Annals of R.S.C.B., ISSN:1583-6258, Vol. 25, Issue 5, 2021, Pages. 1380 - 1393
Received 15 April 2021; Accepted 05 May 2021.

# Supervised Algorithms of Machine Learning in the Prediction of Cervical Cancer: A Comparative Analysis

**CH.Bhavani [1], Dr.A.Govardhan [2]**

[1]Assistant Professor, CVR College of Engineering, Hyderabad, India. Email Id: vbhavani118@gmail.com

[2] Rector, Jawaharlal Nehru Technological University, Hyderabad, Telangana -500085, India
Email Id: govardhan_cse@jntuh.ac.in

## Abstract:

As reported by World Health Organization (WHO), among women Cancer of Cervix is ranked as fourth predominant type of cancer producing 7.9% of female cancers. The symptom-less nature of the disease is the major challenge for the researchers and lack of early effective diagnosis is the basic reason for the of the disease. Machine learning a subgroup of artificial intelligence has been helpful in many medical diagnoses that let on the machine to gather intelligence from existing data and practice them to draw insights from the large amounts of data. The survey paper incorporates various techniques of machine learning that are suited for predicting cervical cancer.

## Introduction

Cervical cancer, a perilous and 4th highest predictable cancer amidst the female all over the globe. The cervix is a female reproductive system which is in cylindrical pear-shaped hallow organ located at the climax of the uterus, the place where embryo expands. It led way from the uterus to birth canal termed vagina. The bottom part of the cervix lies inside the vagina labelled ectocervix and the overhead part of the cervix labelled endocervix recline above the vagina. The cervical cancer cells usually originate from the point where the ectocervix and endocervix join [1, 2].

As stated by WHO, Cervical cancer ranks 4th among all carcinoma types in women throughout world and is ranked second frequent cancer among women in India. In 2018, world widely approximately 570000 women were undergone cervical cancer diagnosis and about 311000 women died of disease [3]. In India, the estimated new cases of cervical cancer are approximately 96,922 and deaths are 60,078 [4]. Approximately 1 lakh total cases were estimated in 2016, 1.04 lakh total cases were expected in 2021 [5]. Having no proper awareness, lack of expert physicians and equipment's, lack of early detection were considered as the main reasons for this type of cancer in the low-incomed countries [6].

As cancer of cervix is asymptomatic in nature, predicting it in the early stage has become a major challenge for the researchers. As it does not exhibit any symptoms in the early stage and symptoms appear only in the later stage which is considered as advanced cervical cancer. The general symptoms comprise abnormal menstruation, irregular menstruation, heavy menstruation, vaginal bleeding, increased vaginal discharge, unexplained persistent pelvic pain or spotting, fatigue, leg-pains, loss of weight, loss of appetite, bone fractures and back pain [7, 27].

Cervical cancer arises from a type of virus termed as “Human Papillomavirus” (HPV) that are commonly found worldwide. There are about 100 varieties of HPV types, 14 of these are treated as cancer causing also called as high-risk types. HPV is disseminated through sexual contact. Majority of the adult women will be affected with this kind of virus abruptly after the beginning of sexual activity. Apart from this virus cigarette smoking, contraceptive usage, multiple pregnancies are also considered as the origin of cervical cancer [8, 9].

The threat of cervical cancer increases, if a female blighted with HPV has a habit of tobacco/cigarette smoking. The women using oral contraceptives or pills that control the birth have 3 times higher threat of attacking cervical cancer than the women who does not use contraceptives. Similarly, the women with multiple pregnancies may have higher chances of cervix cancer than the HPV-infected women without pregnancy [10].

## Staging of Cervical Cancer: FIGO staging of cervical cancer

### Stage I:

At this level, the cancer has-been began and identified only in the cervix.

Considering the tumor size and extensive point of tumor invasion, stage 1 is further split-up into two stages IA and IB.

- Considering the extensive point of invasion, stage IA has been further diverged into IA1 and IA2.
- In stage IA1, a minute proportion of cancer disease can be identified in the tissues of cervix when dissected in the microscope. The deepest invasion of tumor will not be larger than 3mm in this stage.
  In stage IA2, the extensive point of tumor invasion will be larger than 3mm but not greater than 5mm, which is going to be examined with a microscope in the cervix tissues.

- Considering the extensive point of invasion, stage IB has been further split-up into IB1, IB2 and IB3.
  - In phase IB1, the tumor dimension will be smaller than or equal to 2 cm and extensive point of invasion of tumor is larger than 5mm diameter.
  - In phase IB2, the dimension of the lump will be greater than 2cm and will be less than 4cm.
  - In stage IB3, the tumor size will be more than 4cm.

### Stage II:

In level II, the cancer has expanded to the tissues encircling the uterus and two-thirds upper part of the vagina.

With reference to the distance of infection, stage II had been further diverged as two stages – IIA, IIB.

- In stage IIA, the cancer disease has infected towards the overlying part of vagina and to the tissues encircling the embryo/uterus. By taking the dimension of the tumor, stage IIA had been split into two stages.
  - In IIA1 stage, the dimension of the tumor will be smaller than 4cm
  - In IIA2 stage, the dimension of the tumor will be larger than 4cm.

