ENSEMBLE MODELLING FOR THE PREDICTION OF CERVICAL CANCER BY ANALYZING DATA BALANCING TECHNIQUES

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

Article History: Received: 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

Abstract:

Cervical cancer is a disease that affects women and has a high fatality rate. Risk factors may help in advancing the cervical cancer early detection approach. Nonetheless, screening for cervical cancer at an earlier stage may reduce the risk of death and other complications. Regrettably, existing prediction algorithms need clinically relevant physiological and biochemical features, limiting their use to a narrower situation. To improve diagnostic, that would use a feature set with a decreased probability of occurrence in conjunction with three ensemble-based classification algorithms. The paper focuses on cervical cancer detection, which employs the advanced machine learning approach stacked unified machine learning (SUML) to improve the prediction models' performance. Stacking suitable machine learning employs a different set of learning algorithms. The screening data were arbitrarily divided into two groups: Training data accounted for 80% of the total and was used to construct the algorithm; testing data accounted for 20% of the total and evaluated the methods' validity. The random forest (RF) model and AdaBoost were employed to classify cervical cancer prognostic indicators. Furthermore, in previous cervical cancer detection studies that used fewer risk indicators, the accuracy of the recommended models is significant. As part of this research, we chose three of the most well-known machine learning algorithms and evaluated their accuracy in predicting cases of cervical cancer.

Index Terms: Cervical Cancer, Machine Learning (ML), Stacked Ensemble, stacked unified machine learning (SUM L).

[1]Department of Computer Science, CVR College of Engineering,
[2]Rector & professor, Department of Computer Science, JNTU Hyderabad

Email: [1]ch.bhavani@cvr.ac.in, [2]Govardhan_cse@jntuh.ac.in

DOI: 10.31838/ecb/2023.12.s3.259

1. Introduction

According to data [1] when detected early, cervical cancer has a 5-year survival rate, with estimates ranging from 80% to 90% [2]. When the disease reaches stage 4, the proportion of cured people falls to 10% [3]. As a result, cervical screening is critical for detecting cancer at an early stage and, as a result, reducing morbidity and death associated with the disease. Cervical cancer frequency and mortality rates vary significantly between countries, with highly developed countries having lower rates due to greater screening and immunization efforts [4]. In contrast, Developing countries' medical systems regularly neglect screening. This shows that detecting high-risk cervical cancer patients is more important to enhance screening intervals and better employ medical resources [5]. This type of cancer currently requires two tests: The patient must first get a pap smear or Papanicolaou test [6]. During this examination, cells are carefully scraped from the cervix and adjacent areas using argument so that they may be analyzed under a highly advanced microscope. This method makes identifying abnormal cells, including cancer cells, quite simple. The next step is a comprehensive colposcopy examination [7]. Numerous studies have found age-related differences in cancer risk. Although cervical cancer is preventable, most women are unaware of its etiology, health risks, prevention, and treatment, owing to their origins and education. Rich countries account for 95% of cervical cancer mortality [8]. Sexual interaction may also spread HPV. A higher risk of cervical cancer has been linked to a person's first sexual experience, age, the number of sexual partners, and the use of contraception [9]. Addressing these risk factors may prevent malignant tumors. Cervical cancer screening reduces infections and enhances cancer treatment [10]. Early-stage cancer diagnostic alternatives are also required. Hospitals have historically employed statistical approaches to define and evaluate the information because there is a small amount of data to work with, and the data is not highly sophisticated. Nonetheless, in this age of big data, data volume and complexity are increasing exponentially. Statistical techniques struggle to assess and exploit vast amounts of internal data appropriately and efficiently. In recent years, researchers in the field of medicine have begun to embrace various machine learning approaches, such as random forest (RF), support vector machine (SVM), decision tree (DT), neural network (NN), and others, due to the rapid rise of machine learning and data mining.

