Predicting Cervical Cancer Cases Resulting in Biopsies Using Machine Learning Techniques

Tenali Pranuthi

11714610

KM032B62

Lovely Professional University

Abstract:

There are various algorithms and methodologies used for automated screening of cervical cancer by segmenting and classifying cervical cancer cells into different categories. This study presents a critical review of different research papers published that integrated ML methods in screening cervical cancer via different approaches analyzed in terms of typical metrics like dataset size, drawbacks, accuracy etc. An attempt has been made to furnish the reader with an insight of Machine Learning algorithms like SVM (Support Vector Machines), k-NN (k-Nearest Neighbors), RFT (Random Forest Trees), for feature extraction and classification. This paper also covers the publicly available datasets related to cervical cancer. It presents a holistic review on the computational methods that have evolved over the period of time, in detection of malignant cells.

In this paper, we are going to train our model using various machine learning techniques and all the models thus made are compared in terms of accuracy, precision and recall.

Keywords:

Ensembling, Supervised learning, unsupervised learning, Random Forest, Decision trees, cancer, biopsy.

Introduction:

Cervical cancer is a malignant tumour starting in the cells of a woman’s cervix, and possibly spreading or metastasizing to other parts of her body. The cervix is part of a woman’s reproductive system, located below the uterus. In most cervical cancer cases, the tumours develop from precancerous changes in the cervix, and can take several years to develop.

About 13,800 new cases of invasive cervical cancer will be diagnosed. About 4,290 women will die from cervical cancer. In the detection of cervical cancer, machine learning techniques have been of much help contributing to the medical stream.

In the paper titled New Features of Cervical Cells for Cervical Cancer Diagnostic System Using Neural Network by Mustafa et al [3], it has been stated that though Pap test is the most popular and effective test for cervical cancer, Pap test does not always produce good diagnostic performance. In the paper Preprocessing for Automating Early Detection of Cervical Cancer, by Debasis Bhattacharyya et al.[4] states that In Cervigram, cervix region occupies about half of the raw cervigram image. Other parts of the image contain inconsequential information. This irrelevant information can muddle automatic identification of the tissues within the cervix. Asselin et al.[5] discuss the imaging methods available to provide appropriate biomarkers of tumor structure and function using selective regions of interest (ROI), Cluster analysis and Histogram analysis. Turid Torheim et.al[6] present the paper with texture analysis methods and classification by using SVM to identify the cured and relapsed images. S. Jagadeeswari and S. Malarkhodi[7] presented a paper on classification by using an Artificial Neural Network to identify the normal and abnormal tumor images with Fourier transform and Gaussian low pass filter. Rupinderpal and Rajneet presented a noise removal method using discrete wavelet transform. Here, we have tried to detect the cancer using ML techniques along with ensembling.

Literature Review:

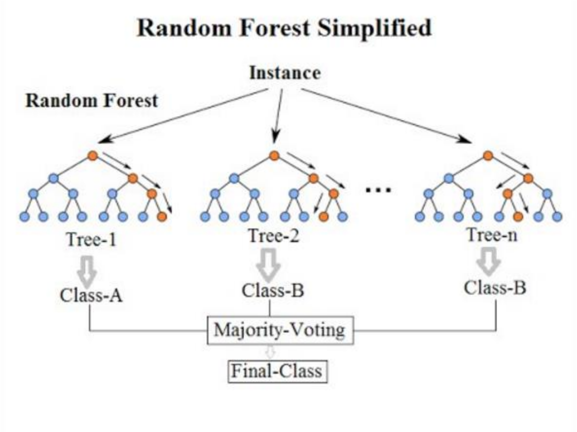
Supervised Learning Techniques:

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

The supervised machine learning techniques that we used here are Logistic Regression, Random Forest Classifier, SVM.

1. Random Forest:

Random Forest is an ensemble algorithm which creates many decision trees (a forest), and applies them to multiple subsets of the dataset, creating multiple classification results. The Random Forest Classifier uses a voting system to make its final classification prediction, with each tree voting, and chooses the class with the most votes. An alternative voting measure is using weights to assign the impact of a decision tree’s result, with trees with high errors getting low weightings, and vice versa. In this voting system, trees with low error rates have a higher impact on the final classification decision.



