# Classification Model to Detect PCOS from Pap smear Images.

Abhinav Rawat

## Table of contents

| Chapter # | Title | Page # |
|---|---|---|
| 1. | Introduction | 3 |
| 1.1 | Overview | 3 |
| 1.2 | Objective of the Study | 3 |
| 2. | Methodology | 4 |
| 2.1 | Data Preparation | 4 |
| 2.2 | Model and Training | 4 |
| 2.3 | Results | 5 |
| 3. | Discussion and Conclusion | 6 |

## Chapter 1

**1. Introduction:** Polycystic ovary syndrome (PCOS) is a hormonal imbalance that occurs ovaries produce unusually high levels of hormones generally androgens. This leads to an imbalance in reproductive hormones causing irregular menstrual cycles, missed periods and unpredictable ovulation. PCOS is one of the most common causes of infertility in women, it can also increase risk of other health conditions, up to 15% of women of reproductive age have PCOS. The Papanicolaou test also know as Pap test or Pap smear, is a method of cervical screening used to detect PCOS. The collected cells are examined under a microscope to look for abnormalities. PCOS is one of the most common endocrine disorder observed in reproductive-aged woman, despite costing 8 billion USD in healthcare costs, 75% of patients with PCOS are unidentified in clinical practice. This confusion surrounding PCOS diagnosis is due to the broad symptomology. One solution which we propose is to implement machine learning algorithms on pap smear images of positive and negatively identified patients of PCOS.

**1.1 Overview:** The pap smear PCOS detection is a machine learning initiative designed to enhance the diagnosis of PCOS automatically predicting the presence of PCOS in pap smear images of patients. Leveraging advanced Machine Learning and Polar coordianate transformation techniques we extract data to be used for identification of PCOS from images, unlike Deep learning we will isolate information from the images and not feed the images into the algorithm. Since we are looking for speed along with accuracy. Using a systematic approach that includes reading image data, find key pixel locations, extracting pixel intensities of interest, performing radial polar transformation, extracting image information, apply SMOTE a data balancing algorithm, training model on classification algorithms, making predictions.

**1.2 Objective of the Study:** This project aims to classify pap smear PCOS images correctly into respective categories of PCOS +ve and PCOS -ve.

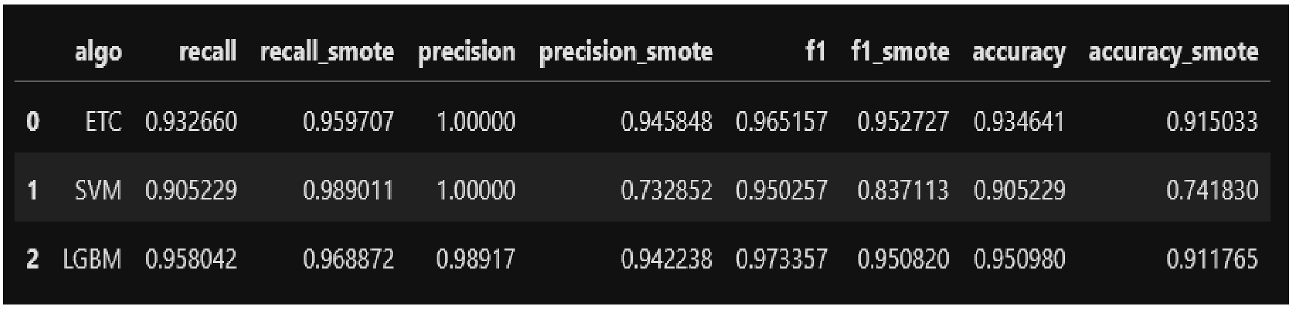
# Chapter 2

**2. Methodology:** The dataset consists of pap smear images which are labelled as positive or negative for PCOS, the images are '.bmp' format. All the work is performed using Python utilizing its vast library support for reading image data (Opencv), manipulating images with numpy and scipy, matplotlib.pyplot for plots and observing image data, sckit-learn to machine learning. Since the data is heavily skewed for positive labels, SMOTE was applied to balance the data. Synthetic Minority Oversampling Technique or SMOTE for short, is an approach to address imbalanced datasets by oversampling the minority class. One might duplicate the existing minority class features but that will not add any new information to the model. Instead SMOTE synthesises new examples from existing example, this is a type of data augmentation for the minority class. This technique is performed after image feature extraction since it is difficult to create new images from scratch.

