Abnormality Detection and Severity Classification of Cells based on Features Extracted From Papanicolaou Smear Images using Machine Learning

Abhinaav R
Student
Department of Biomedical Engineeringing
PSG College of Technology
Coimbatore, India
contactabhinaav@gmail.com

Dr. D. Brindha
Assistant Professor (Sl. Gr)
Department of Biomedical Engineeringing
PSG College of Technology
Coimbatore, India
brn.bme@psgtech.ac.in

Abstract— A Papanicolaou Smear (PAP) test is a screening method developed for cervical cancer that involves the microscopic examination of cervical cells carefully extracted and spread out as a smear and stained specially. A Pap test reveals premalignant and malignant changes and the changes that are due to non-carcinogenic conditions like inflammation. The diagnosis of this test are based upon key features of the nucleus and cytoplasm of the affected cell or the cell under observation. This work is aimed at devising a classification algorithm using supervised methods to efficiently classify the affected cells from normal cells and further group the affected cells Logistic Regression. [9] All algorithms and models were trained and validated using the Azure Machine Learning Studio.

Keywords—PAP Smear, Classification Algorithm, cervical cancer, supervised data analysis, Azure ML

I. INTRODUCTION

This work is concentrated towards developing an efficient classifier to identify an affected or abnormal cell and identify the severity if affected. The proposed algorithm is designed for the numeric data extracted from the set of images after initial processing techniques. For training and validating the algorithm designed, all the data used in this work is taken from the Pap-smear Benchmark Data for Pattern Classification [1] [11]. The dataset consists of 917 smear images, classified by specialist cyto-technicians and doctors. Each cell is characterized by 20 numeric data features and the cells fall into 7 classes [1]. The algorithm used to classify the data is explained below along with graphical illustrations and explanatory data. This benchmark data provides the necessary standardized data for effectively diagnosing the accuracy if the proposed classifier. The algorithm devised makes use of all available data in the dataset without any exceptions, thus improving the liability of it.

II. PAP SMEAR TEST

The importance of the PAP smear test is that it is a simple procedure that can be used to find the cervical cancer and also the stage of cancer. The test is performed by collecting cells from the cervix and then making a thin smear over a glass slide. This smear is stained using different chemicals and the results are observed under a microscope. The pap smear data base consists of 917 such microscopy images collected for special cases from different locations.

III. PREVIOUS WORKS

In the process of literature survey, we identified the works of Jonas Norup as a standardized work with several citations. Hence most part of our research were driven towards improving the accuracy of classification in comparison to the works done by Jonas Norup. The works of his research is devising a classifier network and separating the data using clustering as a whole. His clustering model classified the entre data into 7 clusters. He used both inductive classifier (LMS Algorithm) and transductive classifier (KNN and WKNN [10], Nearest Class Gravity Center NCC [6], Neuro Fuzzy Interference Method for Transductive Reasoning NFI [7], Evolving Clustering Method) and compared the results using each type of classifier. Of all the classifiers, the Nearest Class Gravity center method provided the best results at an error of 5.13% upon classification of all cells. [2]

We also Identified the works of K. Hemalatha and K. U. Rani, Presented in the 2017 IEEE 7th International Advance Computing Conference (IACC), Hyderabad, 2017. They were also an eye-opener and a vast motivation in gathering more and more information in this respect. [11]

IV. FEATURE PROCESSING AND CLASSFICATION

A. Extracted Features present in the Dataset:

The different characteristics of different classes of PAP Smear Cells [2] in Figure:1 shown below.

- Normal - 242 cells :
  - Class1 Superficial squamous epithelial, 74 cells.
  - Class2 Intermediate squamous epithelial, 70 cells
  - Class3 Columnar epithelial, 98 cells.
- Abnormal - 675 cells :
  - Class4 Mild squamous non-keratinizing dysplasia, 182 cells.
  - Class5 Moderate squamous non-keratinizing dysplasia, 146 cells.
  - Class6 Severe squamous non-keratinizing dysplasia, 197 cells.
  - Class7 Squamous cell carcinoma in situ intermediate, 150 cells.

