# **AI Contribution to Cervical Cancer Prediction**

**1. Project Title**

Artificial Intelligence contribution to Cervical Cancer prediction.

Team members: George Adamides, George Koumis, Irene Kyriacou

**2. Results**

**2.1 Proposed Models:**

The two machine learning algorithms that we used are the following:

1. Random Forest
2. XGBoost

1.Random Forest

The first machine learning model that we will use within the framework of this project is called Random Forest and it makes predictions by fusing different decision trees. A random subset of the data and a random subset of the characteristics are used to train each decision tree in the random forest. The final forecast is then generated by combining the predictions of all the trees. This aids in lowering overfitting and improving model accuracy. Random forests are well renowned for their robustness and capacity to handle high-dimensional data and are frequently used for classification and regression applications. [1]

2.XGBoost (eXtreme Gradient Boosting)

The second machine learning model that is going to be investigated is a well-liked machine learning technique for supervised learning tasks, particularly for classification and regression issues, and in particular the XGBoost. To build a stronger model, it iteratively joins weak decision trees using a gradient boosting framework. In order to reduce overfitting and enhance generalization, XGBoost additionally incorporates regularization techniques. It is now a standard approach for many data science contests and is extensively utilized in business for a variety of purposes, including recommendation systems, fraud detection, and customer churn prediction.[2]

**2.2 Evaluation Metrics for the models**

**Without PCA feature selection:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Random Forest** | | **ACTUAL VALUE** | |
| **POSITIVE** | **NEGATIVE** |
| **PREDICTION VALUE** | **POSITIVE** | **11 (TP)** | **46 (FP)** |
| **NEGATIVE** | **7 (FN)** | **108 (TN)** |

Table Random Forest algorithm without PCA feature selection

|  |  |  |  |
| --- | --- | --- | --- |
| **XGB** | | **ACTUAL VALUE** | |
| **POSITIVE** | **NEGATIVE** |
| **PREDICTION VALUE** | **POSITIVE** | **6 (TP)** | **26 (FP)** |
| **NEGATIVE** | **12 (FN)** | **128 (TN)** |

Table XGBoost algorithm without PCA feature selection.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **AUC** | **ACCURACY** | **SENSITIVITY** | **SPECIFICITY** | **F1 score for positives** | **F1 score for negatives** | **F-beta** |
| **Random Forest** | **0.65** | **0.69** | **0.61** | **0.7** | **0.29** | **0.8** | **0.65** |
| **XGB** | **0.58** | **0.78** | **0.33** | **0.83** | **0.24** | **0.87** | **0.58** |

Table Performance metrics for algorithms without the use of PCA method.

**With PCA feature selection:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Random Forest** | | **ACTUAL VALUE** | |
| **POSITIVE** | **NEGATIVE** |
| **PREDICTION VALUE** | **POSITIVE** | **13 (TP)** | **63 (FP)** |
| **NEGATIVE** | **5 (FN)** | **91 (TN)** |

Table Random Forest algorithm with PCA feature selection.

|  |  |  |  |
| --- | --- | --- | --- |
| **XGB** | | **ACTUAL VALUE** | |
| **POSITIVE** | **NEGATIVE** |
| **PREDICTION VALUE** | **POSITIVE** | **9 (TP)** | **39 (FP)** |
| **NEGATIVE** | **9 (FN)** | **115 (TN)** |

Table XGBoost algorithm with PCA feature selection.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **AUC** | **ACCURACY** | **SENSITIVITY** | **SPECIFICITY** | **F1 score for positives** | **F1 score for negatives** | **F-beta** |
| **Random Forest** | **0.65** | **0.6** | **0.72** | **0.59** | **0.27** | **0.72** | **0.64** |
| **XGB** | **0.62** | **0.72** | **0.5** | **0.75** | **0.27** | **0.82** | **0.62** |

Table Performance metrics for algorithms with the use of PCA method.

In the above tables we can find different metrics that evaluate the performance of the different methods/models that have been used. First of all we have the Area Under the Curve (AUC) that gives an aggregated measure of performance, then the accuracy that indicates how many times the model was correct (true predictions). Secondly, we have the sensitivity that evaluates the model’s ability to predict true positive values for each category. The prediction of the true positives for a variable is a very important metric to take into consideration, because when it comes to the prediction of the disease it’s disastrous to diagnose as negative a person that is in fact positive. In that way he will no longer proceed to further tests, and he will remain undiagnosed until probably too late. So, if the model predicts the True Positives exceptionally well then it does not leave undiagnosed the patients and eliminates the prediction of the false negative cases. Specificity on the other hand shows the model’s ability to predict true negatives. Further we process with the calculation of the F scores: F1 score for positives(known as F1 score), F1 for negatives. Those F scores evaluate the predictive skills of a model combining the precision and the sensitivity/ specificity (respectively to F1 score for positives and F1 for negatives). [3]

