A Cognizant Study of Machine Learning in Predicting Cervical Cancer at various Levels-A Data Mining concept

S. Geeitha[1] and M. Thangamani[2]
[1]*Assistant Professor, Department of Information Technology,*
Mahendra Engineering College for women Tiruchengode, India.
[2]*Professor, Department of Computer Science Engineering,*
Kongu Engineering College, Tamilnadu, India.

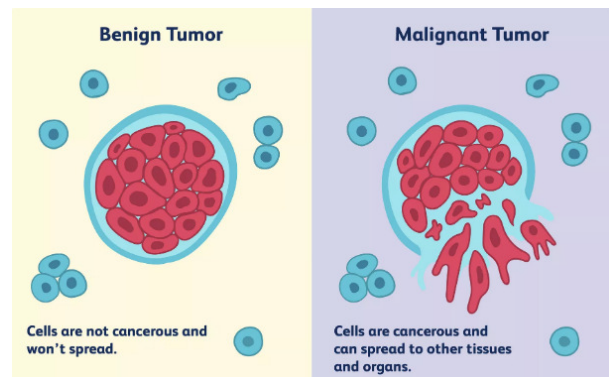*(Corresponding author: Geeitha S.)*

ABSTRACT: A lot of machine learning tools are being exploited in the biomedical field for discovering various patterns and predicting cervical cancer among women. This paper is designed to present the current research undergone in cervical cancer analysis. The main objective of this paper is to procure the knowledge of machine learning algorithms and data mining techniques in handling the cervical cancer dataset. In this study, many machine learning algorithms in analyzing gene expressions and screening cervical cancer by integrating different data mining techniques have been discussed and provide results on comparison of different statistical parameters of various mining techniques. The paper also reveals the importance of feature selection and classification techniques in categorizing the benign and malignant cervical cancer genes and discusses the various parameters that measure the prediction accuracy using data mining tools.

Keywords: Cervical Cancer, Data mining, Machine Learning, Gene Expression, prediction.

Abbreviations: CC, Cervical Cancer, HPV, Human Papilloma Virus, SVM, Support Vector Machine, SVMRFE, Support vector machine Recursive Feature Elimination, SVM PCA, Support vector machine principal component analysis, LR, Logistic Regression, DT, Decision tree, RF, Random Forest, GA, Genetic Algorithm, Knn, K nearest neighbor, NB, Naïve Bayes, NN, Neural Network, CART, Classification and Regression Tree, CC, Classifier chains, LC, label combination, BR, binary relevance, CDN, conditional dependency network, BAHSIC, Backward Elimination Hilbert-Schmidt Independence Criterion, EVD, Extreme Value Distribution based gene selection, SVD, Singular Value Decomposition Entropy, PAT, pap-smear analysis tool, SMOTE, Synthetic Minority Oversampling Technique, IUD, Intra uterine device, HIV, Human Immune Virus, HPV, Human Papilloma Virus, FCM, Fuzzy C means.

I. INTRODUCTION

Generally, cancer disease has been determined as a great cause of mortality and morbidity in the field of the medical era [1, 2]. Nearly 8.2 million among the world's population die due to cancer disease where the death rate falls to 13% worldwide. This death rate is because of the lack of screening services in the under developing countries compared to the developed countries where it varies between 26% and 90% respectively. It has been expected that in 2030, cancer incidences may fall up to 22 million [3, 4]. Cervical cancer is considered is known to the third most serious malignancy among worldwide. This cancer occurs due to the abnormal growth of the tissues and organs in the cervix. The origin of cervical cancer initiates with squamous cells [5]. It then gradually develops to the precancerous stage known as cervical dysplasia. If it remains undetected, it leads to the malignancy stage where it spreads to intestines, liver and other parts of the body uncontrollably otherwise it is detected as a benign stage where it might not cause cancer but with treatment, it can be diffused. Fig. 1 depicts the benign and malignant stage of cancer cells. Cervical cancer has a major effect especially in women belonging to poor resources that lack public health infrastructure to carry out the screening process [6].

Fig. 1. Cancerous and Non-Cancerous Tumor-Illustrated by Joshua Seong, 2017.