The Random Forest Classifier splits the dataset into a training set and testing set by sampling with replacement, until approximately one-third of the data is remaining, which is used for testing the classifier. Before applying the classifier to the data, you must determine how many trees each forest should contain, and the minimum number of nodes required in order for the tree to split. Advantages:

1.It’s works well with noisy data and it reduces overfitting. Since the end result is an average or majority vote of multiple classification results, the classifier has a significantly lower chance of overfitting the data.

2.Since there are multiple forests, not every forest is necessarily affected by noisy data.

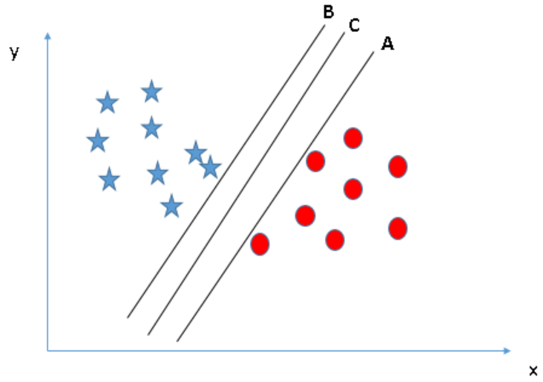
Disadvantages:

1.They are much more complex than normal decision trees, thus are harder to understand and visualize.

2.Because there are many more trees being created and used than in a normal decision tree, it is more computationally expensive.

2. Support Vector Machine (SVM):

A Support Vector Machine is a classification technique which attempts to separate the different classes of data by finding a decision boundary which maximizes the margin. The SVM represents this boundary by using support vectors. Although there are infinitely many different hyperplanes that could be selected to separate the data, the hyperplane with the largest margin often performs better, as it leaves more room for any perturbations to the decision boundary without having an impact on the classification.



Advantages:

1.They work well with high-dimensional data, avoid the curse of dimensionality, and they still work well in cases where there are more dimensions than there are samples of data.

Disadvantage:

1.They are harder to analyze, as they do not give out a probability score.

3. Logistic Regression:

This is a learning technique employed when the output of training data is in the form of groups called classes.

In our model, the output is 0 if the patient is negative with cancer and 1 if she is diagnosed positive.

Advantages of Logistic Regression:

1. Logistic Regression performs well when the dataset is linearly separable.

2. Logistic regression is less prone to over-fitting but it can overfit in high dimensional datasets. You should consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios.

Disadvantages:

1. Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. Most of the time data would be a jumbled mess.

2. Logistic Regression can only be used to predict discrete functions. Therefore, the dependent variable of Logistic Regression is restricted to the discrete number set. This restriction itself is problematic, as it is prohibitive to the prediction of continuous data.

The unsupervised technique used is K means clustering.

K-Means Clustering:

It is an unsupervised approach to classifying data which tries to make clusters of similar data. Each data point is compared to randomly selected centroids, and placed in the neighborhood of its nearest cluster (using Euclidean distance). The number of clusters must be defined at the beginning of the model. After selecting the initial centroids, the distances are computed and the data is assigned to centroid, and the centroids are recomputed multiple times until they don’t move around anymore.

K-means clustering performs well and is easy to understand the visualization of the data, However, k-means clustering doesn’t perform well when the data is of different sizes or densities, and has problems when the data contains outliers.

Advantages:

