XAI on Prediction of Cervical Cancer Using ML Classifiers with Missing Value Imputation

***Abstract*— Cancer has disrupted human life, causing premature deaths and cervical cancer has posed a massive threat to women's life over a long time. Prior prediction of this from basic diagnosis can be life-saving. In basic diagnosis patients are not always comfortable sharing all personal info. So, here we have approached by using ensembled learning for imputation of missing data and then used Logistic Regression, Multinomial Naïve Bayes, K- Nearest Neighbor, Random Forest, Stochastic Gradient Descent, Support Vector Machine, Decision Tree Classifier, and XGBoost Classifier for prediction where Multinomial Naïve Bayes has performed out others. On each algorithm explainable AI Lime has been used for interpretability where 'Schiller' has shown the most relevancy.**

Keywords—ML, XAI, Lime, Imputation, Cervical, Classifier.

I. INTRODUCTION

Cervical Cancer is one of the most common and serious ones among women. A lot of women die of this every year around the world. Early treatment of this cancer may result in recovering the patient from this deadly disease. So if we were able to predict this disease early just by having the basic diagnosis then it would be very fruitful and starting early treatment would be lifesaving. There some works have been done on different machine learning approaches for early predicting this. Like some have applied random forest for the prediction of this cancer with different feature extraction techniques [1]. Then there's been the use of the Gaussian Mixture Model for sampling with Neural Network in the identification of precancerous cells by precise removal of highlights [2]. Building models like logistic regression, decision tree, k nearest neighbor (KNN), random forest, and support vector machine using input features like lifestyle, habits, medical history of the patients, and sexual practice for predicting [3]. Another work where ensembled learning method has been applied for explainability and interpretability with Lime and Shapley for predicting this cervical cancer [4]. Comparison of performance of different ML classifiers (LR, Random Forest, NB, SVM, SGD, and KNN) for performance analysis [5]. Another method has generated models for predicting disease recurrence and mortality in people with early cervical carcinoma following primary RH. Because the anticipated value is determined using logistic regression models comprised of preoperative variables, it is based on preoperative parameters such as the surgical technique [6]. Here in our paper we have proposed an approach for comparing eight classifiers algorithms considering the Logistic Regression (LR), Multinomial Naïve Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), Decision Tree Classifier (DT), and XGBoost Classifier (XGB) with missing value imputation and further explained by explainable AI.

II. WORKING METHOD

A. Methodology

Here, in Fig. 1 for Predicting the detection of Cervical cancer we have taken the dataset of 858 patients containing the demographic information with previous medical records, which counts 36 information (feature) columns from the Hospital Universitario de Caracas [7]. After having the dataset, it has been checked for missing values. Some missing values are obvious for patients not willing of sharing all personal information at basic diagnosis.

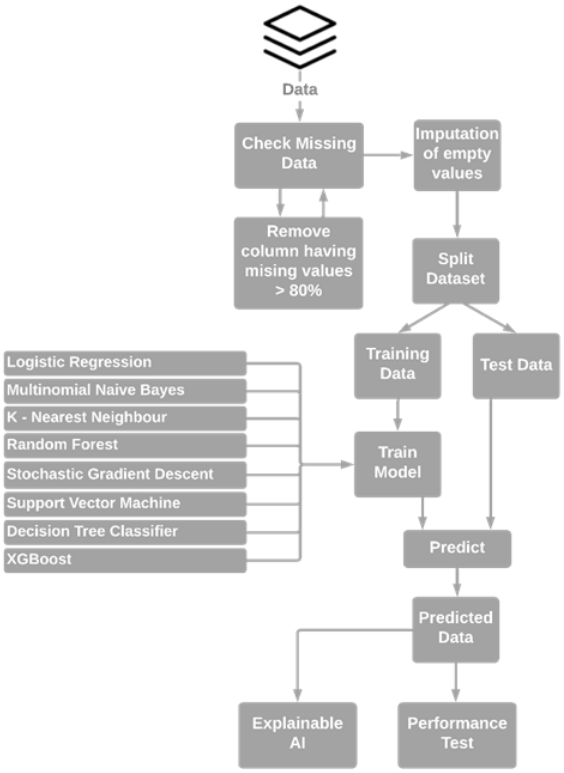

Fig. 1. Methodology.