- In stage IIB, the cancer has infected to the tissues encircling the uterus but not to the parts which are above the vagina.

## Stage III:

In this level, the infection of the cancer has reached the underneath part of the vagina either-or near the wall of the pelvic either-or has affected kidney either-or involvement of lymph nodes were exhibited. Calculating the distance of infection, the stage III had been diverted into IIIA, IIIB, IIIC.

- If the cancer disease has infected to the bottom part of vagina, not towards the pelvic walls then it was considered to be in stage IIIA.
- If the cancer had infected towards the pelvic wall either-or the tumor had grown massive in size which will be able to block ureters or caused the kidneys to enlarge or stop functioning.
- IIIC, is sub categorized into IIIC1 and IIIC2, depending on extent of lymph node infection.
    - In IIIC1 phase, cancer disease has infected in pelvis lymph nodes.
    - In level IIIC2, cancer disease has infected to abdomen lymph nodes near aorta.

## Stage IV:

In level IV, the cancer disease has infected exceeding the pelvis and other parts of body. Based on cancer infection phase IV is partitioned as IVA, IVB.

- In level IVA, cancer disease has infected to nearby organs like bladder or rectum.
- In level IVB, cancer disease has infected various parts of body such as lungs, liver, distant lymph nodes and bones [11].

Screening is the primary objective for cervical cancer. A perfect screening test is a particular one with minimal invading, effortless for achieving, justifiable to subject, economic and fruitful in diagnosing the disease the disease at early stage when treatment is effortless for ailment. Four screening methods are at hand for cervical cancer, which encompass cytology also termed as Pap smear test (Papanicolaou) is an uncomplicated, painless, quick screening stratagem for testing cervical cancer, this procedure incorporates, collecting cells arising out of women’s cervix and are unroll over the electron-microscope and are examined in order to find the abnormal cells. The procedure is sightly uncomfortable but does not has a long-term pain, the procedure has some drawbacks like having no proper patient obedience, having no proper care after the test, inadequate clinical components and time taking process. [12, 13]. Biopsy is a medical procedure, where a piece of live tissue of the cervix is cut and sent to pathologist for diagnosis [14]. The Hinselmann test will be done by applying Iodine solution for inspecting the cervix visually. Lugol’s iodine solution will be applied to inspect the cervix succeeding smudging Lugol’s iodine perception rate of doubtful area over the cervix, this is also termed as Schiller test [15].

A tremendous escalation in the populace of the world created a pressure on healthcare sector for providing quality treatment and healthcare services. Machine Learning, subgroup of Artificial Intelligence takes a crucial part in numerous health-correlated issues. It is crystal clear that, implementing machine learning approaches increases accuracy very successfully in cancer prediction or susceptibility. In an article called “A.I. versus M.D.”, Sebastian Trump a computer scientist quoted “exactly as medicines build human muscles a thousand times stronger, machines will model the human brain a thousand times better powerful”.

## Literature Review

Literature review presents a brief summary of previous works done by many experimenters in the predicting the cervical cancer and also briefs the machine learning techniques that have employed in their works [16]. The review includes narration of various approaches and algorithms for filling the missing data, solving the data imbalancing problem, feature reduction techniques of supervised machine learning.

[17] the authors have used a dataset which has been collected from the hospital ‘Universitario de Caracas’ in Caracas has 32 features or variables, 4 dependent variables. The dataset has missing values considering it as imbalanced dataset; Oversampling has been applied for preprocessing the data. As cervical cancer is hard to detect due to lack of symptoms in early stage, the authors has applied standard **SVM** but because of low dimensionality they did not able to get the clear disjunction of hyperplane. As there are massive number of features, it led to overpriced computational cost and harms the results due to the existence of noise. The improved version of SVM method **SVM-RFE** (recursive feature extraction), in this method first they applied SVM to every feature and completed the procedure of training and the features with little relevance are eliminated due to which the computational costs has been reduced. Another improved version **SVM – PCA** an effective statistical multi-variate method has been applied. They claimed that SVM is capable of classifying the 30 features, whereas two versions of SVM, REF & PCA are capable of doing same with hardly 8 features.

[18], the authors have followed a three-step process. In the initial step, they have chosen a dataset with 32 features. As second step, data cleaning has been done using ignoring missing values and imputation using decision tree methods. The data imbalancing problem has been solved using SMOTE (Synthetic Minority Oversampling Technique), a new strategy put forward by Chawla et al. The third step was feature reduction and classifier building, a feature reduction technology Principal Component Analysis (PCA) having 11 principal integrant has been practiced to bring down the count of features and working hours. For modelling, four models implemented on target variables, popular ensemble method, **voting classifier** has been used which combines logistic regression, RF and DT. A validation method, 10-fold cross validation was applied for preventing the problem of overfitting. Authors claimed, the usage of voting classifier, PCA & SMOTE strategies helped them in raising the accuracy, ROC-AUC (Receivers operating characteristic – Area Under Curve) Curve and sensitivity.