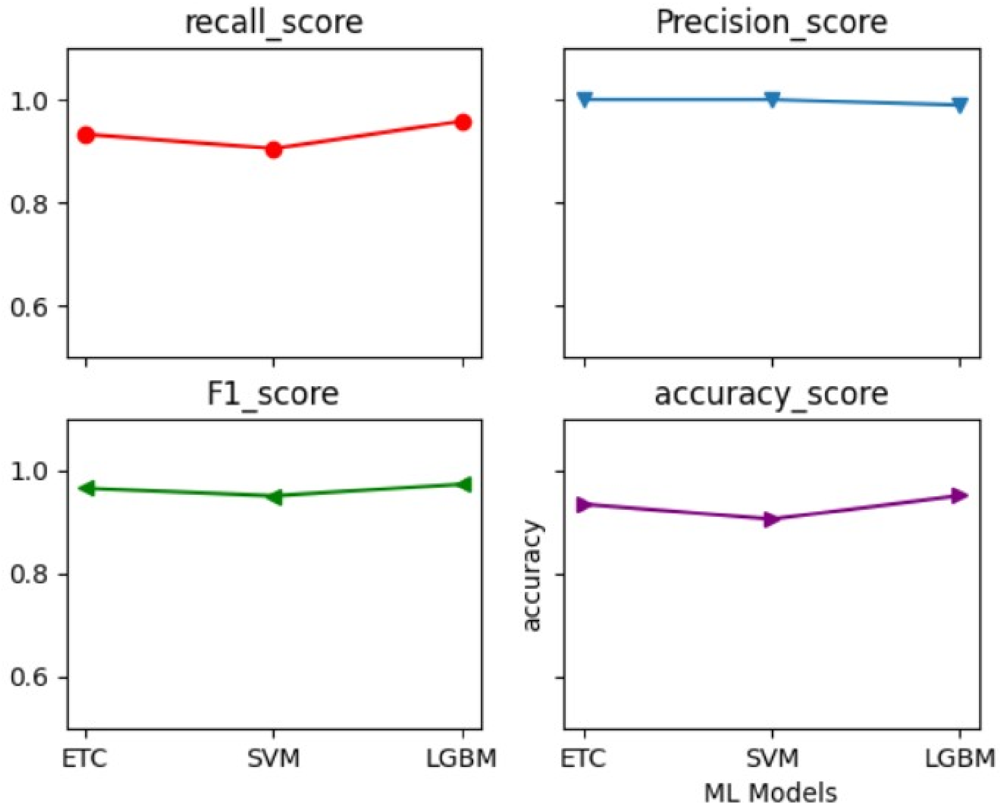
**2.1 Data Preparation:** The data consists of around 1500 images labelled as positive PCOS and negative on PCOS. To extract the image pixels we first performed Radial Polar Transformation, when we invert the pixel values in the images the cells are the brightest points in the images either PCOS cells or normal cells. We apply radial polar transformation and extract the top 50 pixel values, using these values a feature data frame was created with each sample giving 50 pixel values and a label which categorises sample to be positive or negative on PCOS. Since we are building a machine learning model these are converted to labels '1' and '0'. As observed the data is heavy biased towards positive PCOS as most of the data stored is of positive detection rather than negative detection in clinical practice, we will have to perform data augmentation technique to generate new examples for minority label samples. This was achieved using SMOTE, as described above, after which data is divided into training and testing splits 80% & 20% respectively. SMOTE is performed only on the training data, since we would like to predict on real world samples rather than augmented samples.