Finally, the group of research revealed that the data set from the UCI repository had many missing values, indicating that previous studies had overlooked at least two components. Patient privacy concerns prevented values from being recorded. After deleting two features with many missing values, SVM-PCA worked well. SMO and SMOTE-RF were high performances. Oversampling was used to correct the UCI cervical risk factor data difference. Deep learning may be helpful when there is inadequate data from the biopsy and other screening techniques. Age, first sexual encounter, number of pregnancies, smoking, hormonal contraception, intrauterine devices (IUD), sexually transmitted diseases (STDs), notably genital warts, and HPV infections are the most important risk factors. The significant machine learning classifier findings for cervical cancer prediction require additional research and refinement. Machine learning classifier results inspired this criterion. This research classified cervical cancer using SVM, Ada Boost, and RF ensemble-based classifiers. Identifying cancer risk factors is as important as separating malignant from noncancerous instances. SMOTE is also used to balance data classes since it's unbalanced.

The paper continues as follows. Initially section 2 first conduct a brief evaluation of relevant prior studies. Section 3 lays out the framework for our materials and practices. Section 4 analyses the test results, which show that our procedure is at the front of modern methodology. In the 5 section we serves as the paper's conclusion, we will finally provide a discussion.

Related Work

Many researchers investigated the Cox proportional hazard model alongside other predictive models, such as those based on machine learning and deep learning, in the early phases of the field's growth.

Modern hospitals can collect, retain, and exchange data thanks to the digital revolution and machine learning algorithms [11]. Decision trees [12] enhance cervical cancer detection. The experiment revealed the decision tree's accuracy: The biopsy test scored 92.54%, cytology 92.80%, Hinselmann 94.41%, and Schiller 90.44%. Bayes Net's cervical cancer categorization was the most accurate, identifying 97.26% of instances. MLP (95.89%) and k-Nearest Neighbour (95.89%) followed. MLP and k-Nearest Neighbour [13] have also been used to detect cervical cancer correctly. Bayes Net categorization accuracy was best after experiments. When the results of each model's test sets were added together, Naive Bayes came out on top with an accuracy score of 81%, followed by C4.5 (72%) and ID3 (69%).

The Iterative Dichotomous (ID3), C4.5, and Naive Bayes models were all tested to evaluate how well they predicted cervical cancer [14]. It is crucial to ascertain the frequency with which a model predicts an illness when the patient has that disease and the frequency with which it predicts no disease when the patient does not have that condition due to the sensitive nature of the medical diagnosis. This needs to be done along with determining

whether the model is accurate. The main explanation for this is that medical diagnosis is a sensitive process. The sensitivity and specificity of many models, including those covered in this section, are not shown; instead, they provide their degrees of accuracy or ratings based on prior research. Recently, several new research has been undertaken to look at the many approaches that may be used to determine the risk of cervical cancer [15–20].

2. Methodology

AdaBoost is a fast, efficient, and difficult-to-overfit classification algorithm, particularly for high-dimensional data, yet, it can only provide label categorization. The SVM algorithm with a linear kernel function can give a hyperplane depicting malware detection; however, this method's success depends on feature selection correctness. The RF model is better because it produces the most accurate data. We anticipate being able to make a prediction regarding cervical cancer based on the significant data of the models, which is shown in Figure 1.

Dataset:
UCI's repository provided the study's dataset. This 21-system has 858 scenarios and 32 attributes. Because some ladies were reluctant to share details, the data needed to include many variables. It was also unbalanced, with most instances not being malignant. Hinselmann, Schiller, cytology, and biopsy variables comprised the dataset. Each target variable gives a specific cervical cancer test.

Data Preprocessing:
All of the data linked with each predictive characteristic could not be obtained. It was anticipated that around 20%-30% of clinical prognostic data and 0%-15% of behavioral data needed to be included. It was essential to infer missing data from existing data. This replaced values for missing data. The vast amount of missing data prevented the use of typical methods for filling in the mean and median. These techniques cannot verify data. The filler values are mostly not real numbers. This reduces model accuracy. The data set is also class-imbalanced. The +e data set target labels had 35 Hinselmann, 74 Schiller, 44 Cytology, and 55 Biopsy records.
SMOTE corrected the class discrepancy. There are several ways to perform resampling on an imbalanced dataset, including SMOTE and the Bootstrap Method. To guarantee that our dataset is representative of the entire population, we will use a technique known as the synthetic minority oversampling technique (SMOTE), which will generate additional copies of our undersampled data at random. As can be seen from the counters for each sentiment category both before and after the SMOTE resampling, the data are now in a previously unknown balanced state.

Splitting Dataset:
We split our dataset into 80:20 portions for the training and test set.