[19], the researchers collected the data from Shohada Hospital which contained 145 records with 23 attributes, they analyzed various algorithms which includes Support Vector Machines (SVM), MLP, QUEST, RBF (Radial basis function network) and C&R Tree. The algorithms were evaluated based on accuracy, sensitivity, specificity and area under the curve (AUC). Investigations have concluded that the most appropriate predictors can be pin down by applying decision trees. Authors found QUEST and C&R Tree has outperformed. They identified that the cervical cancer can be shut out by boosting individual health and socio-cultural level of the diseased person.

In [20] image processing has been used for pap smear images and a screening system was developed which is a computer-assisted system, since manual screening of images obtained from pap smear test in the microscope is suffering from poor copiability. They followed six step process, image acquisition followed by enhancement of image, segmenting the cell to detect shape of nuclei and separating overlapping cytoplasm, extracting and selecting the features has been done by Random Forest, finally for classification Bagging Ensemble classifier has been adopted which integrated the consequences of linear discriminant (LD), boosted Trees, SVM, KNN, Bagged Trees on SIPaKMed & Herlev datasets. They claimed that they got better accuracy in the binary and multi class problems.

Kayalvizhi and Kanimozhi[21], has conducted research on various supervised and unsupervised algorithms for predicting cervical cancer. The methods that have been analyzed are Linear Regression (LR), Deep Neural Network (DNN), Decision Tree, SVM, Random Forest (RF). SMOTE has been employed for balancing the dataset. The authors focused on Recall to shrink the type-I error as type-II error is costlier. They concluded RF and DNN are the overall best model, if all 4 dependent variables are combined together.

Fazal et.al [22] used outlier detection algorithms DBSCAN and iforest to eliminate the differently behaving data objects from the dataset and raised the number of samples in the dataset for solving data imbalance using SMOTE and SMOTETomek techniques. The chi-square has been implemented for selecting the features and selected 10 features for further analysis. They concluded the model developed using Random Forest+SMOTETomek+iforest has better accuracy. They also developed a mobile application for collecting the risk or probable factor responses from users and gives the result about the disease.

The [23] authors have used the classifiers Naive Bayes (NB), KNN, MLP Neural Network, Sequential Minimal Optimization (SMO), C4.5 Decision Tree, Random Forest (RF), Simple Linear Regression. Authors asserted RF was best and NB was worst in performance.

The [24] study accomplished in this paper was based on an ensemble approach where voting strategy was used and result of the model has been illuminated by conventional classifiers Linear Regression, KNN, Decision Tree, MLP, SVM. A genomic sequencing dataset was exercised and focused on recall, precision and F-score. Data was accumulated from a questionnaire of 472 patients at a Chinese hospital.

[25] Authors have set their site on building a classification model using Random Forest algorithm and SMOTE for data imbalance's, PCA & RFE for feature reduction. They constructed an RSOnto ontology for visualizing the performance progress in classification at a better rate.

[26], data from UCI having 4 dependent variables have been combined to form a single dependent variable which contained summation values of four target variables. The non-zero values were replaced with 1, since models have given poor performance with original summated values. Parameter's tunning was done for optimal evaluation. Hill Climbing algorithm has been carried out for feature selection.

Jaswinder Singh, Sandeep Sharma. Et. Al.,[27], put forward a model by collecting the data from wireless sensors to the finger tips and feet of the patient and concerted it to digital form using Ardino chip for predicting the stage or level of Cervical Cancer. Six ML algorithms Naïve Bayes, Function-based-logistic SMO, Lazy-based-LWL, meta-based-iterative-classifier-optimizer, Rule based Decision Tree & Tree based Decision Stump were implemented to train the data and evaluated true-positive's, false-positive's, Matthews-correlation-coefficient (MCC), f-measure, and precision. The observation was Tree-based-Decision Stump was better among all.

The authors of [28] used dataset that includes a massive amount of missing data which are also misaligned. To tackle these issues NOCB, LOCF, filling using the median and filling using mode were tailored. Correlation has been used for selecting the features and to avoid the bias in data, the data was divided into doubtless and doubtful biopsy and then break the isolated sets randomly as 2 sets called training set and testing set. Decision Tree, SVM, Naïve Bayes, RF, Logistic Regression (LR) and Neural Network were applied contrasting the recall, F1-Score, accuracy and precision and identified SVM, LR with NOCB preprocessing is the superior.

The authors [29] used SMOTE for data imbalance problem and carried out tests for four screening methods Schiller, Cytology, Biopsy and Hinselmann individually using Boosted Decision Tree, Decision Jungle, Decision Forest. Boosted Decision Tree a variant of Decision Tree has obtained higher performance, AUROC than the sensitivity, precision and specificity. Hinselmann screening method has got better accuracy.

[30] have used C5.0, Random Forest Tree, SVM, RPART and KNN. Filter and wrapper methods were implemented for selecting the features; R-language was implemented for data balancing. The dataset was reduced to 26 independent variables or features and combined with a new feature which has obtained by combining all four dependent variables. They affirmed C5.0, Random Forest were shown better accuracy.