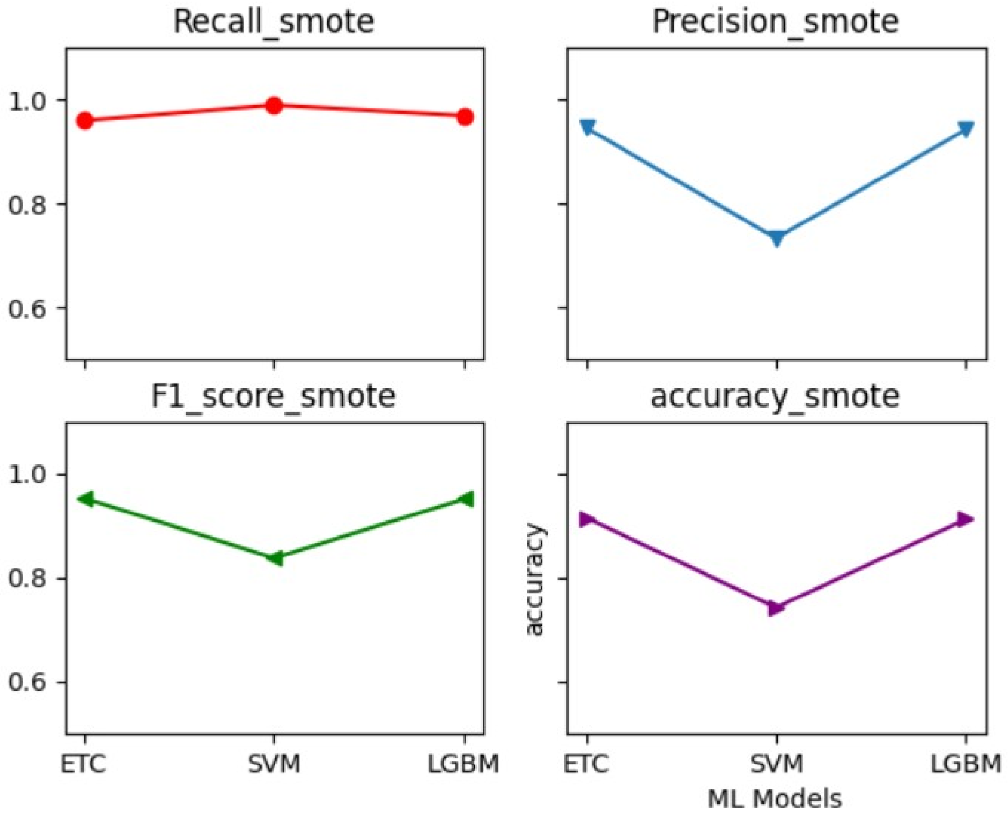
**2.2 Model and Training:** In this project we will train Eight different machine learning algorithms to determine which performs best and later fine tune hyper parameters for the best three. The different algorithms used for this study are ExtraTree, DecisionTree, Kneighbors, SGD, XGB, LGBM, RandomForest and SVM. Training model is performed on balanced data and ExtraTree Classifier, LGBM Classifer, RandomForest Classifier and SVM perform the best, hyper parameters are fine tuned for these classifiers to improve the prediction capabilities.

**2.3 Results:** After hyper parameter tuning the models improved its prediction accuracy from 90% to 95%, time taken to the entire model was close to 2 minutes from feature extraction to model predictions. ExtraTree and RandomForest Classifiers did not show improvements after hyperparameter tuning. We ran the models on pre data augmentation and post, we observe that the models post data augmentation showed better performances since the model can get away by predicting all positives as unbalanced data is majorly positive, we can observe that when we look at recall and F1 scores.



| | algo | recall | recall_smote | precision | precision_smote | f1 | f1_smote | accuracy | accuracy_smote |
|---|---|---|---|---|---|---|---|---|---|
| 0 | ETC | 0.932660 | 0.959707 | 1.00000 | 0.945848 | 0.965157 | 0.952727 | 0.934641 | 0.915033 |
| 1 | SVM | 0.905229 | 0.989011 | 1.00000 | 0.732852 | 0.950257 | 0.837113 | 0.905229 | 0.741830 |
| 2 | LGBM | 0.958042 | 0.968872 | 0.98917 | 0.942238 | 0.973357 | 0.950820 | 0.950980 | 0.911765 |





## Chapter 3

**3. Discussion and Conclusion:** From the results we can observe that the best algorithm for classifying pap smear images for PCOS are Extra Tree classifier and LGBMClassifier, where as Support Vector Machines (SVM) has a reduction is prediction accuracy after the data augmentation, this can be due to the doubling of samples as SVM's do not perform well on large data. As theorized precision score took a hit after data augmentation since its easier for the model to put all predictions as 1's since there are very few 0's and likelihood of the model being wrong reduces, but when there is a more balanced data the models under performs. SVM usually underperforms when there are more samples, a similar observation can be made for light GBM models as well, XGB Classifier shows little to no difference in accuracy as it does not get effected by the size of the data. Few interesting observations were made, a) we found three classification models which classify pap smear PCOS pretty accurately (upwards of 90%), b) Size of the data effected the ability of the model, c) SMOTE data augmentation effected the model slightly.