1. Relatively simple to implement.

2. Scales to large data sets.

3. Guarantees convergence.

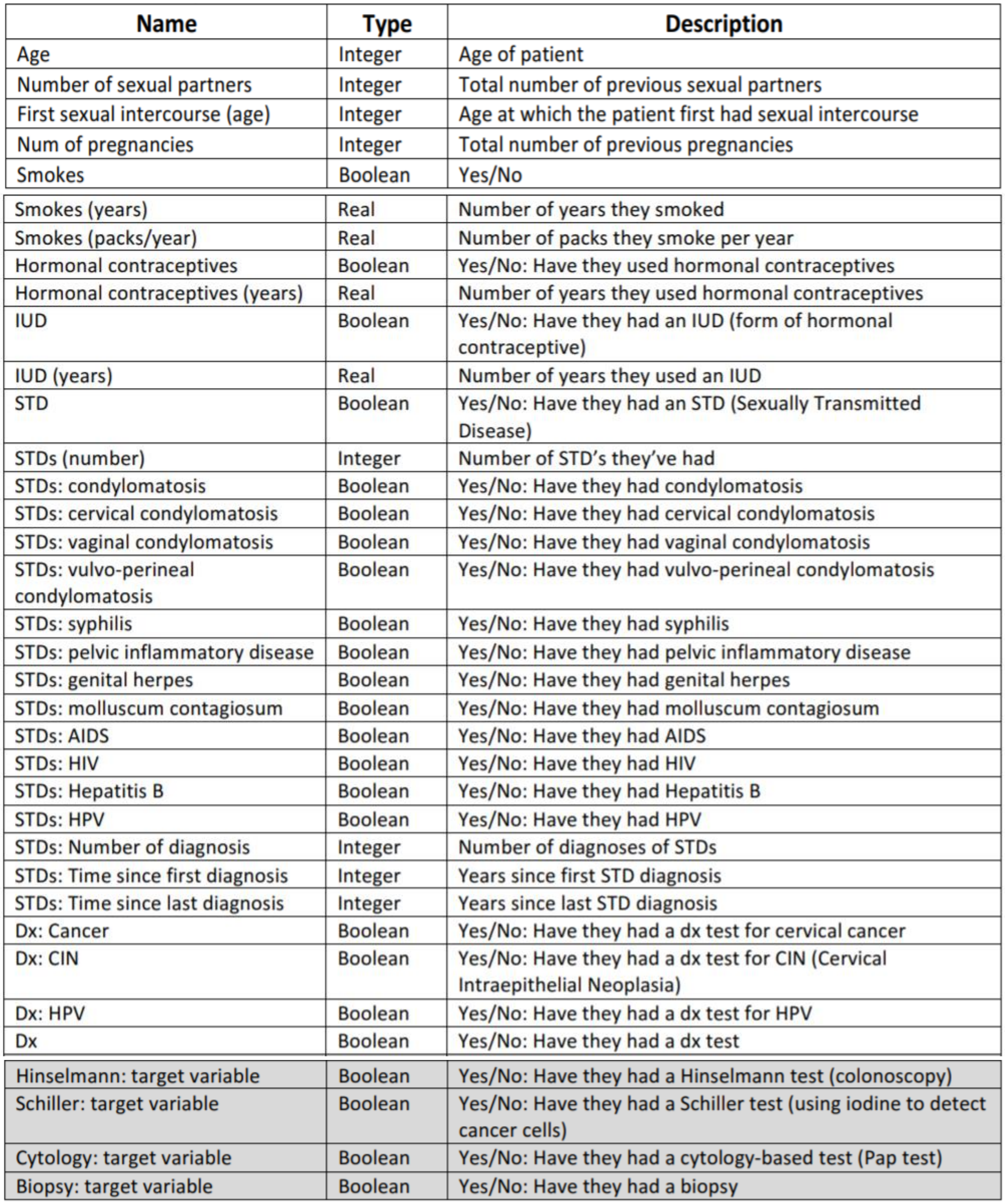
Disadvantages:

1. Choosing k manually.

2. Being dependent on initial values.

About the dataset:

There are 36 attributes in the dataset, consisting of 32 risk factors, and 4 target variables (the last four attributes):

