IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

EARLIER PREDICTION AND DETECTION OF CERVICAL CANCER USING BAGGING CLASSIFIER WITH LOGISTIC REGRESSION ALGORITHMS

Dr D J Samatha Naidu[1] B. Haritha [2]
Principal ,Annamacharya PG College of Computer Studies,Rajampet[1]
MCA Student,Annamacharya PG College of Computer Studies,Rajampet [2]

Abstract : CERVICAL CANCER is one of such diseases which cannot be diagnosed earlier stages but this symptom nature of the diseases has become the challenging for the patient researchers and practitioners. A human body usually produces millions of cells to replace the inactive and dead cells in the body. If this body. If this production of cell produces and predictably, they will turn into tumors. This type of cancer which will be formed of the service of women in their women. In this earlier stages if we diagnostic then it is to cure than risk factories high. The existing works focusing on stack enable algorithms. The using that they can identify In the proposed system proposing bagging classifier technique which takes heterogeneous base learners along with logistic regression, random forest and meta learners SVM algorithms are used for predicting the cervical cancer from various risk factors, the applying proposed algorithms data balancing and acquire the future knowledge extension can be done to improve to overall accuracy of final model. Imputation of missing values as to be consider for biopic variables.Various data balancing techniques work be implemented type-2 errors has been included.The computational time can be reduced compared to any other algorithms such as LR, KNN, DTD, RF.

I INTRODUCTION

A human body usually produces millions of cells to replace the inactive and dead cells in the body. If this production of cells increases unpredictably, they will be turned into tumors. Not all tumors will be malignant and also goes off from the body. Some tumors which remain in the body turn to be malignant . Cervical cancer is a type of cancer which will be formed at cervix of women in their womb. This type of cancer is curable if identified at early in the stage. In course of time these cells travel through the tissues and effects the nearby organs in the body and converts into advanced stage of the cancer. The HPV is the main cause for this type of cancer.As stated by WHO, Cervical cancer ranks 4thamongall carcinoma types in women throughout world and is ranked second frequent cancer among women in India. According to the Indian statistics, 527,624 new cases are being appended every year, to this India contributes 122,844 cases every year. Globally1/3rd of cervical cancer deaths are being accounted by India with 1.6% aggregative risk of developing the cervical cancer and 1.0% aggregative deaths. Having no proper awareness, lack of expert physicians and equipment's, lack of early detection were considered as the main causes for this type of cancer especially in the low- and middle income countries. It will be very difficult for diagnosing diseases manually but now a days, machine learning approach is predominantly being used and is playing a key role in the diagnosis and prognosis of any kind of disease. Machine Learning allows the computers to acquire the knowledge from past or previous examples and helps us in predicting the present situation from complex data. Machine Learning takes the advantage of various probabilistic, statistical and optimization techniques for classification of cancerous and non-cancerous instances. In the present work, we have used Stacked ensemble technique with heterogeneous base learners Logistic Regression, SVM, KNN, Bagging Classifier, Random Forest and meta learner SVM.

II Objective of cervical cancer

A)Early Detection: Cervical cancer is highly treatable if detected early, making accurate prediction models crucial for timely intervention and improved patient outcomes.

B)Imbalanced Data: Imbalanced datasets pose a challenge in cervical cancer prediction, where the number of positive cases (cancer patients) is significantly lower than negative cases. SMOTE addresses this issue by generating synthetic samples of the minority class, thereby improving the training process and model performance.

C)Feature Selection: RFERF helps in identifying the most relevant features for prediction, reducing dimensional and enhancing model interpret ability. This is essential for understanding the underlying factors contributing to cervical cancer risk.

D)Predictive Performance: Optimized stacked ensemble techniques combine the strengths of multiple models, leading to improved predictive performance compared to individual classifiers. This enhances the reliability and robustness of the prediction system.

E)Personalized Treatment: Accurate prediction models enable personalized treatment strategies tailored to individual patient characteristics, improving treatment efficacy and patient satisfaction.

F)Research Advancement: The system contributes to the advancement of cervical cancer research by exploring novel machine learning techniques and integrating diverse data sources, paving the way for new insights and innovations in cancer prediction and management.