Cervical Cancers are mostly caused by the Human Papilloma Virus (HPV) commonly defined as a sexually transmitted virus. It remains a major cause for cervix, vaginal cancer, in and around other parts of the cervix [7]. Recent tests prove that HPV testing provides good protection against malignant cervical cancer rather than Pap smear screening [8, 9]. There occur many risk factors that influence cervical cancer such as HPV infection, sexual contraceptive, age, smoke, etc., among which the former factor plays a vital role in developing cervical cancer [10].

Screening plays a major part in identifying the progress of cervical cancer which mainly includes four tests namely Pap smear test, biopsy test, schiller test, and Hinslemann test. With the help of these screening associated with the risk factors, cervical cancer can be easily analyzed.

In machine learning a large volume of data can be well normalized using preprocessing analysis which can be easily accessible for cancer detection. Cervical data set can be downloaded from Herlev University hospital, kaggle, Bilkent university and Hacetepe university hospital, and risk factor dataset can be downloaded from UCI repository (from Hospital Universitario de Caracas) [11] containing Geo accession number for each series of dataset such as GSE5787, GSE3578, GSE10372, etc. with different expression platform that comprises Affymetrix Human Genome, Human Whole Genome Bioarray, SentrixHuman-6 Expression Bead Chip respectively. There exists some predefined gene dataset titled BIOCARTA, KEGG and REACTOME belonging to the category Pathway [12]. Besides text datasets, image data sets are also available such as the Herlev dataset, Haceptte data set. Machine learning algorithms have become a growing trend in the biomedical field. In this paper, we present a review of various machine learning models implemented in cervical cancer predictions and its prognosis factor. Feature selection in cancer genes is made dependent on certain treatment which paves the way to a clinical decision for cervical cancer patients [13]. Here, we discuss various types of machine learning tools integrated along with the cervical cancer dataset and also examine the performance measure of various algorithms implemented relevant to the feature analyzes.

Generally, machine learning models have occupied its space in the biomedical area with many tremendous applications where it makes it search in the n-dimensional space for the given dataset samples using a variety of techniques and algorithms [14]. Process analytics model can be implemented with a machine which is self-automated and used for taking dynamic decision that can be enhanced using big data with IOT [15]. Supervised learning includes a labeled set of training data whereas unsupervised learning includes non labeled dataset. Machine learning methods comprises many different techniques for converting the raw data suitable for further process such as preprocessing to fit properly into the model. It also includes some approaches such as dimensionality reduction, feature extraction and feature selection using both classification and clustering algorithms. Support Vector Machine (SVM), Logistic Regression, Random Forest, Decision Tree (DT), Naïve Bayes (NB), k-Nearest Neighbour (kNN), Neural Network (NN), Genetic Algorithm (GA) are categorized as classification techniques in data mining [16]. K-means clustering, Density-based spatial clustering, hierarchical clustering, Expectation maximization clustering comes under clustering techniques.
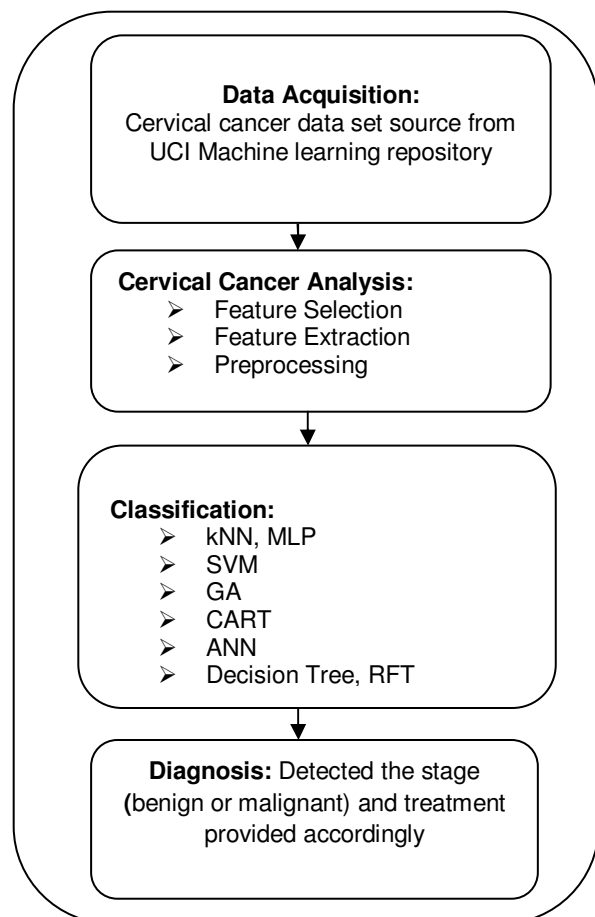
II. MATERIALS AND METHODS

Many surveys on cervical cancer data are taken from Geo data set which contains thousands of samples with different series under different platform id from which each sample contains thousands of subsamples with different stages of cancer such as benign and malignant tumors. Table 1 shows comparison analysis of different classification techniques used for cancer classification.