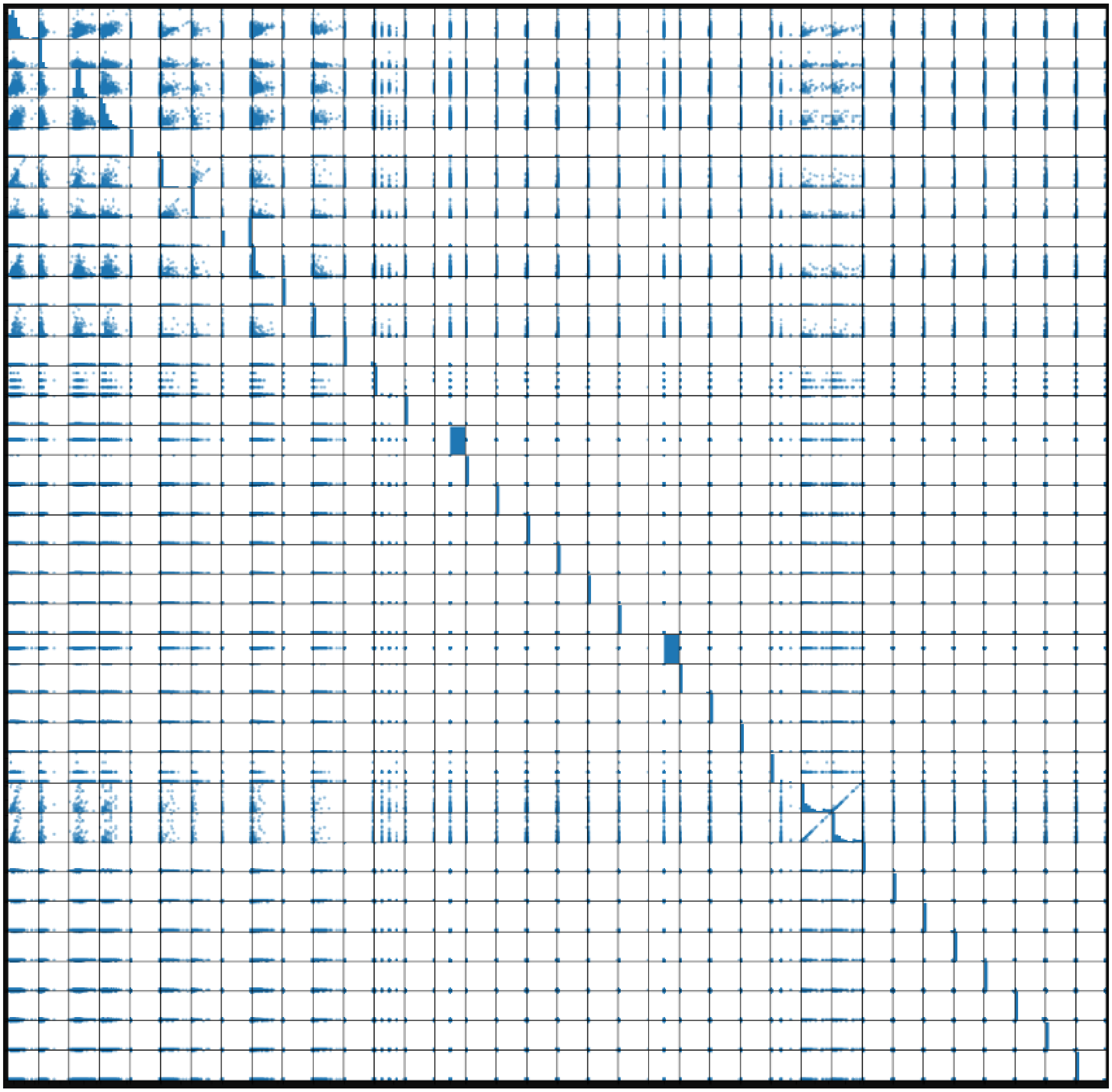


Name	Type	Description
Age	Integer	Age of patient
Number of sexual partners	Integer	Total number of previous sexual partners
First sexual intercourse (age)	Integer	Age at which the patient first had sexual intercourse
Num of pregnancies	Integer	Total number of previous pregnancies
Smokes	Boolean	Yes/No
Smokes (years)	Real	Number of years they smoked
Smokes (packs/year)	Real	Number of packs they smoke per year
Hormonal contraceptives	Boolean	Yes/No: Have they used hormonal contraceptives
Hormonal contraceptives (years)	Real	Number of years they used hormonal contraceptives
IUD	Boolean	Yes/No: Have they had an IUD (form of hormonal contraceptive)
IUD (years)	Real	Number of years they used an IUD
STD	Boolean	Yes/No: Have they had an STD (Sexually Transmitted Disease)
STDs (number)	Integer	Number of STD's they've had
STDs: condylomatosis	Boolean	Yes/No: Have they had condylomatosis
STDs: cervical condylomatosis	Boolean	Yes/No: Have they had cervical condylomatosis
STDs: vaginal condylomatosis	Boolean	Yes/No: Have they had vaginal condylomatosis
STDs: vulvo-perineal condylomatosis	Boolean	Yes/No: Have they had vulvo-perineal condylomatosis
STDs: syphilis	Boolean	Yes/No: Have they had syphilis
STDs: pelvic inflammatory disease	Boolean	Yes/No: Have they had pelvic inflammatory disease
STDs: genital herpes	Boolean	Yes/No: Have they had genital herpes
STDs: molluscum contagiosum	Boolean	Yes/No: Have they had molluscum contagiosum
STDs: AIDS	Boolean	Yes/No: Have they had AIDS
STDs: HIV	Boolean	Yes/No: Have they had HIV
STDs: Hepatitis B	Boolean	Yes/No: Have they had Hepatitis B
STDs: HPV	Boolean	Yes/No: Have they had HPV
STDs: Number of diagnosis	Integer	Number of diagnoses of STDs
STDs: Time since first diagnosis	Integer	Years since first STD diagnosis
STDs: Time since last diagnosis	Integer	Years since last STD diagnosis