III RESEARCH METHODOLOGY

Women should be screened for cervical cancer every 5–10 years starting at age 30. Women living with HIV should be screened every 3 years starting at age 25. The global strategy encourages a minimum of two lifetime screens with a high-performance HPV test by age 35 and again by age 45 years. Precancers rarely cause symptoms, which is why regular cervical cancer screening is important, even if you have been vaccinated against HPV. Self-collection of a sample for HPV testing, which may be a preferred choice for women, has been shown to be as reliable as samples collected by healthcare providers.

A.Population and Sample

We have implemented Support Vector Machine, Bagging Decision Tree, Logistic Regression, K-Nearest Neighbor and Random Forest Classifier individually by hyper parameter tuning through which we can find the best parameters of these algorithms and identified the various scores and later we have built the stacked ensemble classifier which will be taking the heterogeneous base classifiers in the level I and forms a new data set at level II and we have observed the increase in the accuracy. In the experiments, we have performed hyper parameters tuning for the five different algorithms and employed stacked ensemble classifier and we have come to a conclusion that, stacking ensemble algorithm is performing better as it is trained on heterogeneous supervised algorithms

B. Data and Sources of Data

Cervical cancer program monitoring requires totals or counts (i.e. aggregate data) that summarize the delivery and outcomes of services provided to individual women. Summary data from each facility and laboratory are further aggregated to create datasets for district, regional, and national level monitoring. The ability to exchange information among the systems that collect and manage health data (HIS) is fundamental to quality data aggregation. However, in most low-resource settings, systems are fragmented and lack this necessary interoperability. Information exchange and data aggregation are further limited by the absence of national unique personal identifiers. Manual aggregation processes in paper-based information systems present an additional obstacle to ensuring the quality and timeliness of data for decision making.

C. Theoretical framework

In a study, Fernandesetal proposed a framework which helped in reducing the number of dimensions and classified the data using an artificial neural network (ANN). However, they have not explained how the null-values were dealt. As a conclusion, they made a differentiation between the baseline model and the proposed model, which contained a deep-fed neural network and acquired a finer accuracy than the baseline

D. Logistic Regression (LR)

Logistic Regression, a technique of machine learning has been borrowed from statistics for binary classification problem and has been growling used in healthcare research, particularly in the last two decades. The LR uses logistic function to squash the output of a linear equation between 0 and 1. It also establishes the association between one dependent binary variable and one or more independent variables. It uses the following sigmoid or logit function.

SigðtÞ ¼ 1 1 þ e x ð2Þ

LR is distinguished by not assuming a linear relationship between the dependent and independent variables but by displaying a relationship between the output aaaaaand predictive values. Furthermore, predictors are not required to have equal variance in each group or normal distribution.

E. Bagging classifier (BC)

Bagging classifier is an ensemble machine learning algorithm which helps in enhancing the performance of the model, avoids overfitting and also reduces the variance. It is usually applied with decision trees and leads to improvements of unstable procedures. It will train the base classifiers parallel with a training set which has been generated randomly with replacement. It reduces overfitting by using average of voting of outcomes from the base learners, which increases the bias and reduces the variance by maintaining the bias-variance trade-off.

F. K Nearest Neighbor classifier (KNN)

KNN is an easy-to-implement, simple machine learning algorithm based on supervised learning. It is a non-parametric and lazy-learner algorithm, since t does not learn immediately from the training data instead it learns at the time of classification and performs the action on dataset. It is used for both classification

G.Statistical tools and econometric models

Cervical cancer is the fourth most common cancer in women globally with around 660 00 new cases and around 350 000 deaths in 2022. The highest rates of cervical cancer incidence and mortality are in low- and middle-income countries.