For compensating empty values, we have implemented a machine-learning imputation approach. For predicting risk, we have defined the biopsy as our output class. Then the dataset is defined into two parts as training and test. The training set is used to train the model with our machine learning algorithms one by one and later used to predict using both training data and test data for executing performance analytical tasks like accuracy, recall, precision, f1-score, and error matrices like mean squared error (MSE) and root mean squared error (RMSE). As the data are labeled, the algorithms we have used are all supervised. Here on the predicted data, we have applied the explainable AI (XAI) to understand, how our predicted outputs are related to the input features.

B. Dataset

The dataset we have selected for analyzing the early detection of cervical cancer has been collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela [7]. This data set contains 36 feature columns which include statistical data of patients along with their habits, some previous medical records, and so on. This dataset contains records of 858 patients. As we have said this dataset contains patients' personal information so at basic diagnosis patients are not willing to share all the personal information which causes a lot of missing data. In Fig. 2 we can see the percentage of missing data in our dataset.

Age	0.000000
Numberofsexualpartners	3.030303
Firstsexualintercourse	0.815851
Numofpregnancies	6.526807
Smokes	1.515152
Smokesyears	1.515152
Smokespacksyear	1.515152
HormonalContraceptives	12.587413
HormonalContraceptivesyears	12.587413
IUD	13.636364
IUDyears	13.636364
STDs	12.237762
STDsnumber	12.237762
STDscondylomatosis	12.237762
STDscervicalcondylomatosis	12.237762
STDsvaginalcondylomatosis	12.237762
STDsvulvoperinealcondylomatosis	12.237762
STDssyphilis	12.237762
STDspelvicinflammatorydisease	12.237762
STDsgenitalherpes	12.237762
STDsmolluscumcontagiosum	12.237762
STDsAIDS	12.237762
STDsHIV	12.237762
STDsHepatitisB	12.237762
STDsHPV	12.237762
STDsNumberofdiagnosis	0.000000
STDs: Time since first diagnosis	91.724942
STDs: Time since last diagnosis	91.724942
DxCancer	0.000000
DxCIN	0.000000
DxHPV	0.000000
Dx	0.000000
Hinselmann	0.000000
Schiller	0.000000
Citology	0.000000
Biopsy	0.000000

Fig. 2. Missing values percentage in each column.

In Fig. 3 we have placed the proportion of our output class 'Biopsy' data where it is visible that how much the much of positives (1) and negatives (0) are in imbalanced proportion.

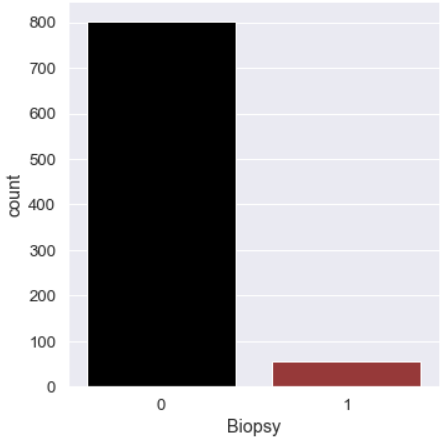

Fig. 3. Biopsy.

C. *Imputation of missing data*

Before performing the imputation of missing data, we have cleaned the dataset by removing two of the columns named 'STDs: Time since the first diagnosis' and 'STDs: Time since first diagnosis' which having missing values of more than 91%, can be identified in the fig. 2. This results in 34 remaining columns.