[31] Presented various datasets that are publicly available. SVM, Hierarchal Clustering, C5.0, Genetic Algorithms, PNN, CNN, Decision Tree, GLCM were employed for building the model with accuracy as predictor and concluded saying convNet-T a variant of CNN was superior in classifying the cancer image cells and least square version of SVM and SoftMax type of regression for classification of records was used.

In the study [32], a dataset with 858 records consisting of 32 features. As a result of preprocessing two features were pin pointed as containing 92% missing values were excluded. Over & under Sampling methods, wrapper methods like Selection of Features hinged on Sequential Forward Selection (SFS) & Sequential backward Selection (SBS) were employed to fill remaining missing data. Decision Tree, Logistic Regression (LR), SVM and KNN classifiers are implemented for building the model, among these Decision tree is identified as exhibiting higher accuracy by choosing common features in the middle Decision Tree and KNN.

The authors [33] used a data from UCI repository, unbalancing problem was solved using SMOTE. For down scaling the number of features, RFE and PCA were implemented. They proved that the model performance has been enhanced by using RFE-SMOTE with 10-fold cross validation with evaluation metric sensitivity, accuracy, specificity, PPA, NPA for all four dependent variables Cytology, Biopsy, Hinselmann and Schiller before SMOTE and after SMOTE results were analyzed.

The study [34] employed borderline SMOTE for data imbalance problem and found effective. SVM, Random Forest and XGBoost were used and claimed XGBoost, Random Forest was performing well than the SVM with regard to sensitivity and specificity while ensuring high accuracy. They identified top 5 features, count of partners of sex, count of pregnancies, age, age at which first sexual intercourse happened and usage of contraceptives as causes of the disease.

[35] used pap smear image dataset with 20 features and a clustering algorithm Fuzzy C-means for preprocessing the dataset in MATLAB software using KNN and ANN. The classifier accuracy has taken as the criteria of evaluation for assessing the used algorithms and concluded the future scope of work as identifying the stage or level of the disease.

The authors [36] employed KNN, MLP, Bayes Net and focused on false positives and false negatives and given classified instances rates as 1.37%, 1.71% and 2.05% respectively for the algorithms employed. In authors opinion KNN was better with 86 neighbors but authors have not discussed about data balancing and feature selection techniques.

[37] Collected the data from NCBI which encompasses of 500 records with 61 attributes. The authors have analyzed the outcomes using RF, CART and RF Tree with K-Means an hybrid algorithm using

MATLAB software for coding and claimed RF Tree with K-Means has been out performed by exhibiting an accuracy of 96.77%.

The authors of [38] demonstrated various stages or levels of cervical cancer. The implementation was done using Naïve Bayes, SMO, LWL, Iterative classifier, Decision Table, Decision Stump using WEKA tool and concluded Iterative Classifier Optimizer and Decision Stump were outperformed, Decision Table was the worst performer.

**Table 1: Comparison of various approaches surveyed**

| Reference | Dataset | No. Of Attributes | Approaches and Methods used | Pre-Processing Method | Evaluation Metrics | Advantage | Limitation |
|---|---|---|---|---|---|---|---|
| [19] | Shohada Hospital | 23 | SVM<br>QUEST<br>C&R TREE<br>MLP-ANN<br>RBF-ANN | Filling missing values using mean, median and mode | Sensitivity, Accuracy, specificity and AUC | Identified new risk factors | Small dataset was used to build the model, basic data imputations were used. |
| [22] | UCI Repository | 36 | DBSCAN + SMOTETomek + RF<br>DBSCAN + SMOTE + RF | Feature - Extraction<br>Outlier Detection<br>Data Balancing | Precision<br>Recall<br>True Negative Rate<br>F1 Score<br>Accuracy | They have come with a Mobile application | Application was employed based on only one dataset. |
| [23] | UCI Repository-reduced | 8 | Naïve Bayes<br>C4.5<br>KNN<br>SMO<br>Random Forest<br>MLP<br>SLR | SMOTE<br>PCA | Accuracy<br>Precision<br>Recall | Identified a reduced UCI repository dataset. | Simple data imputation methods were used. |
| [24] | Chinese Hospital<br>UCI | 50<br>32 | Logistic Regression<br>Decision Tree<br>SVM<br>MLP<br>KNN | Firefly – algorithm for selecting the Features | Precision<br>Recall<br>F1 Score<br>Accuracy | For improving the prediction, a Data correction mechanism was proposed. | Limited by Experimental support, colposcopy images not included. |
| [25] | UCI Repository | 32 | RF<br>RF – PCA<br>RF - RFE | SMOTE<br>PCA<br>RFE | Accuracy<br>Specificity<br>Sensitivity<br>PPA<br>NPA | Represented ontological representation | SMOTE was giving good accuracy for higher dimensionality since it results in overlapping as adjacent nodes were |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | not considered. |
| [26] | UCI Repository | 33(includes combined target variable) | KNN<br>Decision Tree<br>Random Forest | Parameter Tunning<br>Hill climbing algorithm – feature selection | Confusion Matrix<br>Accuracy<br>Classification Error Rate<br>Precision<br>Recall<br>F1 Score | | Data balancing techniques were not discussed. |
| [30] | UCI Repository | 26 | C5.0<br>Random Forest<br>KNN<br>SVM<br>Recursive partitioning and regression (rpart) | RFE, Boruta – An algorithm for selecting the features in R language, Simulated annealing-feature selection method | Accuracy<br>AUC | For reducing the number of features Optimized feature selection techniques were used. | Related information of dataset (like factorized or not factorized) was not provided, replacement of variables was done on the principal of assumptions. |
| [35] | Herlev dataset | | KNN<br>ANN | For classifying the cells Coupled fuzzy based segmentation techniques with KNN was used. | Accuracy | An automated, comprehensive machine learning approaches has improved identification of cytoplasm of the cervix cell. | ANN was not given proper accuracy and might be improved. |