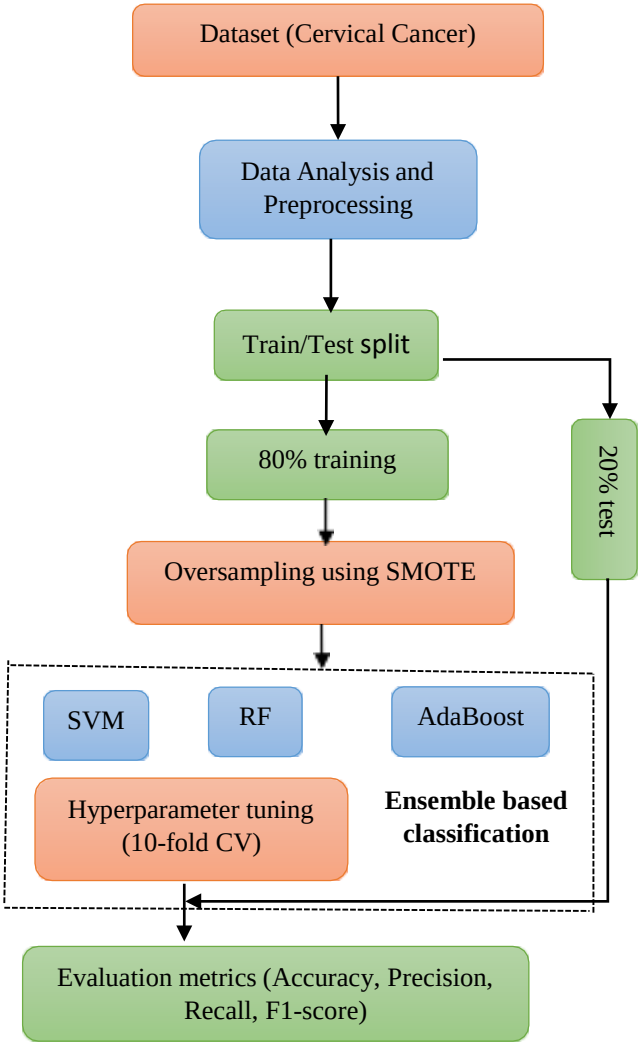


Figure 1: Cervical cancer prediction framework using machine learning

Stacked Ensemble-Based Classification Methods.
The following paragraph will describe the many available strategies.
We don't know which model best matches our data. As a result, we will need to test every available categorization model to determine the models that perform the best. The initial step in selecting the most successful model is to do many cross-validation processes. The Confusion Matrix and the F1 Score will be our primary measures, with the remaining metrics as secondary.

Random forest
Primary bagging-based ensemble method. Classifier operation: The classifier will generate k samples of D using the bootstrap approach, and each sample will be given the tag D_i. D_i uses replacement to sample the tuples of D. It's likely that some D tuples won't make it into D_i. when sampling with replacement, while others will be repeated. The classifier builds a decision tree based on each D_i. The creation of a "forest" made up of k decision trees is chosen. The random forest classifier has the highest accuracy compared to a single decision tree.

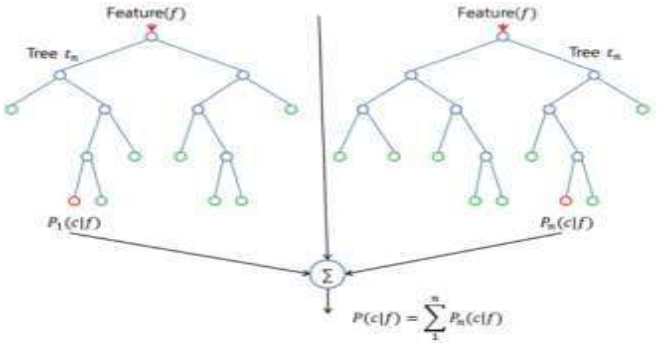


Figure 2: Tree structure of Random Forest model

Scikit-learn makes use of a method known as "decision trees," often known as CART (for "Classification and Regression Trees"). To categorize a tuple with an unknown category, X, each tree casts one vote in the form of its class prediction. The student in Class X who receives significant support can vote. The Gini index is used in the CART technique of tree construction. The Gini index for D may be determined using the formula:

$$\text{Gini(D)} = 1 - \sum_{i=1}^{m} p_i^2 \qquad (1)$$

Where p_i is the probability that a tuple in D belongs to class C_i, the Gini index calculates D's impureness level. The index value represents how successfully D was partitioned; the lower the number, the better.