There are different approaches for imputation like median [8], mean [9], or mode [10]. As we are dealing with medical data here, it will be too gross to use these techniques. So here, we have used the ensembled approach for that we have used the random forest (RF). In our dataset some features are categorical, and some are of discrete numerical values for different patients. For predicting empty categorical values RF classifier and for numerical values RF regressor has been used.

D. *Training the model using classifiers*

Training has been done on different machine learning classifier algorithms. Before training the whole dataset is divided into input and output features. Here biopsy has been defined to be our output labeled class, so here biopsy 1 means cancer positive, and 0 represents no cervical cancer. Then the whole dataset has been divided into test data and training data. We have defined the training data sequentially as 80, 70, 60, 40, and 20 where the rest will be test data respectively as 20, 30, 40, 60, and 80. For training the model we have used several classifiers which are Logistic Regression (LR), Multinomial Naïve Bayes (NB), K- Nearest Neighbor (KNN), Random Forest (RF), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM), Decision Tree Classifier (DT), and XGBoost Classifier (XGB).

E. *Performance test*

For performance tests, we have made a prediction on both the test data and trained data to measure if it's overfitting or not. For this, we have used measured the MSE and RMSE [13]. We have also calculated the True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) [12] and concurrently calculated the accuracy, precision, recall, and f1-score [13].

F. *Deploying Explainability*

We have depicted earlier that we have used 34 features for training the model. By observing the predicted values, we cannot tell which features are having an impact on the resulted output. Here we need the explainability. For clarifying the outputs of the classifiers, we have used Local Interpretable Model-agnostic Explanations (LIME) [11].

LIME: Lime is a model-agnostic, this interprets how much the model is trustworthy in providing prediction. If we define a class of potentially interpretable models as g $\in$ G, L (f, g, π x) be a measurement of the unfaithfulness of g is in approximating f in the locality. Here f(x) is the probability that x belongs to a certain class and πx is a proximity measure. If complexity is represented by Ω(g) then the explanation can be represented as [11],

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

III. Experimental Result Analysis

A. *Classifiers Performance Analysis*

Here, we have got the performance analytics of different classifier models after predicting on the ensemble approached imputed dataset on cervical cancer. In Table 1, we have placed the confusion matrix outputs which are True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Using these matrices we have evaluated the precision, recall, f1-score, and accuracy shown in Table 2 where eight ML classifiers we have used in the training model here for comparative predictability analysis. Here in Table 2, we can see that the Multinomial Naïve Bayes (NB) is the best performer according to the accuracy and f1-score. We need to take into account both the accuracy and f1-score because f1 indicates the correct detection rate where accuracy gives a gross of correct identification overall observations. So here from statistical data, this can be stated that NB has been followed by the KNN, XGB, and others where DT is comparatively the worst performer. With lower precision (percentage of positive observations) value comes a higher

recall (percentage of genuine positive observations) value can be seen in DT, LR, SVM, and SGD. In RF, XGB, KNN, and Multinomial NB both precision and recall values are pretty decent so they come up with better accuracy and f1-score. In Table 3. We have placed the error values which are Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) calculated for both training and test data. Here, all the models for MSE show a little higher error in test data than training data, as test data are yet unknown to the model. For RMSE values are the same in both training data and test data. As the error rate in MSE and RMSE are both relatively lower so the models can be represented as a good fit. For the error rate in test data considering both MSE and RMSE, RF can be said as the best fit then followed by XGB, NB, KNN, and others.

TABLE I. CONFUSION MATRIX OUTPUT TP, FP, FN, & TN FOR ALL CLASSIFIERS

Classifiers	TP	FP	FN	TN
DT	155	10	2	5
LR	157	8	2	5
SVM	157	8	2	5
SGD	158	7	2	5
RF	160	5	3	4
XGB	160	5	2	5
KNN	162	3	4	3
NB	161	4	2	5

TABLE II. ACCURACY (ACC), PRECISION, RECALL, AND F1-SCORE.