## Research Methodology:

We have conducted analysis of a dataset which has been collected from UCI repository with 858 instances and 36 features. The features have both integral and Boolean that is categorical kind of data and also it was consisting many missing values. We have identified that, two of the attributes were having more than 91% of missing values. So, as a first step, we have dropped those from the dataset. As a second step, preprocessing the data, we have filled the missing values by implementing simple imputer technique; strategy mean was used for integer attributes and strategy most frequent for categorical attributes. The dataset was consisting of four target variables; among them we have taken Biopsy as target variable. The dataset was highly imbalance, so we have implemented the basic supervised algorithms of machine learning Logistic Regression (LR), Decision Tree (DT), Naïve Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machines (SVM) before solving the imbalance problem and after solving the imbalance problem using a combination of over sampling and under sample technique SMOTETomek.

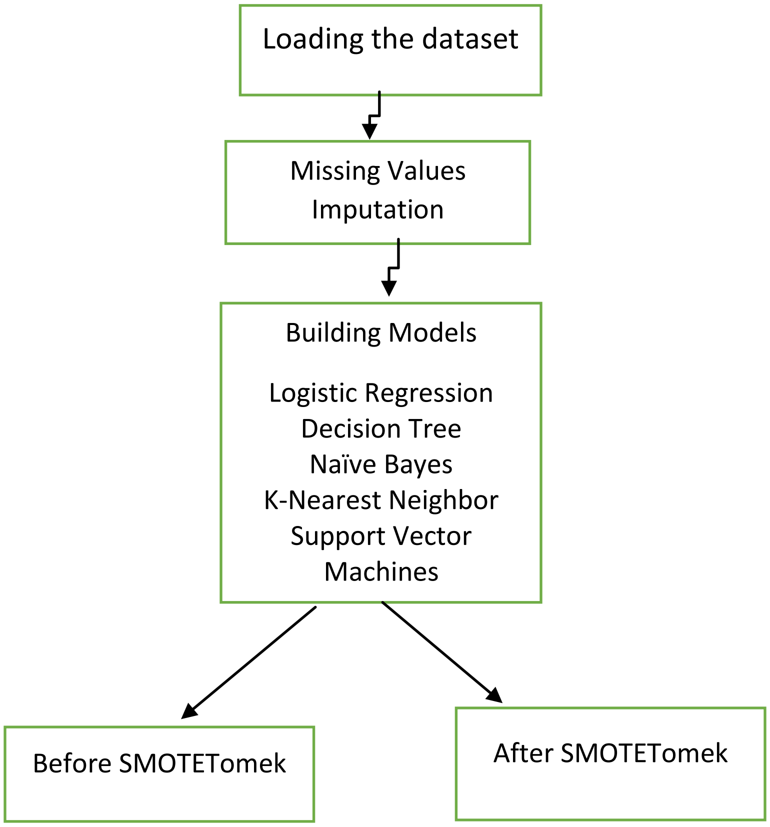


**Fig 1: Process flow of the work**

The following models were used as part of the prediction:

1) Logistic Regression

Logistic Regression is used for both prediction and classification, which is statistical model used when dependent variable is a binary one. It used when the data points are spread only between 0 and 1, which will be using the sigmoid function or logit function. It is used to predict the relationship between dependent variable and one or more independent variables. It is a probabilistic based supervised algorithm [41].

2) Decision Tree

Decision Tree is a tree like structure, where the attributes will form the nodes of the tree and the class labels are leaf nodes. The split attribute will be identified by using the Gain, Gini Index or Information Gain measures. So, when we get an instance to predict the class, we will traverse through the tree until we reach one of the leave nodes [42].

3) Naïve Bayes

Naïve Bayes will classify all the entities of a dataset by using bayes theorem, a probabilistic approach. Most of the times, this algorithm will be used in healthcare as it is suited for high dimensionality data. As healthcare data have lots of features and the usage of this algorithm will be an added advantage [43].

4) K-Nearest Neighbor

KNN is a non-parametric, simple supervised algorithm which is also considered as lazy-learner algorithm as it does not learn from the dataset immediately but learns as when it requires. It used for both classification and regression.

5) Support Vector Machines

SVM is a most popular supervised machine learning algorithm which is used for classifying the data by using a kernel method for transforming the data from low dimensionality to higher dimensionality. We have used linear SVM among the various variants of SVM which is extremely fast and efficient and also produce better accuracy for the prediction [44].