Support vector machine:

SVMs can categorize both linear and nonlinear data sets. If the data can be divided into linear categories, the support vector machine can find the linear optimum separating hyperplane, also known as the linear kernel. This decision boundary separates the data for each class. The equation for a separating hyperplane is as follows: WX+b=0, where W is a weight vector containing w_1, w_2,...,w_n values, and X is a training tuple whrere B is a scalar. The best strategy to increase hyperplane efficiency is to decrease "W," which may be determined using the following formula:

$$\sum_{i=1}^{n} \alpha_i y_i x_i \qquad (2)$$

where αi are numeric parameters and yi are support vector labels, Xi=

$$\sum_{i=1}^{n} w_i x_i \geq 1 \qquad (3)$$

If y_i= -1 then

$$\sum_{i=1}^{n} w_i x_i \geq -1 \qquad (4)$$

If the data cannot be separated linearly, the SVM will employ nonlinear mapping to increase the dimension of the data. For this experiment,

$$K(X_i, X_j) = e^{-\gamma \|X_i - X_j\|^2/2} \qquad (5)$$

where γ is a free parameter that takes the scikit-learn default value for our experiment, and X_i and X_j are support vectors and testing tuples, respectively. The X_i is thought to be support vectors. X_j is presently researching tuples. The picture on the next page depicts a classification example using the SVM based on the linear kernel and the RBF kernel.

AdaBoost:

Progressive learning using ensembles creates a meta-classifier by combining the results of multiple less-accurate classifiers. It's called "progressive learning with ensembles." The AdaBoost algorithm, which boosts data samples, relies on adaptive sampling to give erroneously classified events large weights. The next iteration will select misclassified data to improve model training. Weighted voting is predicted and judged. AdaBoost uses decision tree stumps and forecasts better than bagging [24]. It also errors less. Each dataset sample starts with the same weight. Assume x is the dataset's sample count and y is the desired result. To reach a binary state, which may be

0 or 1, depends on the situation. After incorporating some data set records into the prototype decision tree stump, predictions can be made. The sample weights will be updated after the first prediction. Misclassified data samples were given more weight in the study. Next iteration, we'll choose samples with the highest weights. After reducing mistakes or reaching a target, the approach will be employed until it is no longer needed. Repeat until the error rate is satisfactory.

The step forward stage follows the combination stage of AdaBoost. Both phases use iterative and sequential methods. The first iteration assigns each training set instance a fixed weight. Error rates change the weights as iterations rise. The preceding phrase applies. Error-prone cases are weighted higher. The following equation illustrates the difficulty of classifying data into binary classes using T samples as training:

$$\{(x_i, y_i)\}_{i=1,}^{T} \text{ with } y_i \epsilon \{0,1\} \quad (6)$$

C is weak classifier linear combination. +e classifiers form a

$$C(x) = \sum_{n=1}^{N} w_n c_n(x) \quad (7)$$

where N is the total number of weak classifiers, w is the weights, and C(x) is the weak classifiers. Each cycle is used to train the classifier based on its previous performance.

$$C(x)_t = C(x)_{t-1} + w_n c_n(x) \quad (8)$$

where C(x)t is the t-iteration classifier. Classifier performance at t-1 is C(x).

This equation calculates weights:

$$w_n = \frac{1}{2} \ln \left(\frac{1 - \epsilon_n}{\epsilon_n} \right) \quad (9)$$

Model Building: We employ K-Fold Cross Validation (CV) on our early dataset (before resampling) since the CV is unaffected by an unbalanced dataset because it splits the dataset and considers each validation. In other words, the fact that the dataset is unbalanced does not affect CV. We should obtain results comparable to the original ones if we apply the CV to the balanced dataset generated by resampling.

3. Results and Discussions

The model was implemented in Python 3.8.0, and the environment in which it was done was Jupyter Notebook. The classifiers and other critical built-in tools came from the Ski-learn package; however, the XGBoost ensemble came from a separate library, specifically version 1.2.0 of the XGBoost library. The total number of evaluation criteria used was accuracy, sensitivity (recall), specificity (precision), positive predictive accuracy (PPA), and negative predictive accuracy (NPA). To discriminate between training and testing data, k-fold cross-validation with a factor of 10 is used.