Classifiers	Acc	Precision	Recall	f1-score
DT	92.44	93.93	98.1	95.97
LR	94.18	95.15	98.74	96.91
SVM	94.18	95.15	98.74	96.91
SGD	94.76	95.75	98.75	97.23
RF	95.34	96.96	98.15	97.56
XGB	95.93	96.96	98.76	97.85
KNN	95.93	**98.18**	97.59	97.88
NB	**96.51**	97.57	**98.77**	**98.17**

TABLE III. MSE AND RMSE OF TRAINING AND TEST DATA FOR ALL CLASSIFIERS.

Classifiers	Train Data		Test Data	
	MSE	*RMSE*	*MSE*	*RMSE*
DT	0.0	0.0	0.0755	0.0
LR	0.0291	0.1707	0.0581	0.1707
SVM	0.0174	0.1322	0.0581	0.1322
SGD	0.0306	0.1749	0.0755	0.1749
RF	0.0102	0.1010	0.0465	0.1010
XGB	0.0	0.0	0.0639	0.0
KNN	0.0481	0.2193	0.0406	0.2193
NB	0.0466	0.2159	0.0348	0.2159

Then we have made some gradual changes in the size of training data 80%, 70%, 60%, 40%, and 20% respectively of the whole dataset. In Fig. 4 on the x-axis, we have placed the training data in percentage and accuracy is on the y-axis. Here in Fig. 4, we can see that NB has a drastic fall in performance with decreasing training data. RF, XGB, and KNN have a little bit of a slower decrease rate in the performance. But for DT its accuracy has increased which is an unusual thing to happen. Same way for SGD, SVM, and LR accuracy has first decreased with lower training data and then suddenly gone upwards at 40 and 20 percent training data, these are also unusual behavior. As we have seen in Fig. 3 that our dataset is an unbalanced one according to the proportion of output class so when training data decreases it gets a comparatively larger data of only one class to train with and making predictions on the same kind of class on a larger quantity with increasing test data size accordingly. So, this is how this unusual behavior can be explained.

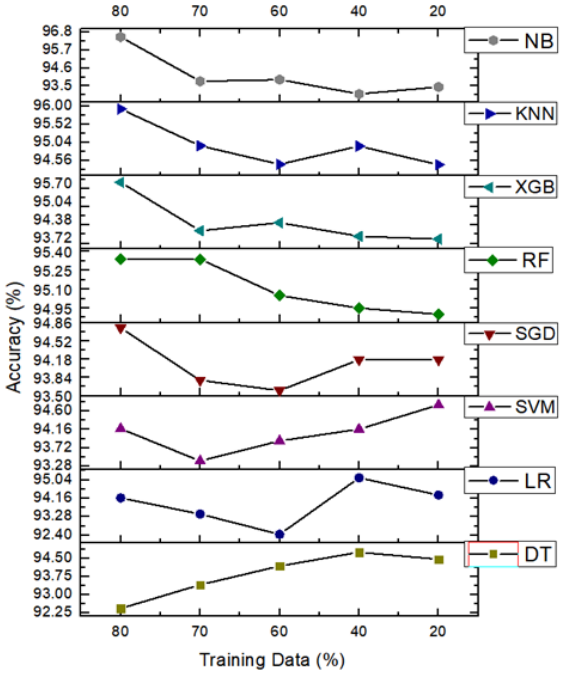

Fig. 4. Change of accuracy with the changing size of training data for all used classifiers.

B. *Feature explanation from LIME*

Here, from Fig. 5 to Fig 12 the output data of lime has been presented. When the blue ones are high means the biopsy is '0', cervical negative and the orange-colored ones are meaning biopsy '1' means cervical cancer positive..