At this point we would like to introduce a new metric called F-beta that has not been introduced in previous assignments. The Fbeta-measure is a generalization of the F-measure that adds a configuration parameter called beta. A default beta value is 1.0, which is the same as the F-measure. A smaller beta value, such as 0.5, gives more weight to precision and less to recall, whereas a larger beta value, such as 2.0, gives less weight to precision and more weight to recall in the calculation of the score.[4]

**3. Discussion**

**3.1 Main findings**

For a more easy discussion of the results a summary table can be found below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Without PCA | | With PCA | |
|  | Random Forest | XGB | Random Forest | XGB |
| **AUC** | 0.65 | 0.58 | 0.65 | 0.62 |
| **Accuracy** | 0.69 | 0.78 | 0.6 | 0.72 |
| **Sensitivity** | 0.61 | 0.33 | 0.72 | 0.5 |
| **Specificity** | 0.7 | 0.83 | 0.59 | 0.75 |
| **F1 score for positives** | 0.29 | 0.24 | 0.27 | 0.27 |
| **F1 score for negatives** | 0.8 | 0.87 | 0.72 | 0.82 |
| **F-beta** | 0.65 | 0.58 | 0.64 | 0.62 |

Table Concentrated result's table.

As we can see from Table 7, the different models with or without the PCA feature selection perform differently on the various tests. For instance, when it comes to accuracy the XGB model without PCA, performs the best with a score of 0.78. This comes to a contrast with the performance of the same model in the sensitivity metric that scores only 0.33, that is not only the lower, compared to the performance of the other models but is a very low score in general. Some of the above metrics like F1 for positives, or sensitivity have extremely low values in some cases. Those results are unacceptable for a model and therefore cannot be evaluated furthermore. For that reason, we proceed with the cross validation (k=5) method.

The main reason to use cross-validation was to evaluate the generalization ability of our model and to get a more accurate estimate of how well the model will perform on new data, because the model is tested on multiple different subsets of the data.

A very interesting turn of events occurred as we can see from Table 8 below. Our results in this case -cross validation- show a remarkable improvement compared to the Table 7 above. For this reason, we will proceed with the results discussion using the results from the cross validation.

**Cross validation (K=5) table:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Without PCA** | | **With PCA** | |
|  | **Random Forest** | **XGB** | **Random Forest** | **XGB** |
| **AUC** | 0.88 | 0.96 | 0.85 | 0.94 |
| **Accuracy** | 0.79 | 0.90 | 0.76 | 0.88 |
| **Sensitivity** | 0.73 | 0.88 | 0.84 | 0.92 |
| **Specificity** | 0.86 | 0.93 | 0.69 | 0.83 |
| **F1 score for positives** | 0.77 | 0.88 | 0.79 | 0.89 |
| **F1 score for negatives** | 0.81 | 0.92 | 0.75 | 0.88 |
| **F-beta** | 0.73 | 0.88 | 0.84 | 0.93 |

Table Concentrated result's table after cross validation.

According to the Table 8 all of the models with the cross validation performed surprisingly well especially when we compare them with the Table 7. The lowest value a metric took was 0.69. This was the random forest with PCA when it comes to specificity. Even though that was the lowest nevertheless it was an acceptable value for a performance metric. The highest was the AUC in the XGB without PCA model with a value that reached 0.96. This is an incredibly high score that shows how well our model performs.

As a way to choose which one of the above-mentioned models performs the best, we will take into account which one achieves the best results in most metrics. There is no doubt that the best performing model is the XGB without PCA. We can easily see that it has the highest value in four different metrics, (AUC, accuracy, specificity, F1 score for negatives). The second-best performing model is also the XGBoost with PCA. In this way it is indicated that the XGBoost is a great fit for this kind of problem. Of course, it’s not just the number of the highest value metric but also what this metric represents. XGB using PCA in feature selection, performs exceptionally well in sensitivity-recall (0.92) that is a very important metric to take into consideration. As it was mentioned before, sensitivity(recall) is a really important metric because it indicates the true positives and false positives that are predicted. When a patient get negative results he does not proceed to further tests, that’s why the false negatives should be as low as possible because the patients who have falsely been diagnosed with negative results will not take extra tests and that will have catastrophic consequences regarding his health.

**3.2 Comparison table**

**3.2.1 Comparison table of the procedure followed.**

Throughout our research in scientific articles about predictive models for cervical cancer we came across a few that used the same dataset in order to train their model. Therefore it is very interesting to proceed with the comparison of our models with theirs, not only when it comes to the results but also about the selected features, the methodology that has been followed etc. In the Table 9 we can find the similarities / differences of out model compared to these scientific articles [5]–[7].