Table 1: Survey on Machine learning techniques in classifying cervical cancer data.

Paper Title	Data source	Technique	Proposed Model	Limitations	Performance
Kourou *et al.,* (2015) [17]	Clinical, Pathologic cervical cancer data from 168 patients	SVM, Hold out method, BN	SVM produces more accurate prediction accuracy than compared to other models	Efficient techniques needed for integration of multidimensional & heterogeneous data	68% accuracy
Lakshmi and Krishnaveni (2016) [18]	Images from Herlev University database	Mining techniques such as Multilayer perceptron, Bayes Classifier, SVM classifier, Bayesnet	Dimensionality reduction of a feature set is done with correlation-based techniques with seven class classifier with 11 features	Only 11 features are extracted for classification and these selected features are class - dependent	90% accuracy
Wu and Zhou (2017) [19]	Dataset collected at hospital 'Uiversitario de Caracas'	SVM Classification method with SVM-RFE and SVM-PCA	SVM-RFE and SVM-PCA reduces the number of features that improved the classification speed	Both methods bare high computational costs	93.79% accuracy
Ceylan and Pekel (2017) [20]	UCI Repository	MEKA tool, Multilabel Classification methods(NB,SMO,J48 and RF)	Classifier Chains(CC) & label combination(LC) were implemented with Hamming Loss and Rank Loss that produced different results	The evaluation measures were not satisfied with J48-binary relevance(BR) and J48-CDN(conditional dependency network)	80% accuracy
Benazir and Nagarajan (2018) [21]	UCI Repository	Genetic Algorithm, Particle Swarm Optimization, Ant Colony Optimization	By ANN and MLP back propagation algorithm outperforms in classifying the features into a normal abnormal stage of cervical cancer	Only Mean Square Error and Regression values are compared	Produces less error rate. MSE-1.2365e-13
Kaur and Ginige	Kaggle data set	K-NN, SVM, Logistic	ANN and ensemble-	Annotated data	

(2018) [22]	consisting of 1428 samples with 714 benign and 714 malignant samples with 29 attributes each	regression, ANN, Decision Tree and Ensemble bagged based classifiers	based classifiers outperform other classifiers in terms of Precision, recall, F1score, and accuracy	was not available, Data available in small size	93% accuracy
Tarek *et al.*, (2017) [23]	Leukemia dataset, Colon dataset and Breast cancer dataset	KNN Classifier, Backward Elimination Hilbert-Schmidt Independence Criterion(BAHSIC), Extreme Value Distribution based gene selection(EVD), Singlular Value Decomposition Entropy gene selection (SVD Entropy)	Ensemble system outperforms among the performance of another base classifier	Only one classifier is used as base member	Error rate 0.00%
Alam *et al.*, (2019) [24]	UCI repository	SMOTE Technique, Boosted Decision Tree, Decision jungle , Decision forest	Boosted decision Tree outperformed in prediction ability. Imbalance data set problem solved by SMOTE	AUROC curve disqualified the predictive qualifiers	94% accuracy
William *et al.,* (201) [25]	Harlev pap-smear images, Mbarara Regional Referral Hospital	WEKA segmentation Classifier, Sequential Elimination Approach, Fuzzy C-means Algorithm	PAT reduces the screening work, analyzes full pap smear	The cost of PAT seems to be high to meet the needs of developing countries	98.88% accuracy

Fig. 2. Flow diagram for cervical cancer Data Classification.

Fig. 2 shows the four phases in classifying cervical cancer data that include data acquisition, cervical cancer analysis, classification and diagnosis where the first phase explains the sources of cervical cancer dataset, the second phase depicts three processes that comprise feature selection, feature extraction, preprocessing. The Third phase includes the different classifiers implemented for classifying the cervical gene dataset. And the fourth phase detects the stages of cancer and proceeds with diagnosis and treatment. Cervical cancer dataset can be downloaded from the UCI machine learning repository. The dataset contains 36 attributes among which attributes that hold Boolean values are depicted in Table 2.

