

A Non-Invasive Approach for Early Polycystic Ovary Syndrome (PCOS) Prediction

A thesis

Submitted in partial fulfillment of the requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Submitted by

Mrinmoy Saha Joy	190204017
Nura Zabin	190204003
Sadia Islam	190204010
Marzia Binta Monir	190204114

Supervised by

Ms. Nawshin Tabassum Tanny



Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Dhaka, Bangladesh

April 03, 2024

CANDIDATES' DECLARATION

We, hereby, declare that the thesis presented in this report is the outcome of the investigation performed by us under the supervision of Ms. Nawshin Tabassum Tanny, Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh. The work was spread over two final year courses, CSE4100: Project and Thesis I and CSE4250: Project and Thesis II, in accordance with the course curriculum of the Department for the Bachelor of Science in Computer Science and Engineering program.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Mrinmoy Saha Joy
190204017

Nura Zabin
190204003

Sadia Islam
190204010

Marzia Binta Monir
190204114

CERTIFICATION

This thesis titled, “**A Non-Invasive Approach for Early Polycystic Ovary Syndrome (PCOS) Prediction**”, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in April 03, 2024.

Group Members:

Mrinmoy Saha Joy	190204017
Nura Zabin	190204003
Sadia Islam	190204010
Marzia Binta Monir	190204114

Ms. Nawshin Tabassum Tanny
Lecturer & Supervisor
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Dr. Md. Shahriar Mahbub
Professor & Head
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

ACKNOWLEDGEMENT

The authors would like to extend their heartfelt gratitude to their supervisor, Ms. Nawshin Tabassum Tanny, Lecturer, Department of Computer Science and Engineering, Ahsanullah University of Science and Technology. Her guidance, patience, and expertise were incredibly helpful throughout their research. Authors are really grateful to her that she chose to mentor them and for maintaining faith in their abilities over the course of the year. Finally, they want to thank everyone who has played a role in the completion of this thesis. The support and contributions were really important, and they appreciate all the provided help with this academic work.

Dhaka
April 03, 2024

Mrinmoy Saha Joy

Nura Zabin

Sadia Islam

Marzia Binta Monir

ABSTRACT

Polycystic ovary syndrome(PCOS) is an endocrinological disorder that affects women of reproductive age. Women who have PCOS have abnormal levels of the male hormone, which causes hirsutism(excessive facial and body hair), skin darkening, acne, hair loss, and obesity. Untreated or long-term hormonal imbalance can lead to infertility, endometrial cancer, type 2 diabetes, and cardiovascular diseases. Like many countries, women in Bangladesh also suffer from this medical condition. Unfortunately, a lot of young women with PCOS remain undiagnosed because they are unaware of the symptoms and they are uncomfortable with the complicated diagnosis process. In recent years, machine learning and deep learning algorithms have been used extensively in medical diagnosis. To make the diagnosis easier with early symptoms, we proposed a non-invasive approach using both machine learning and deep learning. To train and test machine and deep learning models, we used a Kaggle dataset of 541 women, where 177 women have PCOS. We used Chi-square and Extra-tree classifiers for feature selection and extracted 10 non-invasive features for each feature selection method. In this thesis, we analyzed machine learning models such as Random Forest Classification, Gradient-Boost Classification, AdaBoost Classification, and SVC. From the analysis of machine learning models, we get that RFC has the highest accuracy of 86.4% from Chi-square method selected features. SVC has the second-highest accuracy of 85.2% from Extra Tree classifier selected features. However, we have also analyzed deep learning models like LSTM and MLP. We get the highest accuracy of 89.91% using Extra Tree classifier selected features using the deep-learning model LSTM. We have developed a web application with non-invasive features that allows any woman to determine whether or not she may have the possibility of having PCOS, this app make the diagnosis more widely available.