H. Descriptive Statistics

cervical cancer risk factors data set used in the study was collected at "Hospital Universitario DE-Caracas" in Caracas, Venezuela and is available on the UCI Machine Learning repository . It consists of 858 records, with some missing values, as several patients did not answer some of the questions due to privacy concerns. the data set contains 32 risk factors and 4 targets, i.e., the diagnosis tests used for cervical cancer. It contains different categories of feature set such as habits, demographic information, history, and Genomics medical records. Features such as age, Dx: Cancer, Dx: CIN, Dx: HPV, and Dx features contains no missing values. Dx: CIN is a change in the walls of cervix and is commonly due to HPV infection; sometimes, it may lead to cancer if it is not treated properly.

I. Ensemble Learning in Healthcare

Previous studies have explored the application of ensemble learning techniques in various healthcare domains, including cancer prediction and diagnosis. Ensemble methods such as bagging, boosting, and stacking have been employed to improve the predictive performance of models in medical datasets.

2. Cervical Cancer Prediction Models

Existing research has developed prediction models for cervical cancer using diverse machine learning algorithms.Some studies have focused on feature selection methods to identify relevant biomarkers or risk factors associated with cervical cancer.

3. Imbalanced Data Handling Techniques

Imbalanced datasets are common in medical data, including cervical cancer datasets where positive cases (cancer instances) are usually a minority.Techniques like SMOTE have been used to address class imbalance issues by generating synthetic samples of the minority class.

4. Feature Selection Methods

Feature selection is crucial for building accurate and interpretable prediction models.RFERF is a novel approach that combines recursive feature elimination with random forest, effectively selecting the most relevant features for prediction.

5. Performance Evaluation Metrics

Studies have employed various evaluation metrics such as accuracy, sensitivity, specificity, and area under the ROC curve to assess the performance of cervical cancer prediction models.

IV. RESULTS AND DISCUSSION

1 Results of Descriptive Statics of Study Variables

Cervical HPV/DNA specimen results for HPV genotype	Number of women aged 35 (n=510)	Percentage %	Number of women aged 45 (n=502)	Percentage %
Negative	478	93.7	478	95.3
Positive	32	6.30%	24	4.8
By HPV genotype				
12 pooled positive	22	4.3	17	3.4

16 positive	9	1.8	5	1
18 positive	1	0.2	2	0.3

Table 1: Descriptive Statics

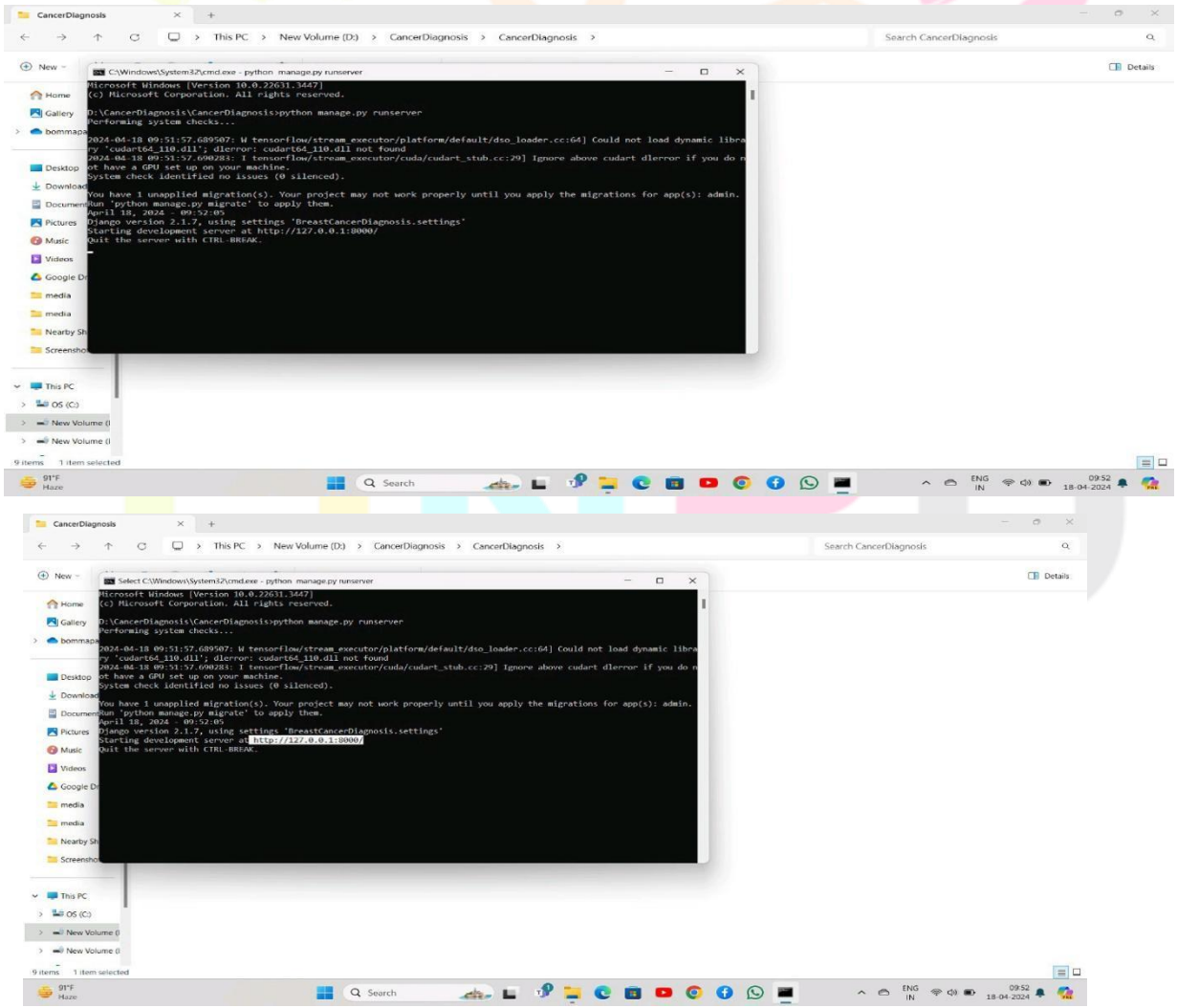
Distribution of High Risk Genotypes According to Cervical HPV/DNA Specimen Results