Table 2: Boolean attributes of cervical cancer dataset.

S.No.	Attributes of Cervical cancer
1.	Smokes
2.	Hormonal Contraceptives
3.	IUD
4.	Condylomatosis
5.	Cervical condylomatosis
6.	Vaginal condylomatosis
7.	Vulvo- perineal condylomatosis
8.	Syphillis
9.	pelvic inflammatory disease
10.	genital herpes
11.	molluscum contagiosum
12.	AIDS
13.	HIV
14.	Hepatitis B
15.	HPV

The proposed method mainly focuses on KNN and SVM besides the existing classifiers such as ANN, GA, CART etc. But they fail to predict the key genes to identify the cervical cancer stage which is done by MLP where gene network is constructed to identify the specific feature of the cancer genes. In preprocessing, balancing the data is handled by SMOTE. The following algorithm is used to predict the accuracy of the classifier from the cancer dataset.
Step 1: Preprocessing using KNN and FCM.
Step 2: Dataset balancing is carried out by SMOTE method.

Step 3: Identifying the key genes using network construction.
Step 4: Calculate the fitness value for training and testing samples.
(i) Choose k samples for which training samples are generated (S_{tr})
(ii) Mean while testing samples are taken separately (S_{ts}).
(iii) Find the correlation between two samples using Euclidean distance such as

$$((Str_1 - St_{s1})^{p} + (S_{tr2} - St_{s2})^{p})^{1/p},$$

where p takes the value 2 as a fitness value.
Step 5: Feature genes are selected with MLP classifier.
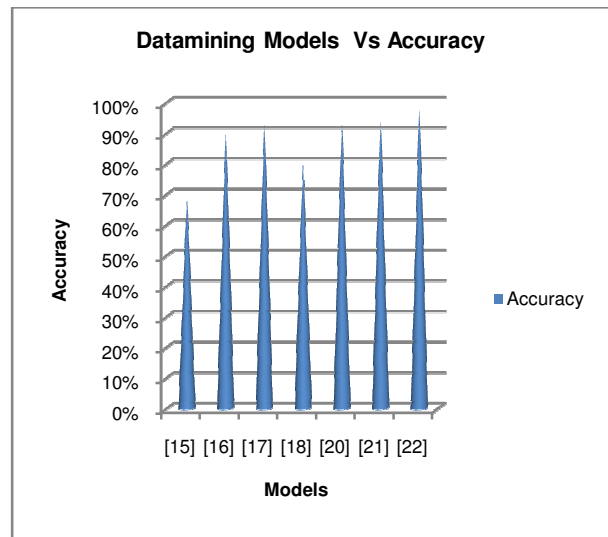Step 6: Cervical cancer genes are classified using an SVM classifier.
Many classification models were performed on the cervical cancer data set and their performance was evaluated based on a few metrics such as precision, recall, accuracy, etc.

III. RESULTS AND DISCUSSION

Table III shows the evaluation of the different classifier models implemented in various papers on the cervical cancer dataset based on the performance parameters. In author carried out 10 independent runs for testing the models using the variables such as pathologic stage, pathologic T, Cell Type and RT (Regulatory T cells) target [15]. In classification accuracy is obtained by computing TPR (True Positive Rate) and TNR (True Negative Rate) [16]. In produced accuracy by classifying the segmented MRI images into normal and abnormal cervical cancers with three different metrics among which accuracy produces 93% [17]. In performance of the classifier is obtained using four measures such as accuracy, exact match ratio, hamming loss and rank loss [18]. In the classifier accuracy is predicted based on correct and incorrect classification of the benign and malignant cervical samples using confusion matrix [20]. In classification is done using 3 or 5 nearest neighbors which is given as input to BAHSIC algorithm and produced the accuracy result of about 94% [21]. In classification accuracy is achieved using different target variables of cervical cancer risk factors and by using SMOTE algorithm that solves the imbalanced dataset problem [22]. Fig. 3 shows the comparative analysis of the various models based on their accuracy.

Table 3: Performance comparison of various Models.