978-1-5386-8260-9/19/$31.00 ©2019 IEEE

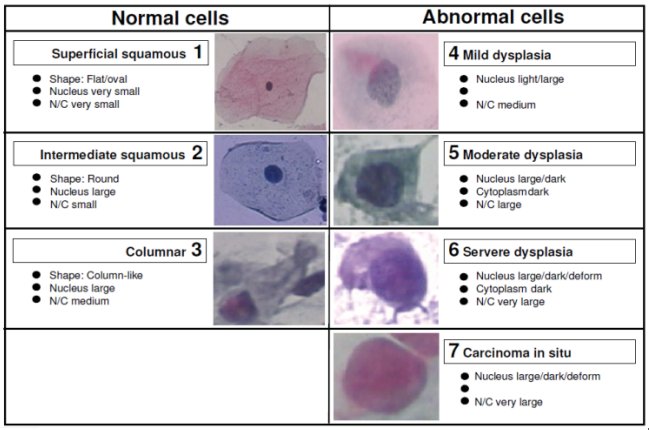

Figure1: Cell type characteristics for single pap-smear cells.

Listed below are the extracted features [2] for pap-smear data. N and C are abbreviation of the Nucleus and Cytoplasm. The letter in brackets refers to the column in the spreadsheet of the new pap-smear database. Column (A) in the spreadsheet is used for images reference.

1. N area(B)
2. C area(C)
3. N/C ratio(D)
4. N brightness(E)
5. C brightness(F)
6. N shortest diameter(G)
7. N longest diameter(H)
8. N elongation(I)
9. N roundness(J)
10. C shortest diameter(K)
11. C longest diameter(L)
12. C elongation(M)
13. C roundness(N)
14. N perimeter(O)
15. C perimeter(P)
16. N relative position(Q)
17. Maxima in N(R)
18. Minima in N(S)
19. Maxima in C(T)
20. Minima in C(U)

B. Design of a Classifier:

Since most of the parameters can be clubbed together like n/c ratio is the ratio of n area and c area. The dimension of the input vector is reduced from 20 to 10. Our design is centrally classified upon the features N/C Ratio, N Brightness, C Brightness, N Elongation, N Roundness, C Elongation, C Roundness, N perimeter, C perimeter, N relative Position(Q).

Initially the classification is done to separate the abnormal cells from normal cells. In this regard two types of classification techniques were used and the more accurate of the two techniques is used to separate the data and send it to the further clustering block to classify the abnormal cells. All processing, training and validation of data were don't using the Azure ML studio. The graphs and numeric data representative of results in this research paper is obtained on Azure ML Studio. The model is trained for randomized 70% of 917 samples and validated for remaining 30%.

The training of the classifier is done through two methods:

1. Two Class Boosted Decision Tree
2. Two Class Logistic Regression

1) Two class Boosted Decision Tree[3]:

This model in Azure ML is used to perform a binary classification algorithm based upon decision tree method. Initially the classification model is created without training for default parameters. I have used a single parameter model giving a specific set of parameters as inputs – the 10 features mentioned earlier.

1. **Maximum number of leaves per tree**: We have used 100 Leaves per tree for increased accuracy at the cost of more computations. For this PAP smear data, it was experimentally known that even if the number of leaves is increased more than 10 there is no change in the results, indicating that 100 leaves is the maximum count required for complete computation.
2. **Minimum number of samples per leaf node**: For increasing the computations per leaf, the minimum number was set to 3 rather than default 5, Indicating 2 * maximum number of leaves per tree computational increase in the model from default values.
3. **Learning rate** : 0.75 was chosen since it was the optimal value for faster convergence.
4. **Number of trees constructed**: (indicates the total number of decision trees to create in the ensemble) 100 trees were chosen since 100 epochs were enough to classify the data efficiently.
5. **Random number seed:** On mentioning the random seed value we can run the algorithm together for data having the same parameters. The random seed is set by default to 0.

Table:1: Final Settings and Values of Two Class Boosted Decision Tree Model in Azure Ml Ml for Classification of PAP Smear Dataset[3]

Setting	**Value**
Number Of Leaves	100
Minimum Leaf Instances	3
Learning Rate	0.75
Number Of Trees	100
Allow Unknown Levels	TRUE
Random Number Seed	0

2) Two Class Logistic Regression Model[3]:

This technique is a predictive algorithm. This two class model classification algorithm is optimized for dichotomous or binary variables. To train this model, we must provide a dataset that contains a label or class column.

In this respect, the classes mentioned in the dataset were included as a parameter- making all values of class < 3 as 0 and other values as 1, i.e., normal cells are assigned a value 0 and abnormal cells are assigned a value 1. This pre-formatted data is fed to the model for training. [5] The same 70% random split data used for two class boosted decision tree is used, for comparing the results of two classifiers. For improved performance a selective feedback was introduced as an input in addition to the input parameters.