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Algorithms Used | Features Selected | Target variable | Handling of imbalance Dataset | Performance Metrics | Use of Cross-Validation |
| Our model | 1. Random Forest 2. XGBoost | 1. Feature selection through correlation heatmap 2. Principal Component Analysis technique (PCA) | A combination of the ‘Biopsy’, ’Hinselmann’,  ’Schiller’,  ’Cytology’ variables | SMOTE-Tomek method | 1. AUC 2. Accuracy 3. Sensitivity 4. Specificity 5. F1-score(0) (positives) 6. F1-score(1) (negatives) 7. F-beta | YES  (Five- fold cross validation) |
| “Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods”[5] | 1. Support vector machine (SVM) 2. XGBoost 3. Random Forest | All features from the Dataset | Diagnosis results of   1. Hinselmann 2. Schiller 3. Cytology 4. Biopsy | Borderline-SMOTE | 1. Accuracy 2. Sensitivity 3. Specificity 4. F1-score(0) (positives) 5. F1-score(1) (negatives) | NO |
| “A Model for Predicting Cervical Cancer Using Machine Learning Algorithms”[6] | 1. **Decision tree classifier (DTC)** 2. **Random Forest (Rf)** 3. **Adaptive Boosting (AB)** 4. **Support vector machine (SVM)** 5. **Radial Basis Function (RBF) Kernel Support Vector Machine (SVM)** 6. **Gradient Boosting (GB)** 7. Logistic regression (LR) 8. XGBoost 9. K-nearest neighbor (KNN) | * Principal Component Analysis technique (PCA) | Target variable – no further information was given | No further information was given | 1. Precision 2. Recall/Sensitivity 3. F1-score 4. Accuracy   (The above metrics have been calculated for both positives and negatives)   1. Mean Squared Error (MSE) 2. Mean Absolute Error (MAE) 3. Root Mean Squared Error (RMSE) 4. R-squared (R2) | YES  (Five- fold cross validation) |
| “Cervical Cancer Identification with Synthetic Minority Oversampling  Technique and PCA Analysis using Random Forest Classifier” [7] | Random Forest (RF) | * Principle Component Analysis (PCA) * Recursive Feature Elimination (RFE) | Diagnosis results of   1. Hinselmann 2. Schiller 3. Cytology 4. Biopsy | SMOTE technique | 1. Accuracy 2. Sensitivity 3. Specificity 4. Positive Predicted Accuracy (PPA) 5. Negative Predicted Accuracy | YES  (10- fold cross Validation) |

Table Comparison table of the used models in different studies

As it can be extracted from Table 9, we can find many similarities in the way other researchers proceed with the handling of the Cervical Cancer - Risk Factors dataset. The differences can be found with bold font in the above-mentioned table. In general terms the same or similar procedures were followed for the handling of imbalanced data, the feature selection process, the performance metrics which were used etc. The main difference in those models was regarding the choice of the target variable because in contrast with our model, the others keep all four target variables. Each one of these studies chose how deeply they wanted to proceed with their research and act accordingly.

**3.2.1 Comparison table of the results**

Our next step is to proceed to compare the performances of the different models in the different studies. First of all, in order to be able to compare the results we need to make sure that we compare similar things. For this reason in the table below, only the results of the mutual performance metrics will be displayed. Also, because no other studies have combined the four target variables into one target like us, we will compare our results for each one of the four different target variables. Finally we will have two separate tables for the results of Random Forest model and XGBoost. At this point it is worth mentioning that the “Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier” article does not involve with the XGBoost model and the “A Model for Predicting Cervical Cancer Using Machine Learning Algorithms” does not represent their results because the level of accuracy was not satisfying.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Accuracy | Sensitivity | Specificity | F1-score(0) (positives) | F1-score(1) (negatives) |
| Our model’s target variable | Without PCA | 0.79 | 0.73 | 0.86 | 0.77 | 0.81 |
| With PCA | 0.76 | 0.84 | 0.69 | 0.79 | 0.75 |
| “Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods”[5] | Hinselmann | 0.97 | 0.95 | 0.99 | 0.98 | 0.97 |
| Schiller | 0.95 | 0.94 | 0.96 | 0.95 | 0.95 |
| Cytology | 0.96 | 0.94 | 0.99 | 0.97 | - |
| Biopsy | 0.97 | 0.96 | 0.99 | 0.97 | 0.97 |
| “A Model for Predicting Cervical Cancer Using Machine Learning Algorithms”[6] | Target variable – prediction of cervical cancer | 1 | 1 | - | 1 | 1 |
| “Cervical Cancer Identification with Synthetic Minority Oversampling  Technique and PCA Analysis using Random Forest Classifier”[7] | Hinselmann | 0.96 | 0.97 | 1 | - | - |
| Schiller | 0.96 | 0.97 | 1 | - | - |
| Cytology | 0.96 | 0.97 | 1 | - | - |
| Biopsy | 0.96 | 0.97 | 1 | - | - |

Table Comparative table of the results using the Random Forest model.