Authors	Data mining models	Accuracy
Santhoshkumar *et al.,* (2019) [15]	SVM, Holdout method, BN	68%
Sarvaiya *et al.,* (2019) [16]	MLP,SVM, BN	90%
Kourou *et al.,* (2015) [17]	SVM-RFE, SVM-PCA	93%
Lakshmi & Krishnaveni (2016) [18]	NB, SMO, J48	80%
Ceylan and Pekel (2017) [20]	ANN, Ensemble baged tree	93%
Benazir and Nagarajan (2018) [21]	SMOTE, Boosted DT,DF	94%
Kaur and Ginige (2018) [22]	WEKA segmentation, FCM	98%

Fig. 3. Performance of various data mining models with cervical cancer dataset.

Machine learning techniques mainly focus on a particular model that performs various processes such as classification, prediction, selection, estimation, etc., The most probably implemented technique in machine learning is a classification method where training and testing are undergone with the given dataset. The former technique mainly produces some errors on the training data and expected errors in the testing data. A classification model is dependent on its training set rather than the training error and also in classifying the instances. If the classification model is properly obtained in the machine learning technique then it becomes easy to estimate its performance by analyzing the metric parameters. Classification accuracy varies in terms of algorithms also. Table 4 shows the performance of the different classifiers. In this study, the performance of the algorithm varies with different parameters. SVM produces 50% precision and 87% recall, KK produces 84% precision and 90% Recall value. ANN produces 93% precision and 76% recall value. Similarly, other algorithms follow. Among all Ensemble bagged tree outperform by producing 95% of precision and 91% of recall. Fig. 4 shows the performance metrics of different classifiers in terms of precision and recall.

Table 4: Classifiers Performance in terms of Precision and Recall.

Classifier	Precision	Recall
SVM	50%	87%
kNN	84%	90%
ANN	93%	76%
Logistic Regression	58%	48%
Ensemble Bagged tree	95%	91%
Decision Tree	88%	89%

Training samples are generated using different classifiers in terms of 10, 20, 30, 40 and 50 and taken for various parametric comparisons [Table 5] and measured in terms of Precision, Recall and Accuracy where for all training dataset, SVM classifier produced higher accuracy compared to other methods. Fig. 5 shows five methods for a different number of data samples that produce the metric measures.
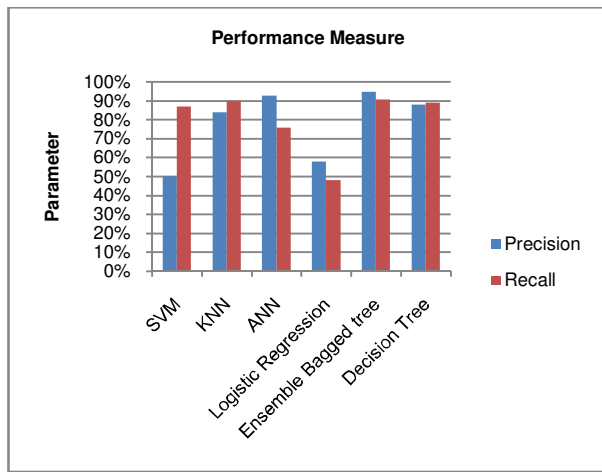
Fig. 4. Comparative Analysis in terms of Precision and Recall.

Table 5: Parameter versus number of Samples.

Training Samples	Methods	Precision	Recall	Accuracy
10	SVM	0.85	0.75	0.95
20	kNN	0.74	0.8	0.87
30	NB	0.65	0.76	0.91
40	CART	0.9	0.67	0.92
50	RF	0.8	0.7	0.9

Fig. 5. Parameter versus Methods.

Table 6 shows the prediction accuracy of different classifiers. SVM and Logistic regression provide 50% to 56% of accuracy. Algorithms such as Naïve Bayes, KNN, ANN and decision tree provide 80% to 88% accuracy. MLP and ensemble bagged tree produces an accuracy of 90% to 93%. CART and Random forest classifier outperforms by obtaining 94% accuracy comparing with other algorithms**.** Fig. 6 shows the classification analysis in terms of prediction accuracy for different data samples.

Table 6: Prediction Accuracy of different classifiers.