## Results and Discussion

Our dataset was divided into training and testing sets at a ratio of 7:3 that is 70 % of data has been considered as training for building the model and 30% of data was used for testing the built model. Accuracy, Recall, Precision, f1-score. ROC_AUC score and confusion matrix were used for comparing the performance of the models before and after applying the SMOTETomek sampling method.

Accuracy is a good performance measure which will summarize the prediction results.

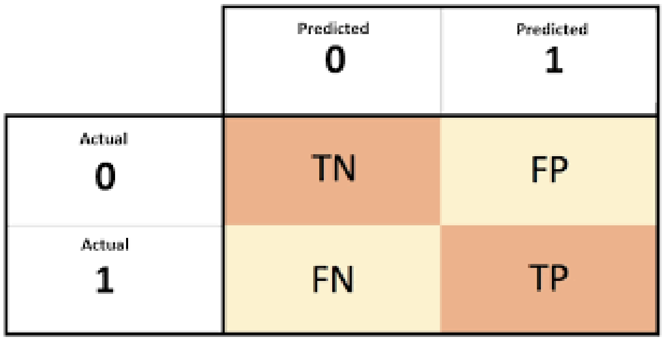
Recall is also called as sensitivity or True Positive Rate.

Precision is also called as Positive Predictive Rate.

F1-score is a tradeoff between precision and recall.

ROC_AUC score will always varies between 0 and 1, it discriminates the threshold created between true positive rate and false positive rate.

## Confusion Matrix:



| | Predicted<br>0 | Predicted<br>1 |
|---|---|---|
| Actual<br>0 | TN | FP |
| Actual<br>1 | FN | TP |

**Table 2: The performance of the models before solving the unbalanced data problem was**

| **Name of the Classifier** | Accuracy | **recall** | Precision | F1-score | Roc_auc score | Confusion matrix |
|---|---|---|---|---|---|---|
| Logistic | 0.957364 | 0.684211 | 0.722222 | 0.702703 | 0.831645 | [[234, 5], |

| | | | | | | |
|---|---|---|---|---|---|---|
| Regression | | | | | | [6, 13]] |
| Decision Tree | 0.934109 | 0.684211 | 0.541667 | 0.604651 | 0.819093 | [[228, 11],<br>[6, 13]] |
| Gaussian Naïve Bayes | 0.112403 | 1.000000 | 0.076613 | 0.142322 | 0.520921 | [[10, 229],<br>[0, 19]] |
| KNN | 0.937984 | 0.263158 | 0.714286 | 0.384615 | 0.627395 | [[ 237, 2],<br>[14, 5]] |
| SVM | 0.965116 | 0.947368 | 0.692308 | 0.800000 | 0.956948 | [[231, 8],<br>[1, 18]] |

**Table 3: The performance of the models after solving the unbalanced data problem by using SMOTETomek**

| **Name of the Classifier** | Accuracy | **recall** | Precision | F1-score | Roc_auc score | Confusion matrix |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.965116 | 0.947368 | 0.692308 | 0.80000 | 0.956948 | [[231, 8],<br>[1, 18]] |
| Decision Tree | 0.953488 | 0.842105 | 0.640000 | 0.727273 | 0.902224 | [[230, 9],<br>[3, 16]] |
| Gaussian Naïve Bayes | 0.112403 | 1.000000 | 0.076613 | 0.142322 | 0.520921 | [[10, 229],<br>[0, 19]] |
| KNN | 0.937984 | 0.263158 | 0.714286 | 0.384615 | 0.627395 | [[ 237, 2],<br>[14, 5]] |
| SVM | 0.969612 | 0.947368 | 0.60000 | 0.734694 | 0.948580 | [[227, 12],<br>[1, 18]] |

From the results, we have concluded that before applying the SMOTETomek, the results were seeming good but we may not rely on these since, the model learns and mostly predicts that patient is non-cancerous as there are huge numbers in dataset. So we have solved the unbalanced problem applied the same algorithms on them and found that SVM, Logistic Regression and Decision Tree were performing better than the other models. The confusion matrix is also revealing the same analysis about the classifiers.

Medical data is very crucial in the predicting and diagnosis of disease by applying machine learning techniques. The pertinency of various algorithms of machine learning in predicting the abnormalities in the data requires studying the handy data and identifying the outcomes grounded on the understanding that has been acquired. The following conclusion has been drawn by considering the existing work that has been performed in the area of prediction of cervical cancer by utilizing the out coming methods of machine learning.

Support Vector Machines are extensively employed by researchers to predict cervical cancer, as the algorithm is very robust. In such models, feeding data can safeguard from immaterial and potentially

Received 15 April 2021; Accepted 05 May 2021.

ambiguous information [39]. Random Forest (RF) plays a vital role in disease identification, as it can deal with overfitting and enhance the accuracy [40]. The dataset is suffering from unbalancing problem SMOTE, SMOTETEK were employed to resolve the problem. For feature selection various algorithms of machine learning were tailored and required features for predicting the cervical cancer were extracted to strength the accuracy of prediction.

