chance of successful treatment in the pre-cancer and cancer stages. Being aware about any symptoms and signs of cervical cancer also can aid in averting diagnostic delays. This research has centred on cervical cancer using traditional machine learning (ML) principles and several conventional machine learning algorithms, like decision tree (DT), logistic regression (LR), support vector machine (SVM). The results of those algorithms are carried out to perceive the most relevant predictors. We have acquired best accuracy compared to the support vector machine algorithm. The findings of this study discovered that the SVM model may be used to discover the important predictors. As the number of essential predictors for analysis decreases, the computational value of the proposed model decreases. The disease may be predicated more appropriately with the usage of machine learning. Furthermore, boosting patients' private health and socio-cultural status can lead to cervical cancer prevention.

## ACKNOWLEDGEMENT

We sincerely thank to our Project Guide and Coordinator who provided unwavering assist, as well as our college for providing cheer throughout the project.

## REFERENCES

[1] Martin, C.M.; Astbury, K.; McEvoy, L.; Toole, S.; Sheils, O.; Leary, J.J. Gene expression profiling in cervical cancer: Identification of novel markers for disease diagnosis and therapy. In Inflammation and Cancer; Springer: Berlin, Germany, 2009; Volume 511, pp. 333–359.

[2] Purnami, S.; Khasanah, P.; Sumartini, S.; Chosuvivatwong, V.; Sriplung, H. Cervical cancer survival prediction using hybrid of SMOTE, CART and smooth support vector machine. AIP Conf. Proc. 2016, 1723, 030017

[3] Yang, X.; Da, M.; Zhang, W.; Qi, Q.; Zhang, C.; Han, S. Role of lactobacillus in cervical cancer. Cancer Manag. Res. 2018, 10, 1219–1229.

[4] Ghoneim, A.; Muhammad, G.; Hossain, M.S. Cervical cancer classification using convolutional neural networks and extreme learning machines. Future Gener. Comput. Syst. 2020, 102, 643–649.

[5] Mao, Y.J.; Lim, H.J.; Ni, M.; Yan, W.H.; Wong, D.W.C.; Cheung, J.C.W. Breast Tumour Classification Using Ultrasound Elastography with Machine Learning: A Systematic Scoping Review. Cancers 2022, 14, 367.

[6] Šarenac, T.; Mikov, M. Cervical cancer, different treatments and importance of bile acids as therapeutic agents in this disease. Front. Pharmacol. 2019, 10, 484–513.

[7] Alam, T.M.; Khan, A.; Iqbal, A.; Abdul, W.; Mushtaq, M. Cervical cancer prediction through

different screening methods using data mining. Int. J. Adv. Comput. Sci. Appl. 2019, 10, 346–357.

[8] Lu, L.; Song, E.; Ghoneim, A.; Alrashoud, M. Machine learning for assisting cervical cancer diagnosis: An ensemble approach. Future Gener. Comput. Syst. 2020, 106, 199–205.

[9] Prabhpreet, K.; Gurvinder, S.; Parminder, K. Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification. Inform. Med. Unlocked 2019, 16, 100151.

[10] Prabhpreet, K.; Gurvinder, S.; Parminder, K. Intellectual detection and validation of automated mammogram breast cancer images by multi-class SVM using deep learning classification. Inform. Med. Unlocked 2019, 16, 100151.