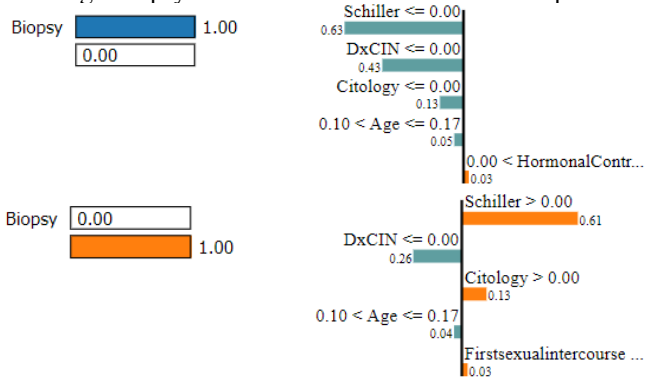

Fig. 5. Decision Tree Classifier (Lime).

In Fig. 5 for Decision Tree, 'Schiller' and 'Citology' are most impactful in both states.

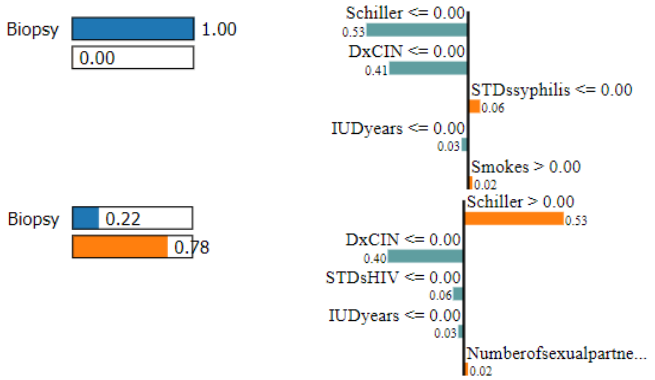

Fig. 6. Logistic Regression (Lime).

In Fig. 6 ‘Schiller’ is the most impactful in both states for Logistic regression.

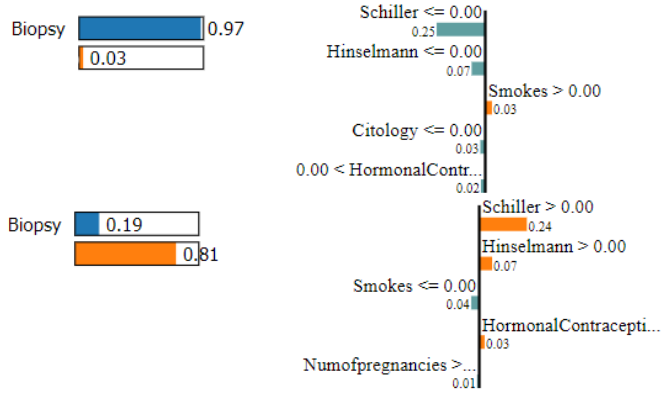

Fig. 7. Support Vector Machine (Lime).

In Fig. 7 for Support Vector Machine, ‘Schiller’ and ‘Hinselmann’ are most impactful in both states.

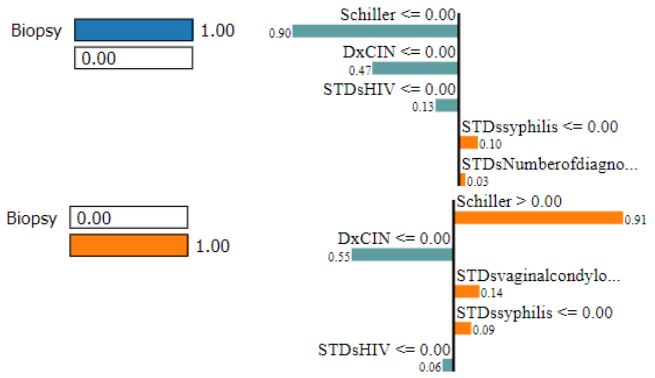

Fig. 8. Stochastic Gradient Descent (Lime).

In Fig. 8 for Stochastic Gradient Descent, ‘Schiller’ is most impactful in both states.