## Conclusion:

ML a subgroup of Artificial Intelligence is emerging as unique prominent field in medical diagnosis. Cervical cancer which is a type of gynecological cancer, can be medicated if identified in the earlier stages. In this paper a brief list of studies and algorithms used by various researchers has been used for predicting the cervical cancer. From the studies, we have identified that, there exists a requirement of a hybrid ensembled approaches in filling the missing values, selecting the features and also in the model building and also enhancing the SVM and a variant of decision Tree which is Random Forest Tree as a future work and also to employ hyper parameter tunning techniques to enhance the performance metrics of the models.

## References:

[1] Organization CDC, Centers for Disease Control and Prevention retrieved from https://www.cdc.gov/cancer/cervical/basic_info/index.htm#:~:text=When%20cancer%20starts%20in%20the,at%20risk%20for%20cervical%20cancer. Last accessed on 22/02/2021

[2] W. H. O. R. Health, W. H. O. C. Diseases, and H. Promotion, Comprehensive Cervical Cancer Control: A Guide to Essential Practice, World Health Organization, Geneva, Switzerland, 2006.

[3] Organization W H 2021 retrieved from https://www.who.int/health-topics/cervical-cancer#tab=tab_1 last accessed 25-03-2021

[4] Balasubramaniam G, Gaidhani RH, Khan A, Saoba S, Mahantshetty U, Maheshwari A. Survival rate of cervical cancer from a study conducted in India. Indian J Med Sci, doi: 10.25259/IJMS_140_2020

[5] Organization NCBI 2021 retrieved from http://nciindia.aiims.edu/en/cancer-statistics# Last accessed 25-03-2021

[6] Mishra, G. A., Pimple, S. A., & Shastri, S. S. (2011). An overview of prevention and early detection of cervical cancers. Indian journal of medical and paediatric oncology: official journal of Indian Society of Medical & Paediatric Oncology, 32(3), 125–132. https://doi.org/10.4103/0971-5851.92808

[7] Organization ASCO retrieved from https://www.cancer.net/cancer-types/cervical-cancer/symptoms-and-signs last accessed 26-03-2021

[8] Organization W H 2021 retrieved from https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer

[9] Bosch FX, de Sanjosé S. Chapter 1: Human papillomavirus and cervical cancer--burden and assessment of causality. J Natl Cancer Inst Monogr. 2003;(31):3-13. doi: 10.1093/oxfordjournals.jncimonographs.a003479. PMID: 12807939.

[10] Organization ASCO retrieved from https://www.cancer.net/cancer-types/cervical-cancer/risk-factors last modified Nov 2020

[11] Organization ASCO retrieved from https://www.cancer.net/cancer-types/cervical-cancer/stages

[12] R. A. Kerkar, "Screening for cervical cancer: an overview." G. Guvenc, A. Akyuz, and C. H. Açikel

[13] "Health belief model scale for cervical cancer and Pap smear test: psychometric testing," Journal of advanced nursing, vol. 67, pp. 428-437, 2011.

[14] M. T. Galgano, P. E. Castle, K. A. Atkins, W. K. Brix, S. R. Nassau, and M. H. Stoler, "Using biomarkers as objective standards in the diagnosis of cervical biopsies," The American journal of surgical pathology, vol. 34, p. 1077, 2010

[15] H. Ramaraju, Y. Nagaveni, and A. Khazi, "Use of Schiller's test versus Pap smear to increase detection rate of cervical dysplasias," International Journal of Reproduction, Contraception, Obstetrics and Gynecology, vol. 5, pp. 1446-1450, 2017

[16] Akshitha Shetty , Vrushika Shah "Survey of cervical cancer Prediction using Machine Learning: A comparative approach" IEEE 9th ICCCNT 2018 July 10-12, 2018, IISC, Bengaluru, India.

[17] W. Wu and H. Zhou, "Data-Driven Diagnosis of Cervical Cancer with Support Vector Machine-Based Approaches," in IEEE Access, vol. 5, pp. 25189-25195, 2017.

[18] Riham Alsmariy, Graham Healy, Hoda Abdelhafez "Predicting Cervical Cancer using Machine Learning Methods" IJACSA thesia.org 2020.

[19] F A, C S, L A. Supervised Algorithms of Machine Learning for the Prediction of Cervical Cancer. J Biomed Phys Eng. 2020 Aug 1;10(4):513-522. doi: 10.31661/jbpe.v0i0.1912-1027. PMID: 32802799; PMCID: PMC7416093.

[20] Win KP, Kitjaidure Y, Hamamoto K, Myo Aung T. Computer-Assisted Screening for Cervical Cancer Using Digital Image Processing of Pap Smear Images. *Applied Sciences*. 2020; 10(5):1800. https://doi.org/10.3390/app10051800

[21] Kayalvizhi. K. R | N Kanimozhi "Prediction of Cervical Cancer using Machine Learning and Deep Learning Algorithms" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-4 | Issue-6, October 2020, pp.426-430, URL: www.ijtsrd.com/papers/ijtsrd33378.pdf

[22] Ijaz, Muhammad Fazal et al. "Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods." *Sensors (Basel, Switzerland)* vol. 20,10 2809. 15 May. 2020, doi:10.3390/s20102809

[23] Razali, Nazim & Mostafa, Salama & Mustapha, Aida & Abd Wahab, Mohd Helmy & Ibrahim, Nurul. (2020). Risk Factors of Cervical Cancer using Classification in Data Mining. Journal of Physics: Conference Series. 1529. 022102. 10.1088/1742-6596/1529/2/022102.