Random Forest:

XGBoost:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Accuracy | Sensitivity | Specificity | F1-score(0) (positives) | F1-score(1) (negatives) |
| Our model’s target variable | Without PCA | 0.90 | 0.88 | 0.93 | 0.88 | 0.92 |
| With PCA | 0.88 | 0.92 | 0.83 | 0.89 | 0.88 |
| “Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods” | Hinselmann | 0.96 | 0.94 | 0.98 | 0.98 | 0.94 |
| Schiller | 0.96 | 0.94 | 0.97 | 0.96 | 0.96 |
| Cytology | 0.96 | 0.98 | 0.94 | 0.96 | 0.96 |
| Biopsy | 0.96 | 0.94 | 0.97 | 0.96 | 0.96 |
| “A Model for Predicting Cervical Cancer Using Machine Learning Algorithms” | Target variable – prediction of cervical cancer | No specific representation of the results was included in this scientific article because not satisfactory level of accuracy was achieved | | | | |

Table Comparative table of the results using the XGBoost model.

When it comes to the Random Forest results we can see that our results are not as ideal as the other studies. Their values fluctuate between 0.69 and 0.86 instead of 0.95 -1 like in the other studies. Also the study [6] reaches the ultimate value that is 1, that shows that their developed model responds 100% to the requested goal.

Although our results are not optimal, they are still acceptable.

On the other hand, our XGBoost model gives better results than Random Forest. Even though the [5] performs better in each one of the performance metrics, our values results are still in a very satisfying level and they indicate a well performing model.

**3.4 Limitations**

Throughout the creation of our model we faced obstacles that we needed to overcome. First of all, we had an imbalanced dataset that was hard to deal with. So, we used a combination of the four tests (Hinselmann, Schiller, Cytology, Biopsy) for our target variable. So, if any of these variables was positive, our target variable is positive, too. This technique has proven effective in addressing the issue of dataset imbalance, resulting in a significant improvement in the dataset's balance compared to its previous state. Also, we had to implement several data-processing techniques like SMOTE Tomek to get the desirable results.

Another limitation we faced was the lack of bibliography using the same procedures as us. To be more specific, when it comes to the target variable we did not have any references in the creation of one mutual target variable that was combining all four. In addition to this, most of the relevant articles were using only the biopsy variable as the target variable. Therefore, as students with no medical background we weren’t completely sure if the steps we were taking would lead us to the right direction. For that reason, when we performed feature selection with the most relevant features, the results were not the desired. And then we proceed to repeat the procedure with the PCA technique.

**4. References:**

[1] S. J. Rigatti, ‘Random Forest’, *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, Jan. 2017, doi: 10.17849/insm-47-01-31-39.1.

[2] W. Dong, Y. Huang, B. Lehane, and G. Ma, ‘XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring’, *Automation in Construction*, vol. 114, p. 103155, Jun. 2020, doi: 10.1016/j.autcon.2020.103155.

[3] Ž. Ð. Vujovic, ‘Classification Model Evaluation Metrics’, *IJACSA*, vol. 12, no. 6, 2021, doi: 10.14569/IJACSA.2021.0120670.

[4] ‘A Gentle Introduction to the Fbeta-Measure for Machine Learning - MachineLearningMastery.com’. https://machinelearningmastery.com/fbeta-measure-for-machine-learning/ (accessed Mar. 25, 2023).

[5] X. Deng, Y. Luo, and C. Wang, ‘Analysis of Risk Factors for Cervical Cancer Based on Machine Learning Methods’, in *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, Nanjing, China, Nov. 2018, pp. 631–635. doi: 10.1109/CCIS.2018.8691126.

[6] N. Al Mudawi and A. Alazeb, ‘A Model for Predicting Cervical Cancer Using Machine Learning Algorithms’, *Sensors*, vol. 22, no. 11, p. 4132, May 2022, doi: 10.3390/s22114132.

[7] R. Geetha, S. Sivasubramanian, M. Kaliappan, S. Vimal, and S. Annamalai, ‘Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier’, *J Med Syst*, vol. 43, no. 9, p. 286, Sep. 2019, doi: 10.1007/s10916-019-1402-6.