LITERATURE SURVEY:

According to the literature survey by Danaei et al.(1) on the diagnosis of PCOS, the steps to predict include reducing overfitting by filtering and performing feature selection on the dataset. After reducing the dimensionality of the dataset, the classification (diagnosis) is done using ensemblers. They have obtained an accuracy of 99,8% using ADABoost, MLP and Random Forest Algorithms.

This approach can also be used to detect other medical illnesses like auto-immune disorders, diabetes, autism spectrum disorder(ASD), etc., thus linking machine learning and human health.

In the diagnosis done by Subha et al.(2) the concepts of Swarm Intelligence and Flashing Firefly algorithm are used to intelligently select the features using statistical ML methods like Chi square and correlation. On this reduced dataset with optimal number of features the classification is done using Random Forest after the train-test validation. The method used to calculate the performance is using the classification metrics namely accuracy, precision, recall, specificity and also using the F1 score.

In the prediction system developed by Denny et al.(3) the feature selection is done using an advanced analytics tool SPSS V 22.0 developed by IBM for multivariate analysis thus reducing the features from 23 to 8. On this dataset multiple classification algorithms are applied to test which algorithm provides the best accuracy. Out of multiple algorithms like CART, SVM, KNN, RF the best accuracy was obtained using Random Forest(89.02%) and is therefore used in prediction of PCOS.

In the work done by Ning-Ning et al.(4), they have linked PCOS detection with the concept of genetics. Gene samples are collected from samples and it is made to undergo a sequence of tests and computations to analyse and understand the difference in the genetic makeup of women with and without PCOS. Most of the analysis is done graphically using area under curve (AUC) and artificial neural networks(ANN). As a product, they have developed a model that uses the RNA genotype as an input and makes the required predictions.

The study by Zhang et al.(5) involves tracking the menstrual cycles of different women and studying the change in the follicular stages via images. Raman Spectroscopy is used here to analyse the fluid in the follicles. PCA was done and various gradient boost algorithms were applied for easy distinguishing between PCOS vs Non-PCOS women. This kind of diagnosis proved to be very accurate and effective.

In the work done by Bharati et al.(6) FH & LSH are considered to be the primary hormones that cause PCOS among the other 8 primary factors. 40 fold cross validation is done on the dataset. A hybrid of two algorithms is employed for classification: Random Forest and Logistic Regression.

Using the RFLR a high testing accuracy is obtained(91%) with a recall of 90%.

The research by Suha et al.(7) involves the usage of computer based systems and convolutional neural networks to detect ovarian cysts instead of physical-manual methods. Feature selection from images is done using transfer learning method. Here, the usage of various ensembling techniques is involved along with bagging and boosting techniques. The model is trained, saved and results are finalised using VGGNet16, a 16 layer object detection technique of CNN model. As it is a very deep convolutional network the accuracy obtained is almost 99.79%

S. Tiwari et al.(8) have utilised out of bag error to analyse the predictions made by Random Forest algorithm after data cleaning and filtering. They have used multiple classification algorithms alongside boosting techniques and meta learners to provide visualisations for factors affecting PCOS.

The dataset used here is scanned ultrasound images. Also, the assessment of performance of multiple algorithms done in this work helps analyse the accuracy of the various ML classification algorithms.

In the analysis done by Nandipati et al.(9) on RapidMiner using KNN-SVM and Random Forest, two important observations were made, one being the accuracy of both algorithms were very high and very similar(94%) and the other observation being that the optimal factors could be chosen based on the performance by the two models.

In the PCOS study done by Thomas et al.(10) data was collected from hospital forms, processed, filtered and classified using MATLAB. Neural Network simulation was done effectively using feed forward propagation. Later in the study hybridisation of NN was done with Naive Bayes classifier to obtain a stronger accuracy of classification, thus concluding that the accuracy produced by hybrid classifiers is higher than a single high performance classifier.