[24] Lu, Jiayi & Song, Enmin & Ghoneim, Ahmed & Alrashoud, Mubarak. (2020). Machine learning for assisting cervical cancer diagnosis: An ensemble approach. Future Generation Computer Systems. 106. 10.1016/j.future.2019.12.033.

[25] Geetha, R. & Sivasubramanian, Sankaranarayanan & Kaliappan.,M.E.,Ph.D, Dr.M & Shanmuganthan, Vimal & Suresh, Annamalai. (2019). Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier. Journal of Medical Systems. 43. 10.1007/s10916-019-1402-6.

[26] Parikh, Dhwaani & Menon, Vineet. (2019). Machine Learning Applied to Cervical Cancer Data. International Journal of Mathematical Sciences and Computing. 5. 53-64. 10.5815/ijmsc.2019.01.05.

[27] Jaswinder Singh & Sandeep Shrama. (2019). Prediction of Cervical Cancer Using Machine Learning Techniques. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 11 (2019) pp. 2570-2577.

[28] Abdullah, F. B. Ashraf and N. S. Momo, "Comparative analysis on Prediction Models with various Data Preprocessings in the Prognosis of Cervical Cancer," 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 2019, pp. 1-6, doi: 10.1109/ICCCNT45670.2019.8944850.

[29] Alam, T.M. & Khan, M.M.A. & Iqbal, Muahammad & Wahab, A.. (2019). Cervical cancer prediction through different screening methods using data mining. International Journal of Advanced Computer Science and Applications. 10. 388-396.

[30] Nithya, B. and V. Ilango. "Evaluation of machine learning based optimized feature selection approaches and classification methods for cervical cancer prediction." *SN Applied Sciences* 1 (2019): 1-16.

[31] Singh, Yasha & Srivastava, Dhruv & Anand, Chandranand & Singh, Dr. (2018). Algorithms for screening of Cervical Cancer: A chronological review.

[32] Alwesabi, Yaseen & Choudhury, Avishek & Won, Daehan. (2018). Classification of Cervical Cancer Dataset.

[33] fayz, Sherif & rizka, Mohamed & Maghraby, Fahima. (2018). Cervical Cancer Diagnosis Using Random Forest Classifier with SMOTE and Feature Reduction Techniques. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2874063.

[34] Deng, Xiaoyu & Luo, Yan & Wang, Cong. (2018). Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods. 631-635. 10.1109/CCIS.2018.8691126.

[35] Priyanka K Malli, Dr. Suvarna Nandyal , " Machine learning Technique for detection of Cervical Cancer using k-NN and Artificial Neural Network" , International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) , Volume 6, Issue 4, July- August 2017 , pp. 145-149 , ISSN 2278-6856.

[36] Ünlerşen, Muhammed & Sabanci, Kadir & Özcan, Muciz. (2017). Determining Cervical Cancer Possibility by Using Machine Learning Methods. International Journal of Latest Research in Engineering and Technology. 3. 65-71.

[37] Raghavendran, R.Vidya & Nasira, G.M.. (2016). Prediction of Cervical Cancer using Hybrid Induction Technique: A Solution for Human Hereditary Disease Patterns. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i30/82085.

[38] Sharma, Sunny and S. Gupta. "Decision Tree approach in Machine Learning for Prediction of Cervical Cancer Stages using WEKA." (2016).

[39] Wang, Hui & Huang, Gang. (2011). Application of support vector machine in cancer diagnosis. Medical oncology (Northwood, London, England). 28 Suppl 1. S613-8. 10.1007/s12032-010-9663-4.

[40] Md. Zahangir Alam, M. Saifur Rahman, M. Sohel Rahman,A Random Forest based predictor for medical data classification using feature ranking, Informatics in Medicine Unlocked,Volume15,2019,100180,ISSN2352-9148,https://doi.org/10.1016/j.imu.2019.100180.

[41] Alzen, J.L., Langdon, L.S. & Otero, V.K. A logistic regression investigation of the relationship between the Learning Assistant model and failure rates in introductory STEM courses. *IJ STEM Ed* **5,** 56 (2018). https://doi.org/10.1186/s40594-018-0152-1

[42] Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang,Prediction performance of improved decision tree-based algorithms: a review,Procedia Manufacturing,Volume 35,2019,Pages 698-703,ISSN 2351-9789,https://doi.org/10.1016/j.promfg.2019.06.011.

[43] (https://www.sciencedirect.com/science/article/pii/S235197891930736X)

[44] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities," IEEE Access, vol. 5, pp. 8869–8879, 2017.

[45] Abdullah and M. S. Hasan, "An application of pre-trained CNN for image classification," 20th Int. Conf. Comput. Inf. Technol. ICCIT 2017, vol. 2018–January, pp. 1–6, 2018