Table:2: Final Settings of Logistic Two class Regression model in Azure Ml for Classification of PAP Smear Dataset[3]

Setting	Value
Optimization Tolerance	0.0000001
L1 Weight	1
L2 Weight	1
Memory Size	20
Quiet	TRUE
Use Threads	TRUE
Allow Unknown Levels	TRUE
Random Number Seed	0

3) *Output of Classifiers:*

The results from the two classifiers are compared and then the result which is more accurate is taken forward to the next section of clustering of abnormal cells. Depending upon the randomized data split, the accuracy of either model varies. So the model which predicts the normal and abnormality of PAP Smear Cells with better accuracy will be chosen by the Algorithm and it is passed upon as an input the clustering model. It was noted that when the classified data containing only the abnormal cells are passed on further to a new classification model, it is able to classify the abnormality in the cells into 4 classes with increased accuracy.The results of the classification models and the ROC curves are given below in Facts and Figures.

C. Design of Multi Class Logistic Regression Classifier[3]:

Since the abnormal cells have to be classified into 4 types of cells, Multi class logistic regression is used [8]. A single parameter model is used and the "class" is given as the label column in the classifier. The classifier is designed in Azure ML using the Multi-class Logistic Regression Classifier using the settings mentioned below in table: The classifier is trained for 60% data from the previous stage (abnormal – normal classifier results). The remaining 40% data is used to validate and score the trained model.

Table:3: Final Settings and Values of Multi Class Logistic Regression Classifier Model in Azure ML Classification of PAP Smear data classified as "Abnormal" [3]

Setting	Value
Optimization Tolerance	1.00E-15
L1 Weight	1
L2 Weight	1
Memory Size	200
Quiet	TRUE
Use Threads	TRUE
Allow Unknown Levels	TRUE
Random Number Seed	0

V. Tables and Figures

The proposed method was able to classify the PAP Smear cells into two classes Normal and Abnormal, with complete accuracy. The results obtained through both methods of classification for one randomized split is given below.

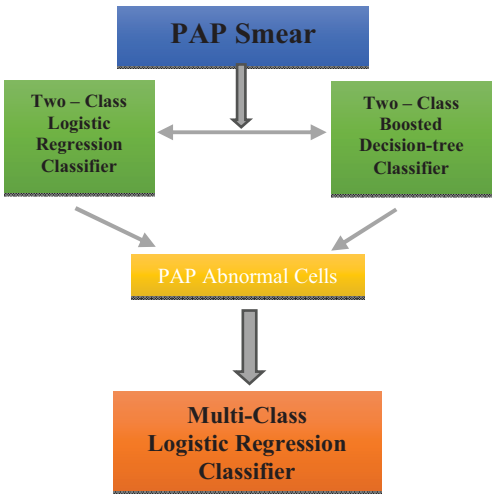

Fig:2: The block diagram representation of the entire experiment designed in AzureML.

1) *Two Class Boosted Decision Tree Model:*

Table:4: Summarized Data metrics of Evaluated Two class Boosted Decision Tree Model [3]

Accuracy	**0.901818**
Precision	0.931034
Recall	0.935644
F-Score	0.933333
AUC (Are under Roc Curve)	**0.954394**
Average Log Loss	0.256489
Training Log Loss	55.67734

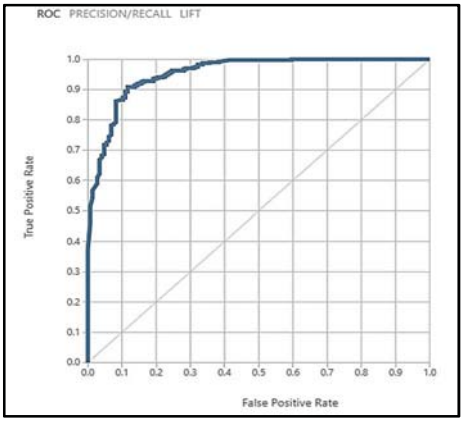

Fig:2: Receiver Operating Characteristics Graph depicting the relations between True Positive and False Positives of the two class Boosted Classification Model Designed to classify PAP Smear Cells in AzureML Studio [3]

2) *Two Class Logistic Regression Model:*

Table:5: Final weights for the features to classify the class of cells as normal or abnormal [3]