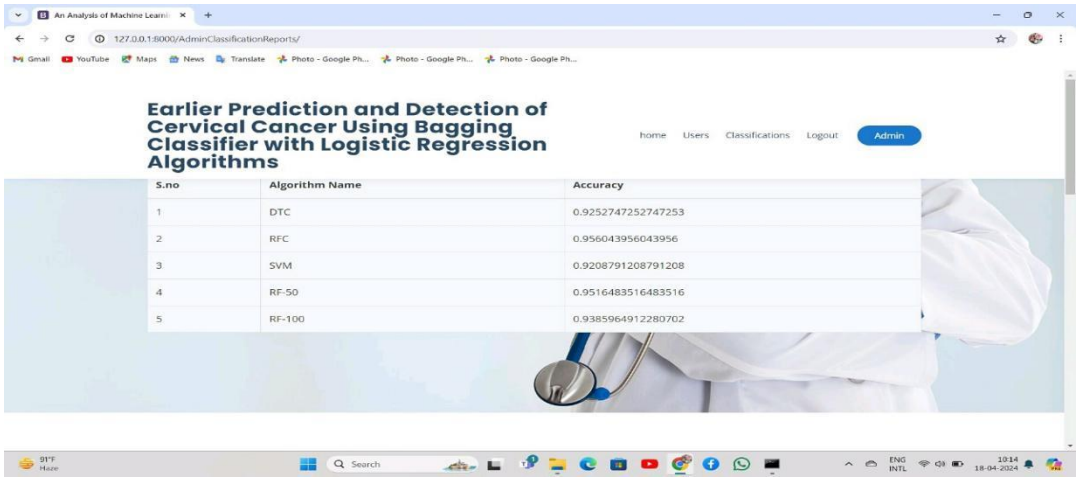
Colposcopy and biopsy was done for participants who screened positive by any method. Of the participants who were screened positive by HPV/DNA, confirmed CIN on Colposcopy was seen in 2.2% (n=11) and 1.8% (n=9) of participants in the 35-year cohort and 45-year cohort, respectively. Colposcopy confirmed CIN in 35-year cohort and 45-year cohort in participants screened positive by PAP and LBC