Dx: Cancer	Boolean	Yes/No: Have they had a dx test for cervical cancer
Dx: CIN	Boolean	Yes/No: Have they had a dx test for CIN (Cervical Intraepithelial Neoplasia)
Dx: HPV	Boolean	Yes/No: Have they had a dx test for HPV
Dx	Boolean	Yes/No: Have they had a dx test
Hinselmann: target variable	Boolean	Yes/No: Have they had a Hinselmann test (colonoscopy)
Schiller: target variable	Boolean	Yes/No: Have they had a Schiller test (using iodine to detect cancer cells)
Cytology: target variable	Boolean	Yes/No: Have they had a cytology-based test (Pap test)
Biopsy: target variable	Boolean	Yes/No: Have they had a biopsy

The dataset, "Cervical Cancer Risk Factors for Biopsy" was obtained from the UCI Repository. The dataset contains habits, demographic information, and medical history of 858 patients from the hospital.

Proposed Methodology:

Visualization of data:



The above is the scattered matrix plot obtained on plotting the features of the dataset as a plot. The scatter matrix is advantageous in revealing strong correlations between specific attributes, or showing that two attributes don't have any sort of correlation. In looking at the "Biopsy" column (the furthest right), I could see many scatterplots that didn't have strong correlations, leading to the fact that there are some potential attributes that are redundant or irrelevant and could be removed from training.

Data Preprocessing:

Imputation:

Because of the large number of missing values in this dataset, I decided to replace the missing values instead of eliminating them. This way I could be sure not to lose any important data, and try to make the most accurate estimate of the missing data.

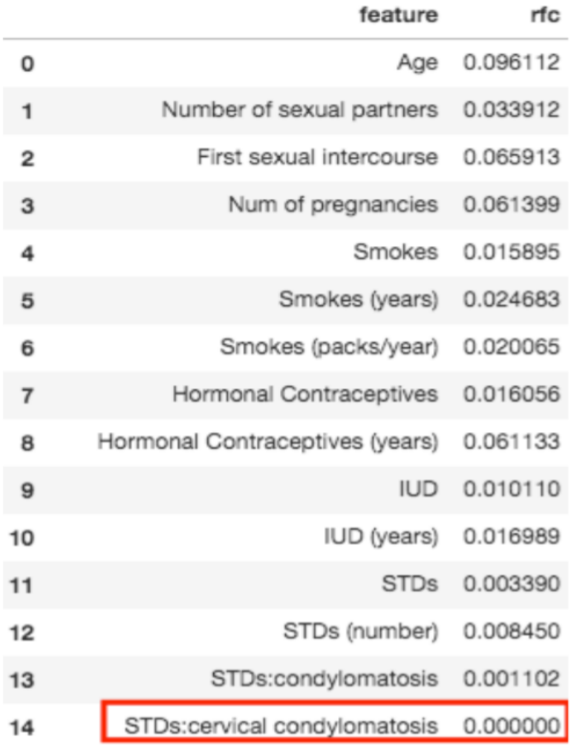
The numerical values are replaced with the median of that particular column and the categorical values are replaced with the mode value.