Feature	**Weight**
Abs (Class)	5.436
Kerne A	3.15763
KernePeri	2.8011
Bias	-2.54001
K/C	1.9037
CytoPeri	-0.726369
CytoElong	0.37822
CytoRund	-0.133906
Cyto_A	-0.118193
KerneRund	-0.0900856

Table:6: Summarized Data metrics of Evaluated Two class Logistic Regression Tree Model [3]

Accuracy	**1**
Precision	1
Recall	1
F-Score	1
AUC	**1**
Average Log Loss	0
Training Log Loss	0

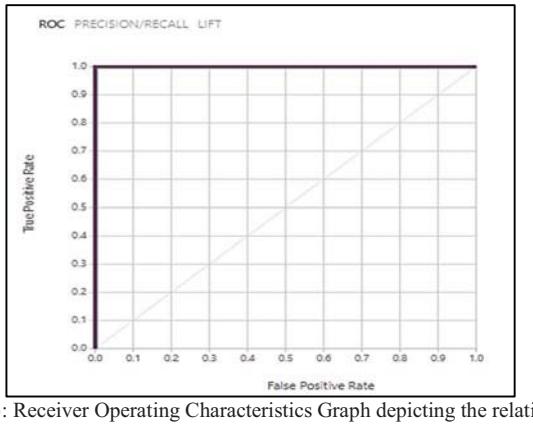

Fig:3: Receiver Operating Characteristics Graph depicting the relations between True Positive and False Positives of the two class Logistic Regression Model Designed to classify PAP Smear Cells in AzureML Studio [3]

3) *Comparison of two classifiers:*

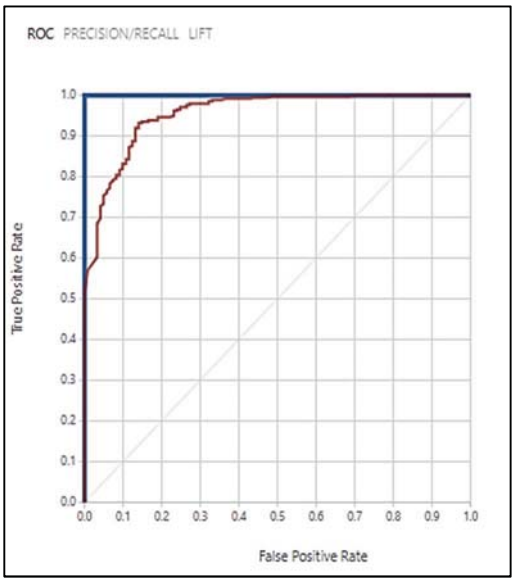

Fig:4: Receiver Operating Characteristics Graph depicting the comparison of the results of the above mentioned techniques. The thick blue line represents the Logistic Classification and the thin red line represents the Boosted Decision Tree. It is clear that the logistic classification has a better ROC and Area Under the Curve [3]

4) *Multi-Class Logisitic Regression Model:*

Table:7: Final weights for the features to classify the abnormal cells into 4 class is given above Multi-Class Logistic Regression Model _AzureML [3]

Feature	**Class 4**	**Class 5**	**Class 6**	**Class 7**
Abs(Class)	-7.85357	-2.24794	1.79112	8.31026
Bias	2.92023	2.13745	0.455197	-5.513
K/C	-0.826212	-0.44732	0	1.30518
CytoShort	0.567079	0.343982	-0.871195	-0.0397847
Cyto_Ycol	0	-0.81126	0.615056	0
CytoLong	0.778555	0	0	-0.359881
Cyto_A	0.734285	0	-0.0740938	0
CytoMin	0.4736	0	0	0
CytoMax	0.470458	0	0	0
KerneShort	0	0.35596	-0.098469	-0.0703439
KernePos	-0.169658	0.336819	0	0
CytoRund	0	0.25577	0	-0.0968374
KerneRund	0	0	0	-0.19101
CytoPeri	0.171117	0	-0.182778	0
Kerne_A	0	0.16613	0	0
KerneLong	0	0.089096	0	0
CytoElong	0	0	-0.075421	0
KerneMax	0	0	0	-0.0415099
Kerne_Ycol	0	0	0	0
KerneElong	0	0	0	0
KernePeri	0	0	0	0
KerneMin	0	0	0	0

Overall accuracy	***0.980296***
Average accuracy	0.990148
Micro-averaged precision	0.980296
Macro-averaged precision	0.983817
Micro-averaged recall	0.980296
Macro-averaged recall	0.976744