Contents

CANDIDATES' DECLARATION	i
CERTIFICATION	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	iv
List of Figures	vii
List of Tables	viii
1 Introduction	1
1.1 Motivation	2
1.2 Objective	2
2 Literature Review	4
2.1 Reviews	4
2.1.1 Accessible Polycystic Ovarian Syndrome Diagnosis Using Machine Learning	4
2.1.2 Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms	5
2.1.3 i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques	6
2.1.4 An Efficient Decision Tree Establishment and Performance Analysis with Different Machine Learning Approaches on Polycystic Ovary Syndrome	6
2.1.5 A Classification of Polycystic Ovary Syndrome Based on Follicle Detection of Ultrasound Images	7
2.1.6 Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence	8
2.1.7 An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image	9
2.1.8 Classification of polycystic ovary based on ultrasound images using competitive neural network	9
2.1.9 Polycystic Ovarian Syndrome Detection Using Deep Learning	10
2.2 Summary	12

2.3	Research Gap	13
3	Background Studies	14
3.1	Feature Selection Methods	14
3.1.1	Chi-Square	14
3.1.2	Extra tree classifier	15
3.2	Training Models	15
3.2.1	Machine Learning Models	15
3.2.2	Deep Learning Models	18
3.3	Evaluation metrics	20
3.3.1	Accuracy	20
3.3.2	Precision	21
3.3.3	Recall	21
3.3.4	F1-Score	21
4	Dataset	22
5	Project Management	23
6	Methodology	24
6.1	Data Pre-processing	25
6.2	Feature Selection	26
6.2.1	Chi-Square	26
6.2.2	Extra Tree Classifier	27
6.3	Used Model	29
6.3.1	Machine Learning Models	29
6.3.2	Deep Learning Models	31
6.4	Web-app Demostation	32
7	Result	34
7.1	Result Analysis	34
7.2	Graph Analysis	36
7.3	Model implementaion on Web-app	37
8	Conclusion and Future Work	39
8.1	Future Work	39
8.2	Conclusion	40
	References	40

List of Figures

3.1	SVC model methodology	16
3.2	Random Forest Algorithm [8]	17
3.3	Adaboost model methodology	17
3.4	LSTM Sturcture [17]	19
3.5	Multilayer Perceptron [5]	20
4.1	Dataset	22
5.1	A Gantt Chart displaying the schedule, milestones, and dependencies through- out the timeline	23
6.1	Methodology of A Non-invasive Approach of PCOS Prediction using Machine Learning	25
6.2	Ranking of features based on Chi-square method	26
6.3	Ranking of features based on Extra tree classifier method	28
7.1	Validation graph for LSTM	36
7.2	Validation graph for MLP	36
7.3	User input of health features	37
7.4	Successful Prediction of PCOS class	37
7.5	Successful Prediction of Non-PCOS class	38

List of Tables

2.1 Literature Review Summary	12
4.1 Class distribution	22
6.1 Top 20 Features by Chi-Square	26
6.2 Top 20 Features by Extra Tree Classifier	27
6.3 Optimal parameter for SVC	29
6.4 Optimal parameters of Random Forest	30
6.5 Optimal parameter of Gradient Boosting Classification	30
6.6 Optimal parameter of AdaBoost Classification	31
6.7 Optimal parameter of LSTM model	31
6.8 Optimal parameter of MLP model	32
7.1 Result analysis for top 10 easily accessible noninvasive features by ML	35
7.2 Result analysis for top 10 easily accessible noninvasive features by Deep Learning	35

Chapter 1

Introduction

Polycystic ovary syndrome(PCOS) is one of the most common endocrinological disorders where an abnormally high level of the male hormone is observed in the female body. This abnormal hormonal level negatively affects normal ovarian processes and can result in the formation of multiple cysts inside the ovary. Furthermore, because this condition is complex, a wide range of symptoms might be seen. The three most common factors associated with PCOS are cystic ovaries, increased testosterone levels, and irregular ovulation. A high level of testosterone in the body results in hyperandrogenism. Acne, skin darkening, hirsutism (facial and body hair), hair loss on the head, infertility, weight gain, and irregular periods are among its most typical symptoms [6]. PCOS is a common hormonal disorder that affects women of reproductive age(15-45). It affects 5 to 10 percent of reproductive-age women, but the prevalence may be higher if broader criteria are used. [22]