Standardization & Normalization:

Due to the fact that some of the attributes in the dataset are binary values and some are not, standardizing and normalizing the dataset is advantageous in evening out the values and helping to solve imbalance problems. Using a built-in function for standardizing/normalizing data in scikit-learn, the dataset now is standardized.

Also there are many attributes in the dataset which has weak correlations with the target and those are removed.

If two columns or attributes that are non-target values are having strong correlation, one of the attribute can be removed in order to reduce the dimensional complexity. Here Random Forest Classifier or PCA can be used to do dimensional reduction.



	feature	rfc
0	Age	0.096112
1	Number of sexual partners	0.033912
2	First sexual intercourse	0.065913
3	Num of pregnancies	0.061399
4	Smokes	0.015895
5	Smokes (years)	0.024683
6	Smokes (packs/year)	0.020065
7	Hormonal Contraceptives	0.016056
8	Hormonal Contraceptives (years)	0.061133
9	IUD	0.010110
10	IUD (years)	0.016989
11	STDs	0.003390
12	STDs (number)	0.008450
13	STDs:condylomatosis	0.001102
14	STDs:cervical condylomatosis	0.000000

Now as the data is ready to use to train the models, we are all set to proceed with training the models.

Decision Tree classifier:
The criterion used here is gini index with parameters set to max_leaf_nodes=None, min_samples_leaf=14, min_samples_split=5, random_state = 1

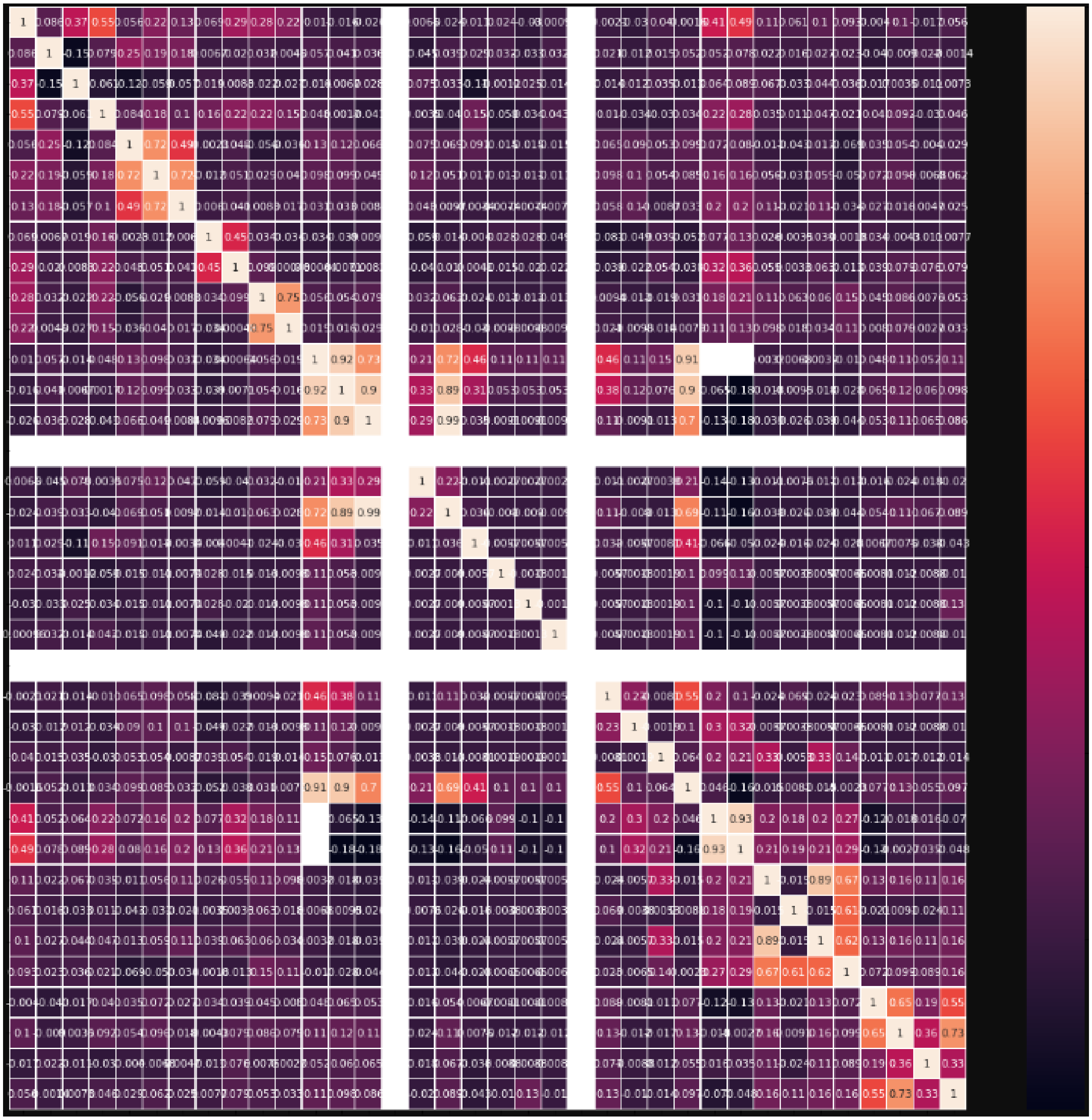
Logistic Regression:
The parameters are set to C=4, penalty='l2'.
And also set to C=4, penalty='l1' in another model.