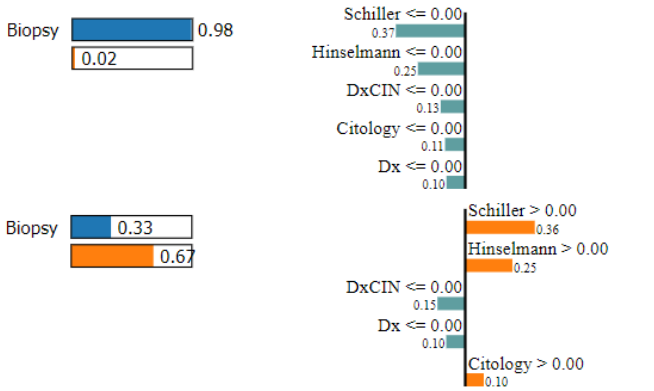

Fig. 9. Random Forest (Lime).

In Fig. 9 for Random Forest, ‘Schiller’ and ‘Hinselmann’ are most impactful in both states.

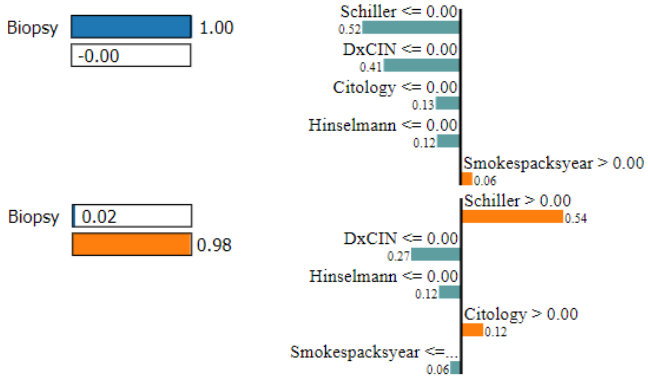

Fig. 10. XGBoost Classifier (Lime).

In Fig. 10 for XGBoost Classifier, ‘Schiller’ and ‘Citology’ are most impactful in both states.

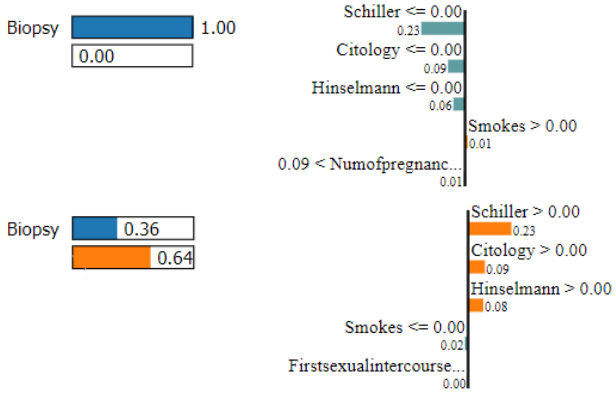

Fig. 11. K-Nearest Neighbour (Lime).

In Fig. 11 for K-Nearest Neighbour, ‘Schiller’, ‘Citology’ and ‘Hinselmann’ are most impactful in both states.

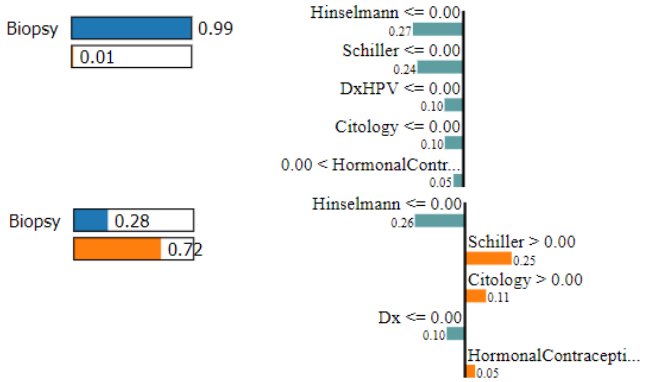

Fig. 12. Multinomial Naïve Bayes (Lime).

In Fig. 12 for Multinomial Naïve Bayes, 'Schiller' and 'Citology' are most impactful in both states.