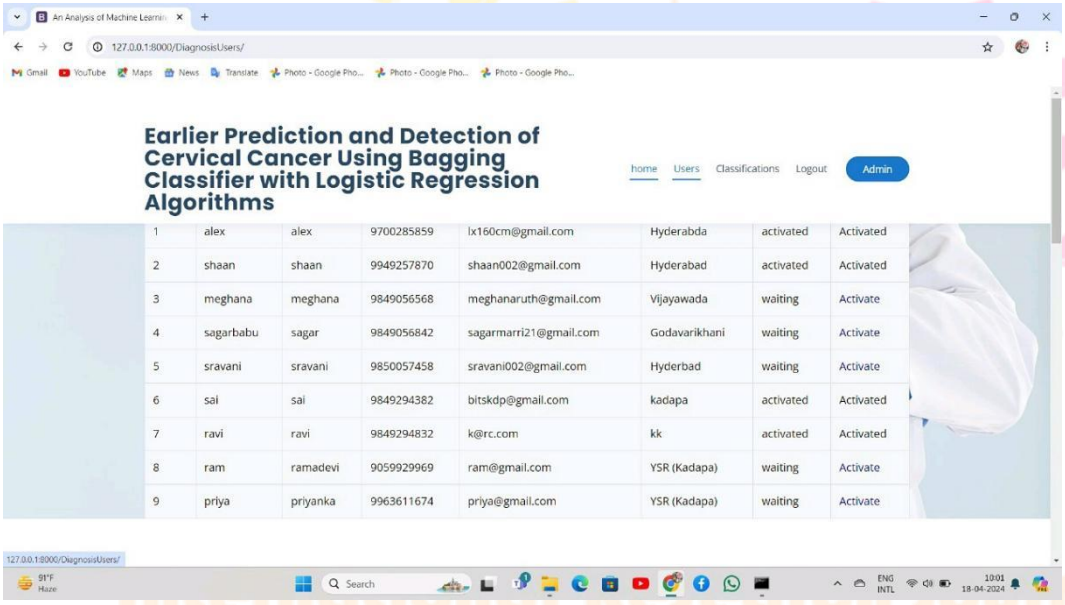
Exclusion Criteria at Clinic Setting

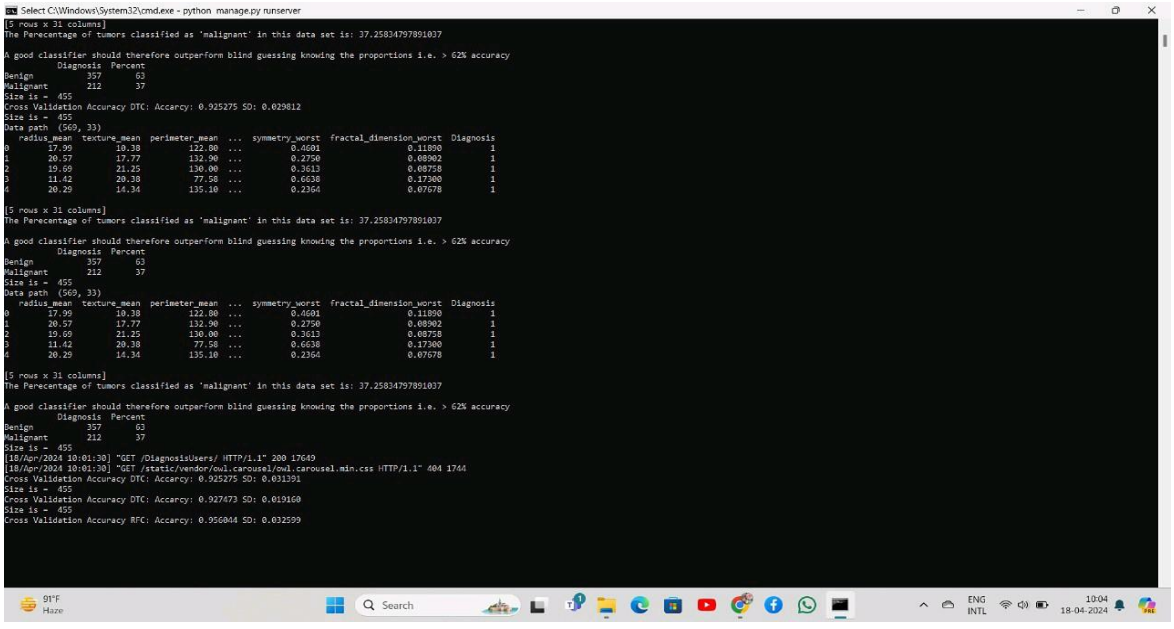
Reason	Number of women aged 35	Number of women aged 45
PV discharge	11	9
Pregnancy	6	0
Cervicitis	8	8
Cervical erosion	6	6
Fungal infection	5	8
Total excluded	36	31