Exploratory Data Analysis: In the example of EDA, we consider that the initial data set, which included 858 rows and 36 columns, contained many null values. The question mark (?) was used to represent non-existent values. Initially, this symbol was altered to "NaN" to make it easier to process. Following the biopsy, it was determined that most patients did not have cancer, and only 6.4% had cervical cancer, as shown in Figure 3.

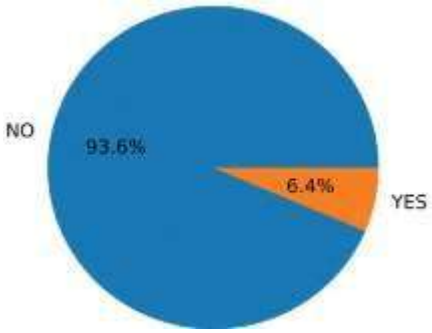
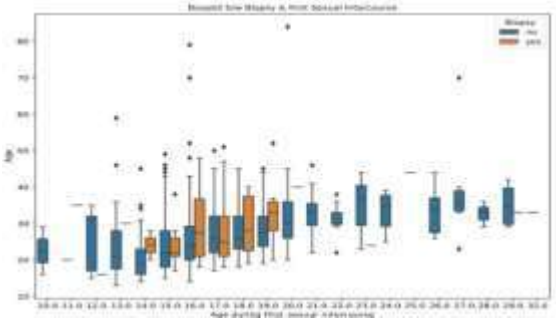
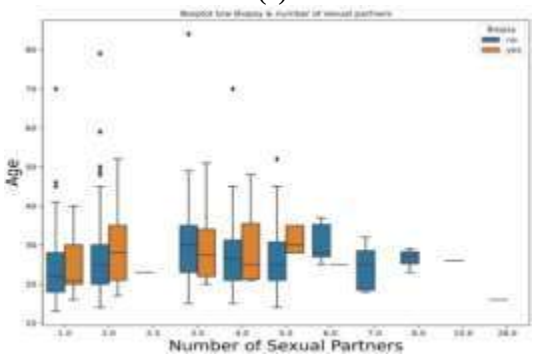


Figure 3: The percentage of positive biopsies

Participation in sexual activity at a young age (between 14 and 19) can result in a positive biopsy result. We went from univariate to multivariate analysis to better understand the problem. Figure 4a shows a box plot showing the relationship between a biopsy's results and a person's age and first sexual contact. This illustration clarifies the link.



(a)



(b)

Figure 4: A multivariate analysis was carried out using box plots

In Figure 4b, a box plot highlights the relationship between age, the number of sexual partners, and the biopsy result. This is done to emphasize the link. According to the data in the graphic, the number of sexual partners is directly associated with an individual's risk of developing cervical cancer.

Evaluation of the Sampling Method's Capability to Make Predictions

Table 1 compares and contrasts the relative performance ratings of several RF sampling strategies. Every example model had been double-checked both internally and externally.

Table 1: The random forest algorithm's prediction capability is evaluated using a range of sampling techniques.

Approach	**Acc**	**Pre**	**Sen**	**Spe**
Undersampling	0.375	0.152	0.167	0.741
Oversampling	0.608	0.159	0.909	0.639
SMOTE	**0.842**	**0.189**	**1.000**	0.62

Figure 5 displays the outcomes of an external validation performed on each classifier. Compared to the other choices, SMOTE-based RF performed very well, earning a score of 0.842 out of 1.00 on our accuracy criteria and the best possible rating on three of our four performance factors.

Compared to the undersampling and oversampling strategies, the precision was 100%, significantly higher than 70%. Consequently, SMOTE was chosen as the best method for the final model to utilize while processing imbalanced data.

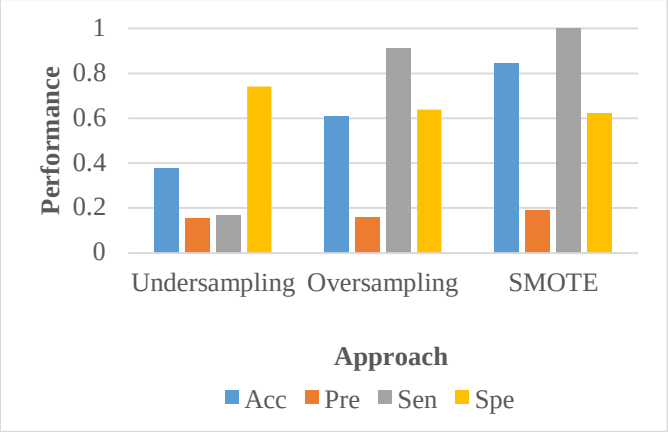


Figure 5: The random forest algorithm's performance in a range of sample schemes.