KNN:
The parameters are set to n_neighbors= 5, p = 2, metric = 'minkowski'.

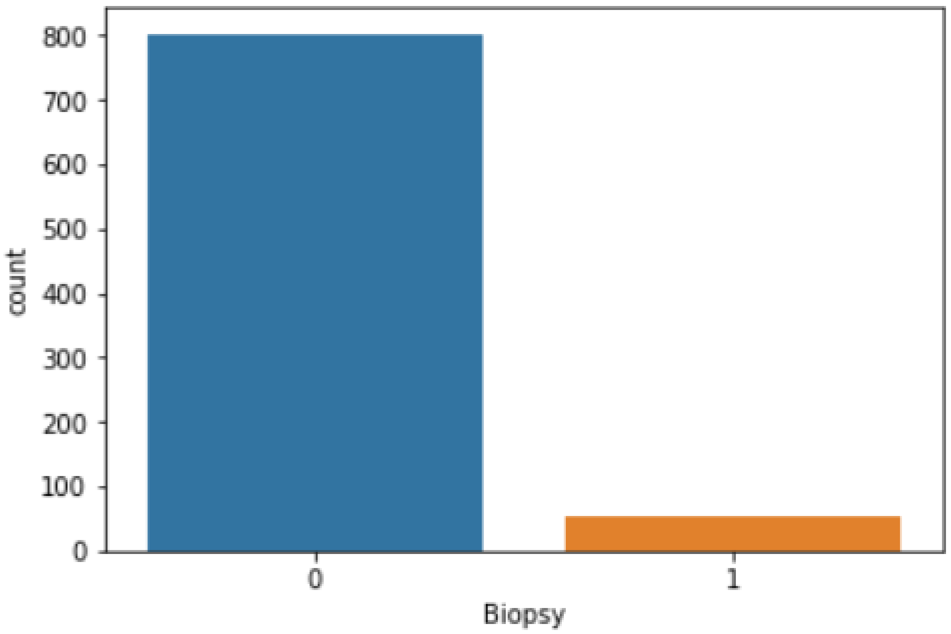
Random Forest Classifier:
The parameters are set to max_depth=5, n_estimators=10, max_features=1, n_jobs=10

SVM:
The kernel is set to "rbf" with C=100, gamma=15

Results and Discussion:



Correlation matrix of the features between datasets



True positives in the data

Comparison of different models used in detecting cervical cancer:

Model	Accuracy	Class	Precision	Recall	F1 score	Support
Decision Tree	95.38%	0	0.97	1.0	0.97	242
		1	1.0	0.52	0.69	16
Logistic Regression	94.5%	0	0.98	0.98	0.96	242
		1	0.73	0.52	0.61	16
Knn	93.75%	0	0.94	1.00	0.97	242
		1	0.01	0.00	0.03	16
K means	76.27%	0	0.93	0.23	0.37	242
		1	0.02	0.06	0.03	16
Random Forest	93.7%	0	0.94	1.00	0.97	242
		1	0	0	0.2	16
SVM	91%	0	0.94	0.96	0.95	242
		1	0.1	0.06	0.08	16
Perceptron	92.63%	0	0.94	0.98	0.96	242
		1	0.2	0.06	0.1	16

As far as the models are all trained and tested, the decision tree model has got the highest accuracy when compared to all the other models.

As the decision trees can avoid overfitting, we don’t see the problem of overfitting in our model.

Among the very first works in cervical cancer detection stands the work by P. Mitra in the year 2000 where staging was done by amalgamation of ID3 and GAs, where GAs were mainly used for refining the architecture. In the work by J. Zhang et al in 2004 the use of SVMs was more desirable because they had small data set of 40 images containing 149 cells. Back in that time medical imaging data set was meagre and SVMs provide accurate result(clearer decision boundary) on small size batch. In paper by R. Vidya and G. M. Nasira [9], CART algorithm was initially implemented to check the feasibility of the task. On realizing pretty successful results, advanced algorithms RFT and RFT with k-NN were implemented leading up to an accuracy of 94.77%.

The Decision Tree model that has been trained here produced an accuracy of 95.3%. So, the decision tree model is said to be highly useful in case of cervical cancer detection using the cervical cancer risk classification dataset.

Conclusion:

In conjunction with more accurate diagnostics, AI has the potential to bring down the cost of unwanted interventions for cervical cancer screening. Early detection will promise a greater rate of patients prognosis especially in case of non-invasive cancer. The papers discussed above made use of independent data sources, consequently a base for comparing algorithms on a single scale was hard to define. Multiple algorithms have been applied for segmenting cell cytoplasm, nuclei and other cell components and classifying cells into different categories. In the view of the known stages of cancer, the accuracy of each algorithm, Decision Tree has proved to yield highest accuracy for classifying cervical cancer risk.

Although it can be difficult to compare the results of the supervised models with the unsupervised model, in this case, it is clear that the Decision Tree Classifier performed better than the other three models. Not only did it have a higher overall accuracy than the Random Forest and Support Vector Machine classifiers, but it also had the highest precision, accuracy, and F1 measures, signifying that it was better at correctly classifying the positive class. In comparing this to the unsupervised K-means clustering model, it’s hard to find an exact difference in performance between the two, but the Gradient Boosting Classifier was able to clearly classify a large amount of the test set correctly, whereas the K-means cluster model struggled to separate the y values from the X values.

References:

[1] S. Sharma, “Cervical Cancer stage prediction using Decision Tree approach of Machine Learning”, International Journal of Advanced Research in Computer and Communication Engineering vol. 5, Issue 4, 2016.

[2] J. Zhang and Y. Liu, “Cervical Cancer Detection Using SVM Based Feature Screening”, 2004.

[3] P. K. Malli, S. 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), 2017

[4] Cervical cancer (risk factor) data set UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29

[5] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html

[6] Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. 'Transfer Learning with Partial Observability Applied to Cervical Cancer Screening.' Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017.

[7] NCCC. Cervical cancer. 2010.

[8] Rama Praba PS, Ranganathan H. Comparing different classifiers for automatic lesion detection in cervix

[9] M. Nunez, "Decision Tree Induction Using Domain Knowledge" in Current Trends in Knowledge Acquisition, Amsterdam:IOS Press, 1990.

[10] J. Kern, G. Dezelic, M. Tezak-Bencic, T. Durrigl, "Medical Decision Making Using Inductive Learning Program", *Proceedings of 1st Congress on Yougoslav Medical Informatics*, pp. 221-228, Dec 6-8, 1990.

[11] "Fact sheet No. 297: Cancer", *World Health Organization*, 01 2007.