RESULTS AND DISCUSSIONS




Earlier Prediction and Detection of Cervical Cancer Using Bagging Classifier with Logistic Regression Algorithms
home
Users
Classifications
Logout
Admin
S.no
Algorithm Name
Accuracy
1
DTC
0.9252747252747253
2
RFC
0.956043956043956
3
SVM
0.9208791208791208
4
RF-50
0.9516483516483516
5
RF-100
0.9385964912280702


Earlier Prediction and Detection of Cervical Cancer Using Bagging Classifier with Logistic Regression Algorithms
home
Users
Classifications
Logout
Admin
1 alex alex 9700285859 lx160cm@gmail.com Hyderabda activated Activated
2 shaan shaan 9949257870 shaan002@gmail.com Hyderabad activated Activated
3 meghana meghana 9849056568 meghanaruth@gmail.com Vijayawada waiting Activate
4 sagarbabu sagar 9849056842 sagarmarri21@gmail.com Godavarikhani waiting Activate
5 sravani sravani 9850057458 sravani002@gmail.com Hyderbad waiting Activate
6 sai sai 9849294382 bitskdp@gmail.com kadapa activated Activated
7 ravi ravi 9849294832 k@rc.com kk activated Activated
8 ram ramadevi 9059929969 ram@gmail.com YSR (Kadapa) waiting Activate
9 priya priyanka 9963611674 priya@gmail.com YSR (Kadapa) waiting Activate


Select C:\Windows\System32\cmd.exe - python manage.py runserver
[5 rows x 31 columns]
The Perecentage of tumors classified as 'malignant' in this data set is: 37.25834797891037
A good classifier should therefore outperform blind guessing knowing the proportions i.e. > 62% accuracy
Diagnosis Percent
Benign 357 63
Malignant 212 37
Size is - 455
Cross Validation Accuracy DTC: Accarcy: 0.925275 SD: 0.029812
Size is - 455
Data path (569, 33)
radius_mean texture_mean perimeter_mean ... symmetry_worst fractal_dimension_worst Diagnosis
0 17.99 10.38 122.80 ... 0.4601 0.11890 1
1 20.57 17.77 132.90 ... 0.2750 0.08902 1
2 19.69 21.25 130.00 ... 0.3613 0.08758 1
3 11.42 20.38 77.58 ... 0.6638 0.17300 1
4 20.29 14.34 135.10 ... 0.2364 0.07678 1
[5 rows x 31 columns]
The Perecentage of tumors classified as 'malignant' in this data set is: 37.25834797891037
A good classifier should therefore outperform blind guessing knowing the proportions i.e. > 62% accuracy
Diagnosis Percent
Benign 357 63
Malignant 212 37
Size is - 455
Data path (569, 33)
radius_mean texture_mean perimeter_mean ... symmetry_worst fractal_dimension_worst Diagnosis
0 17.99 10.38 122.80 ... 0.4601 0.11890 1
1 20.57 17.77 132.90 ... 0.2750 0.08902 1
2 19.69 21.25 130.00 ... 0.3613 0.08758 1
3 11.42 20.38 77.58 ... 0.6638 0.17300 1
4 20.29 14.34 135.10 ... 0.2364 0.07678 1
[5 rows x 31 columns]
The Perecentage of tumors classified as 'malignant' in this data set is: 37.25834797891037
A good classifier should therefore outperform blind guessing knowing the proportions i.e. > 62% accuracy
Diagnosis Percent
Benign 357 63
Malignant 212 37
Size is - 455
[18/Apr/2024 10:01:30] "GET /DiagnosisUsers/ HTTP/1.1" 200 17649
[18/Apr/2024 10:01:30] "GET /static/vendor/owl.carousel/owl.carousel.min.css HTTP/1.1" 404 1744
Cross Validation Accuracy DTC: Accarcy: 0.925275 SD: 0.031391
Size is - 455
Cross Validation Accuracy DTC: Accarcy: 0.927473 SD: 0.019160
Size is - 455
Cross Validation Accuracy RFC: Accarcy: 0.956044 SD: 0.032599