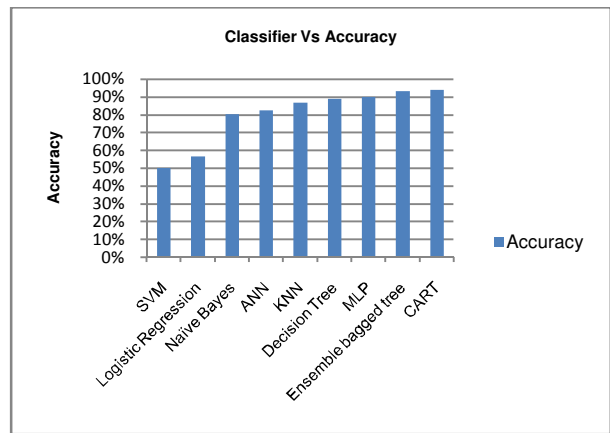| S.No.	Classifiers	Prediction Accuracy
1.	SVM	50%
2.	Logistic Regression	56.70%
3.	ANN	82.40%
4.	kNN	86.70%
5.	Decision Tree	88.90%
6.	Ensemble bagged tree	93.10%
7.	Naïve Bayes	80.20%
8.	CART	94%
9.	MLP	90%
10.	Random Forest	94%

Fig. 6. Classification analysis in cervical cancer Gene data set.

IV. CONCLUSION

In this paper, a comparative analysis of the machine learning models associated with classification techniques is been presented. Machine learning concepts have achieved a significant perception in the research areas by deploying multiple classification methods and their techniques. This paper assesses various classifiers and also machine learning models comparatively by analyzing the performance of methods using metric parameters such as precision, recall, and accuracy. With the cervical cancer dataset downloaded from the UCI browser, many algorithms implemented for feature selection and classification. The proposed method also survey on various models implemented from the year 2015 to 2019 where a different set of algorithms have been proposed in each paper. Machine learning models such as hold out method, Backward Elimination Hilbert-Schmidt Independence Criterion (BAHSIC), Singular value decomposition Entropy gene selection (SVD Entropy), SMOTE technique, J48, Random tree forest, Boosted Decision tree, Sequential elimination approach, and Classifiers such as SVM, Naive Bayes, Logistic regression, KNN, ANN, MLP, WEKA segmentation classifier, etc., have been implemented. For feature selection Particle swarm optimization, Extreme Value Distribution based gene selection (EVD), Ant Colony Optimization have been furnished in this study. The proposed model presents comparative analysis of various classifiers produced in terms of metric measures. Though various algorithms such as SVM, KNN, NB, CART, and Random Forest are implemented for cancer classification, the proposed model proves that SVM produces higher prediction accuracy along with KNN and SMOTE deployed for imbalanced data. The specified classifiers and algorithms reveal the fact that machine learning models have significant implications in medical science, especially in cancer analysis and diagnosis.

V. FUTURE SCOPE

Machine learning is considered as best pathologist for cancer disease prediction as the clinical decision. Besides cervical cancer prediction, feature selection and extraction becomes a major challenge in cancer analysis. Thus this work can be taken for deep learning with large number of dataset with dimensionality reduction.

REFERENCES

[1]. Sylla, B. S., & Wild, C. P. (2012). A million Africans a year dying from cancer by 2030: what can cancer research and control offer to the continent?. *International journal of cancer*, *130*(2), 245-250.

[2]. Jedy-Agba, E., Curado, M. P., Ogunbiyi, O., Oga, E., Fabowale, T., Igbinoba, F., ... & Osinubi, P. (2012). Cancer incidence in Nigeria: a report from population-based cancer registries. *Cancer epidemiology*, *36*(5), 271-278.

[3]. Kamil, N., & Kamil, S. (2015). Global cancer incidences, causes and future predictions for subcontinent region. *Systematic Reviews in Pharmacy*, *6*(1), 13-17.

[4]. Siegel, R. L. Miller, K. D., & Jemal, A. (2017). Cancer statistics, 2017, CA: *A cancer journal for clinicians, 67*(1), 7-30.

[5]. Soumya, D., & Arun Kumar, R. (2011). A brief assessment on cervical cancer. *Journal of Cancer Scinece and Therapy*, *17*, 1-5.

[6]. World Health Organization (2018). Cervical cancer: estimated incidence, mortality and prevalence worldwide in 2012. *International Agency for Research on Cancer website*. http://globocan.iarc.fr /old/FactSheets/cancers/cervix-new.asp.