Table:8: Final metrics of the Multi Class Regression Model [3]

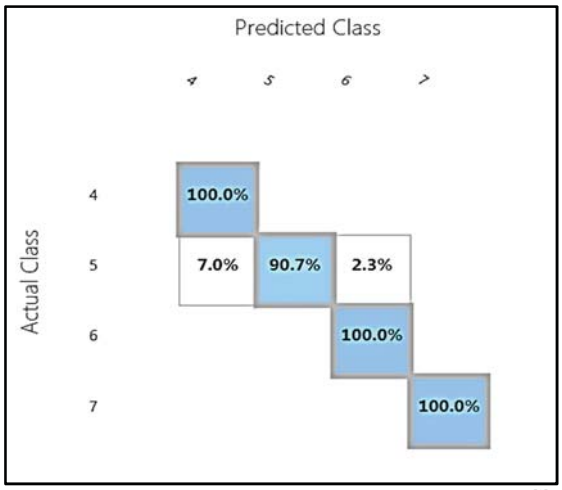

Fig:5: Confusion Matrix depicting the results of the classifier [3]

VI. Results

The entire objective of this work is to find a suitable algorithm which reduces the error in classification of PAP Smear cells. Our idea is to initially classify the cells as normal and abnormal and further classify the abnormal cells into 4 classes. We have achieved an accuracy of **98.03%** in classifying the abnormal cells. Since there is no error in the classification of data as normal and abnormal, the overall error in the model would be **1.97%** which is very low in comparison with the model created by Jonas Norup, which is at **5.13%**. Moreover, this approach to classify data will work for any labelled dataset, where we would only be requiring to change the parameters of the modules used.

The important techniques used is the selective feedback in the logistic regression models and the novelty of designing a two- step classification model in order to achieve the 4 different classes of abnormal cells from the PAP smear dataset with least error.

Table:9: Final results in comparison to NCC method Employed by Jonas Norup[2]

Technique	Overall Error
Nearest Class gravity Center (NCC), Jonas Norup[2]	5.13%
Two Step Logistic Regression Classifier	**1.97%**

Acknowledgment

I am highly indebted to Dr. D. Brindha for her guidance and constant supervision as well as for providing necessary information regarding the research & also for her support in completing the same. I would like to express my gratitude towards the members of The Department of Biomedical Engineering, PSG College of Technology for their kind co-operation and encouragement which helped me in completion of this research work.

Bibliography

[1] Jan Jantzen, Jonas Norup, George Dounias, Beth Bjerregaard,"Pap-smear Benchmark Data For Pattern Classification", Technical University of Denmark, jj@oersted.dtu.dk

[2] Norup, Jonas c960566, "Classification of pap-smear data by transductive neuro-fuzzy methods", May 2, 2005

[3] Azure ML Documentation: https://studio.azureml.net/

[4] PAP Smear Test Uses and Diagnostics : https://www.medicinenet.com/pap_smear/article.html

[5] Shan Suthaharan, Mohammed Alzahrani, Sutharshan Rajasegarar, Christopher Leckie, Marimuthu Palaniswami, "Labelled Data Collection for Anomaly Detection in Wireless Sensor Networks", Proceedings of the Sixth International Conference on Intelligent Sensors Sensor Networks and Information Processing (ISSNIP 2010, Dec 2010).

[6] D Heckerman, D Geiger, D M Chickering. Learning Bayesian networks: The combination of knowledge and statistical data[J].Machine Learning, 1995,20(3): 197-243.

[7] Byriel, J. (1999). Neuro-fuzzy classification of cells in cervial smears. Master's thesis, Technical University of Denmark, Oersted. Dept. of Automation.

[8] Tong-Sheng Chen, Xue-Qin Hu, Shao-Zi Li and Chang-Le Zhou, "Multi-class diagnosis classification on high dimension data by logistic models," *2008 International Conference on Machine Learning and Cybernetics*, Kunming, 2008.

[9] P. Rao and J. Manikandan, "Design and evaluation of logistic regression model for pattern recognition systems," *2016 IEEE Annual India Conference (INDICON)*, Bangalore, 2016.

[10] Zhong Li and K. Najarian, "Automated classification of Pap smear tests using neural networks," *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, Washington, DC, USA, 2001

[11] K. Hemalatha and K. U. Rani, "An Optimal Neural Network Classifier for Cervical Pap Smear Data," *2017 IEEE 7th International Advance Computing Conference (IACC)*, Hyderabad, 2017