We have only presented the top five features for visualization having an impact on the state of the output class. High feature values with the corresponding color have an impact on class outputs. We have exhibited the explanation of LIME of both classes for all the classifiers. So, it can be seen that the most effectual on the output is the 'Schiller' then followed by then 'Citology', 'Hinselmann', 'DxCIN', 'Hormonal Contraceptoives', 'IUDyears', and so on. So, this is how Lime creates transparency in output class about the impact of training features.

IV. Conclusion

Cervical cancer is a disease that develops slowly but can be deadly. By early detection may be a good way to prevent this carcinoma.

In our work, we have used explainable AI to interpret the prediction of cervical cancer. We also used eight ML classifiers which already showed in previous sections. Have compared Performance of those classifiers and Multinomial NB has outperformed other but performance decreases with lower training data. We have performed all these tasks after imputing missing values. Lime has created transparency of input features on prediction outputs. In the future, we can take this work further by comparing different imputing techniques for perfecting the system and creating a balanced dataset for training the models.

References

[1] P. Nagpal, and P. Arora, "Prediction Model for Cervical Cancer in Female Patients Using Machine Learning, In: Sheth A., Sinhal A., Shrivastava A., Pandey A.K. (eds) Intelligent Systems. Algorithms for Intelligent Systems. Springer, Singapore, 2021.

[2] N. S. Benita, S. Vaishnavi, and G. Kalaiarasi, "Survey of Cervical Cancer Prediction Using Machine Learning," In: Mallick P.K., Bhoi A.K., Marques G., Hugo C. de Albuquerque V. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 1317. Springer, Singapore, 2021.

[3] K. Suresh, "Classification Study and Prediction of Cervical Cancer," In: Chiplunkar N., Fukao T. (eds) Advances in Artificial Intelligence and Data Engineering. Advances in Intelligent Systems and Computing, vol 1133. Springer, Singapore, 2021.

[4] F. Curia, Cervical cancer risk prediction with robust ensemble and explainable black boxes method," Health Technol. 11, 875–885, 2021.

[5] M. A. Haque, I. J. Dristy, S. Sharar, A. A. Rasel, "ML Classifier Comparative Performance Analysis of Prediction on Cervical Cancer," International Conference on Electronics, Communications and Information Technology 2021, in press.

[6] S. I. Kim, S. Lee, C. H. Choi, M. Lee, D. H. Suh, H. S. Kim, K. Kim, H. H. Chung, J. H. No, J. Kim, N. H. Park, Y. S. Song, and Y. B. Kim, "Machine Learning Models to Predict Survival Outcomes According to the Surgical Approach of Primary Radical Hysterectomy in Patients with Early Cervical Cancer," Cancers 13, no. 15: 3709, 2021.

[7] Fernandes, Kelwin, Cardoso, Jaime & Fernandes, Jessica, "Cervical cancer (Risk Factors)," UCI Machine Learning Repository, 2017.

[8] A. R. Kapil, "Methods of Missing Value Treatment and their effect on the Accuracy of Classification Models," 2018, 10.13140/RG.2.2.31137.86881.

[9] G. Dr. Madhu, B. L. Bharadwaj, S. Vardhan, and C. Gogulamudi, "A Normalized Mean Algorithm for Imputation of Missing Data Values in Medical Databases," 2020, 10.1007/978-981-15-3172-9_72.

[10] T. Aljuaid, and S. Sasi, "Proper imputation techniques for missing values in data sets", 1-5, 2015. 10.1109/ICDSE.2016.7823957

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," 2016, 1602.04938.

[12] M. Vakili, M. Ghamasari, and M. Rezaei, "Performance Analysis and Comparison of Machine and Deep Learning Algorithms for IoT Data Classification", Researchgate, 2020.

[13] K. Marwan and N. Sanam, "Performance Analysis of RegressionMachine Learning Algorithms for Prediction of Runoff Time", Agrotechnology, 2019.