[7]. Shehata, M. F., & Pater, A. (2011). Human Papillomavirus (HPV) Vaccine: Is it worthwhile. *J Biotechnol Biomaterial*, *1*, 103e.

[8]. Ronco, G., Dillner, J., Elfström, K. M., Tunesi, S., Snijders, P. J., Arbyn, M., ... & Berkhof, J. (2014). Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials. *The lancet*, *383*(9916), 524-532.

[9]. Sankaranarayanan, R., Nene, B. M., Shastri, S. S., Jayant, K., Muwonge, R., Budukh, A. M., ... & Chinoy, R. (2009). HPV screening for cervical cancer in rural India. *New England Journal of Medicine*, *360*(14), 1385-1394.

[10]. Gadducci, A., Barsotti, C., Cosio, S., Domenici, L., & Riccardo Genazzani, A. (2011). Smoking habit, immune suppression, oral contraceptive use, and hormone replacement therapy use and cervical carcinogenesis: a review of the literature. *Gynecological Endocrinology*, *27*(8), 597-604.

[11]. Zhou, L., Zheng, W., Luo, M., Feng, J., Jin, Z., Wang, Y., ... & He, Y. (2014). dbCerEx: A Web-Based Database for the Analysis of Cervical Cancer Transcriptomes. *PloS one*, *9*(6), 1-5.

[12]. Polley, M. Y. C., Freidlin, B., Korn, E. L., Conley, B. A., Abrams, J. S., & McShane, L. M. (2013). Statistical and practical considerations for clinical evaluation of predictive biomarkers. *Journal of the National Cancer Institute*, *105*(22), 1677-1683.

[13]. Niknejad, A., & Petrovic, D. (2013). Introduction to computational intelligence techniques and areas of their applications in medicine. *Med Appl Artif Intell*, 51-70.

[14]. Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, *2*, 59-77.

[15]. Santhoshkumar, S. Mohamed, A. T., & Ramaraj, E. (2019). Process Analytics Model for Health Care using IOT and Big Data Techniques, *International Journal on Emerging Technologies, 10*(4), 197-200.

[16]. Sarvaiya, L. Yadhav, H., & Agrawal, C. (2019). A Literature review of Diagnosis of Heart Disease using Data mining techniques. *International Journal of Electrical, Electronics and Computer Engineering* , *8*(1), 40-45.

[17]. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, *13*, 8-17.

[18]. Lakshmi, G. K. & Krishnaveni, K. (2016). Feature Extraction and Feature set selection for cervical cancer diagnosis. *Indian Journal of Science and Technology, 9*(19), 1-7.

[19]. Wu, W., & Zhou, H. (2017). Data-driven diagnosis of cervical cancer with support vector machine-based approaches. *IEEE Access*, *5*, 25189-25195.

[20]. Ceylan, Z., & Pekel, E. (2017). Comparison of multi-label classification methods for prediagnosis of cervical cancer. *International Journal of Intelligent Systems and Applications in Engineering*, *5*(4), 232-236.

[21]. Benazir, B., & Nagarajan, A. (2018). An Expert System for predicting the cervical cancer using data mining techniques. *International Journal of Pure and Applied Mathematics, 118*(20), 1971-1987.

[22]. Kaur, R., & Ginige, J. A. (2018). Comparative Evaluation of Accuracy of Selected Machine Learning Classification Techniques for Diagnosis of Cancer: A Data Mining Approach. *World Academy of Science, Engineering and Technology: International Journal of Biomedical and Biological Engineering*, *12*(2), 19-25.

[23]. Tarek, S., Elwahab, R. A., & Shoman, M. (2017). Gene expression based cancer classification. *Egyptian Informatics Journal*, *18*(3), 151-159.

[24]. Alam, T. M., Khan, M. M. A., Iqbal, M. A., Abdul, W., & Mushtaq, M. (2019). Cervical cancer prediction through different screening methods using data mining. *IJACSA) International Journal of Advanced Computer Science and Applications*, *10*(2), 388-396.

[25]. William, W., Ware, A., Basaza-Ejiri, A. H., & Obungoloch, J. (2019). A pap-smear analysis tool (PAT) for detection of cervical cancer from pap-smear images. *Biomedical Engineering Online*, *18*(1), 1-22.