ACKNOWLEDGMENT

We thankful to all the referred journal authors for their collaborative study helps me to write this paper.

REFERENCES

[1] American Cancer Society. Cancer Facts & Figures, 2018.
[2] M. Schiffman, P.E. Castle, J. Jeronimo, A.C. Rodriguez, S. Wacholder, Human papillomavirus and cervical cancer, Lancet 370 (9590) (2007) 890–907.
[3] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D.M. Parkin, D. Forman, F. Bray, Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012, Int. J. Cancer 136 (5) (2015) E359–E386.
[4] Monica, Mishra, R. An epidemiological study of cervical and breast screening in India: district-level analysis. BMC Women's Health 20, 225 (2020). Doi: 10.1186/s12905-020-01083-6.
[5] G.A. Mishra, S.A. Pimple, S.S. Shastri, An overview of prevention and early detection of cervical cancers, Indian J. Med. Paediat. Oncol.: Off. J. Indian Soc. Med. Paediatric Oncol. 32 (3) (2011) 125–132, https://doi.org/10.4103/0971- 5851.92808.
[6] M. Rowe, An Introduction to Machine Learning for Clinicians, Acad Med. 94 (10) (2019 Oct) 1433–1436, https://doi.org/10.1097/ACM.0000000000002792,
PMID: 31094727.
[7] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Science & Business Media, New York, 2009.
[8] Y.S. Abu-Mostafa, M. Magdon-Ismail, H.-T. Lin, AMLbook.com, Learning from Data, 2012.
[9] J. De Fauw et al., Clinically Applicable Deep Learning For Diagnosis and Referral in Retinal Disease, Nat Med. 24 (9) (Sep 2018) 1342–1350.
[10] Qazi Mudassar Ilyas, Muneer Ahmad, An Enhanced Ensemble Diagnosis of Cervical Cancer: A Pursuit of Machine Intelligence Towards Sustainable Health, IEEE Access 9 (2021) 12374–12388, https://doi.org/10.1109/ Access.628763910.1109/ACCESS.2021.3049165.
[11] Nazim Razali, Salama Mostafa, Aida Mustapha, Abd Wahab, Mohd Helmy, Nurul Ibrahim, Risk Factors of Cervical Cancer using Classification in Data Mining, J. Phys.: Conf. Ser. 1529 (2020), https://doi.org/10.1088/1742-6596/ 1529/2/022102 022102.
[12] W. Yang, X. Gou, T. Xu, X. Yi, M. Jiang, Cervical Cancer Risk Prediction Model and Analysis of Risk Factors based on Machine Learning, Assoc. Comput. Mach. (2019).
[13] R. Vidya, G.M. Nasira, Prediction of Cervical Cancer using Hybrid Induction Technique : A Solution for Human Hereditary Disease Patterns, Indian J. Sci. Technol. 9 (August) (2016) 1–10.
[14] A A Abdullah, N K Abu Sabri, Wan Khairunizam, I Zunaidi, Z M Razlan, A B Shahriman, ''Development of predictive models for cervical cancer based on gene expression profiling data", in IOP Conf. Series, Mater. Sci. Eng. 557 (2019) 012003, https://doi.org/10.1088/1757-899X/557/1/012003.
[15] X. Deng, Y. Luo, C. Wang, ''Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods," in: Proceedings of 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems, CCIS 2018. pp. 631– 635, 2019.