Comparison with Existing Studies

AdaBoost, extreme SVM, and Random Forest were three ensemble approaches used in the study. In addition, this is the first study of its kind to tackle the issue of feature selection and optimization in the context of cervical cancer using a bioinspired algorithm diagnosis. The study's findings were compared to those of the benchmark studies to determine the importance of the proposed investigation.

Table 2: a comparison of the proposed study to previous studies utilised as standards.

Target class	Authors	Accuracy	Sensitivity	Specificity	PPA	NPA
Hinselmann	Authors in [23]	97.6	96.65	98.54	98.48	96.78
	Authors in [24]	93.97	100	89.96	84.97	100
	Proposed work	98.21	100	98.65	98.65	97.84
Biopsy	Authors in [23]	96.06	94.94	97.76	97.58	94.91
	Authors in [24]	94.13	100	90.21	86.07	100
	Proposed work	95.57	100	91.25	92.14	100

The data set used for cervical cancer diagnosis served as the foundation for the criteria used to select the benchmark studies [25, 26]. Table 2 also compares the proposed strategy to studies that serve as benchmarks in the field of study. On the other hand, several of the earlier investigations' conclusions may have been obtained with fewer features.

4. Conclusion

The screening process for cervical cancer is investigated in this research. To detect cervical cancer cases, this work uses Random Forest, AdaBoost, and SVM. This data collection was made available via the machine learning library at UCI. Experiments focused on different target classes were conducted. In the process of data preparation, we use SMOTE to check for missing values as well as class balance. A comparison of model performance was carried out using SMOTED data, as well as specified qualities and features and chosen qualities and features. Combining the algorithms into many layers resulted in the creation of the stacked model. This particular classifier achieves an accuracy of 95.57%, a specificity of 100%, a positive predictive accuracy of 91.25%, and a negative predictive accuracy of 92.14%.

5. References

Wang L, Zhao Y, Xiong W, Ye W, Zhao W, Hua Y. MicroRNA-449a Is Downregulated in Cervical Cancer and Inhibits Proliferation, Migration, and Invasion. Oncol Res Treat (2019) 42(11):564–71. doi: 10.1159/000502122

Chao X, Li L, Wu M, Ma S, Tan X, Zhong S, et al. Efficacy of Different Surgical Approaches in the Clinical and Survival Outcomes of Patients with EarlyStage Cervical Cancer: Protocol of a Phase III Multicentre Randomised Controlled Trial in China. BMJ Open (2019) 9(7): e029055. doi: 10.1136/ bmjopen-2019-029055

Choi JB, Lee WK, Lee SG, Ryu H, Lee CR, Kang SW, et al. Long-Term Oncologic Outcomes of Papillary Thyroid Microcarcinoma According to the Presence of Clinically Apparent Lymph Node Metastasis: A Large Retrospective Analysis of 5,348 Patients. Cancer Manag Res (2018) 10:2883–91. doi: 10.2147/CMAR.S173853

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin (2018) 68 (6):394–424. doi: 10.3322/caac.21492

Campos NG, Alfaro K, Maza M, Sy S, Melendez M, Masch R, et al. The CostEffectiveness of Human Papillomavirus Self-Collection Among Cervical Cancer Screening non-Attenders in El Salvador. Prev Med (2020) 131:105931. doi: 10.1016/j.ypmed.2019.105931.

Da Silva, D. C., Garnelo, L., & Herkrath, F. J. (2022, April). Barriers to access the pap smear test for cervical cancer screening in rural riverside populations covered by a fluvial priMarchy healthcare team in the Amazon. International Journal of Environmental Research and Public Health, 19(7), 4193. https://doi. org/10.3390/ijerph19074193.

Ren, H., Jia, M., Zhao, S., Li, H., & Fan, S. (2022, February). Factors correlated with the accuracy of colposcopy-directed biopsy: A systematic review and meta-analysis. Journal of Investigative Surgery, 35(2), 284–292. https://doi.org/10.1080/08941939.2020. 1850944

Alquran, H. et al. (2022). Cervical cancer classification using combined machine learning and deep learning approach. Comput. Mater. Contin, 72(3), 5117–5134.

Al-Hashem, M. A. (2021). Performance evaluation of different machine learning classification algorithms for disease diagnosis. Ijehmc, 12(6), 1–28.

Arbyn, M., Castle, P. E., Schiffman, M., Wentzensen, N., Heckman-Stoddard, B., & Sahasrabuddhe, V. V. (2022). Meta-analysis of agreement/concordance statistics in studies comparing self- vs cliniciancollected samples for HPV testing in cervical cancer screening. International Journal of Cancer, 151(2), 308–312. https://doi.org/10.1002/ijc.33967

P. Julian, "The cervical cancer epidemic that screening has prevented in the UK." The Lancet 364.9430:249-256. 2004.

N. P. Pipti, N. P, Kishor, and H. A, "Cervical Cancer Test Identification Classifier Using Decision Tree Method," International Journal of Research in Advent Technology, Vol. 7, No. 4, E-ISSN:2321 – 9637. 2019.

F. Mohammed, U. Kadir, and S. Muciz, "Determining Cervical Cancer Possibility by Using Machine Learning Methods," International Journal of Latest Research in Engineering and Technology. ISSN:2454 – 5031. Vol 03 – Issue 12, pp:65 – 71. 2017.

R. Vidya, and G. M. Nasira, "Predicting Cervical Cancer Using Machine Learning Technologies – An Analysis," Global Journal of Pure and Applied Mathematics. ISSN 0973 –1768, Vol 12, No 3. 2016.

J. V. Kresta, J. F. MacGregor, and T. E. Marlin, ''Multivariate statistical monitoring of process operating performance,'' Can. J. Chem. Eng., vol. 69, no. 1, pp. 35–47, 1991.

J. L. Salmeron, S. A. Rahimi, A. M. Navali, and A. Sadeghpour, ''Medical diagnosis of Rheumatoid Arthritis using data driven PSOFCM with scarce datasets,'' Neuro computing, vol. 232, pp. 104–112, Apr. 2017.

S. Yin and Z. Huang, ''Performance monitoring for vehicle suspension system via fuzzy positivistic C-means clustering based on accelerometer measurements,'' IEEE/ASME Trans. Mechatronics, vol. 20, no. 5, pp. 2613–2620, Oct. 2015.

P. L. Rodrigues, N. F. Rodrigues, J. C. Fonseca, J. Correia-Pinto, and J. L. Vilaca, ''Automatic modeling of pectus excavatum corrective prosthesis using artificial neural networks,'' Med. Eng. Phys., vol. 36, no. 10, pp. 1338–1345, 2014.

S. Yin, H. Gao, J. Qiu, and O. Kaynak, ''Descriptor reduced-order sliding mode observers design for switched systems with sensor and actuator faults,'' Automatica, vol. 76, pp. 282–292, Feb. 2017.

S. Yin, H. Yang, and O. Kaynak, ''Sliding mode observer-based FTC for Markovian jump systems with actuator and sensor faults,'' IEEE Trans. Autom. Control, vol. 62, no. 7, pp. 3551–3558, Jul. 2017.

"Cervical cancer (risk factors) data set," 2020, https://archive. ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors% 29.

S. B. Kotsiantis, "Supervised machine learning: a review of classification techniques," Informatica, vol. 31, pp. 249–268, 2007

W. Wu and H. Zhou, "Data-driven diagnosis of cervical cancer with support vector machine-based approaches," IEEE Access, vol. 5, 2017.

S. F. Abdoh, M. Abo Rizka, and F. A. Maghraby, "Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques," IEEE Access, vol. 6, 2018.

Bhavani, C. H., & Govardhan, A. (2021). Supervised algorithms of machine learning in the prediction of cervical cancer: A comparative analysis. Annals of the Romanian Society for Cell Biology, 1380-1393.

CH. Bhavani, A. Govardhan, Cervical cancer prediction using stacked ensemble algorithm with SMOTE and RFERF, Materials Today: Proceedings, 2021, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.07.269.