PCOS can have a variety of effects on a female's body. Long-term hormonal imbalances may lead to infertility. Moreover, Women with PCOS are more likely to develop type 2 diabetes and cardiovascular disease. Because of hormone imbalances, women with PCOS are more likely to develop endometrial cancer. So, early detection of PCOS is vital, if left untreated, it can lead to these critical medical conditions.

PCOS is a common medical condition that affects women's health significantly. However, diagnosing PCOS requires a combination of clinical and biochemical parameters, which is quite challenging. Traditional diagnostic techniques are frequently time-consuming and costly. So, it is important to develop a method that will be less time-consuming and easier to diagnose PCOS. In recent years, Machine learning algorithms and Deep learning algorithms have become a potential tool for the early detection and diagnosis of many medical illnesses. Both of these processes can be an effective tool to diagnose PCOS at an early stage.

Machine learning algorithms and Deep learning algorithms will be used for a comparative study to early predict PCOS. To make the diagnosis less time-consuming and easily acces-

sible, a non-invasive approach will be implemented. Those extracted non-invasive features will be used to train and test our machine-learning model and deep learning model. Different machine learning models, such as Random Forest Classification, GradientBoost Classification, AdaBoost Classification, and SVM(Support Vector Machine) will be used. And for deep learning, Long-Short Term Memory(LSTM) and Multilayer Perceptron(MLP) model will be used. The dataset of 541 patients from Kaggle, where 177 people have PCOS, will be employed.

The result of this study will be to help people to diagnose their medical condition at an early stage with accessible symptoms. Moreover, it will help those who have PCOS to detect their medical condition before it worsens.

1.1 Motivation

PCOS is also a very common disorder in Bangladesh. Many adolescents and women in their reproductive age suffer from this illness. However, women in Bangladesh are not well aware of this illness and their health condition. Many young girls feel embarrassed talking about their symptoms. Moreover, they do not feel comfortable checking up in hospitals. Thus their symptoms are left undiagnosed which can later lead to a critical medical condition. It will be very helpful for these women if they have an accessible approach where they can easily identify their symptoms without any medical test.

1.2 Objective

The primary objective of this research is to develop a non-invasive methodology for the early prediction of Polycystic Ovary Syndrome (PCOS), a common endocrine disorder in women. The research utilizes a Kaggle dataset comprising 541 instances and 41 features to predict the presence of PCOS. Efforts are directed towards simplifying the diagnostic process by identifying non-invasive features. Various feature selection techniques, including Chi-square and Extra Tree Classifier, are employed to select the most relevant features.

Subsequently, different machine learning algorithms, namely Random Forest Classification, Gradient-Boost Classification, AdaBoost Classification, and Support Vector Machine (SVM), are implemented to model and predict PCOS. Furthermore, deep learning models such as Long-Short Term Memory (LSTM) and Multilayer Perceptron (MLP) are also explored to ascertain their efficacy in PCOS prediction.

A comprehensive comparison is conducted between traditional machine learning models and deep learning models to determine the most effective approach for PCOS prediction.

Additionally, the research aims to develop an accessible tool for PCOS prediction, facilitating early detection of PCOS. Through this research, we endeavor to contribute to the advancement of diagnostic methods for PCOS, and through the developed application users can input relevant health data, and the system will generate a PCOS risk assessment based on learning model trained on extensive datasets. This allows individuals to understand their likelihood of developing PCOS, prompting necessary proactive measures and health awareness.

Chapter 2

Literature Review

There are various kinds of studies and research articles on the classification of PCOS using different kinds of ML models and Deep learning models. To predict PCOS many kinds of features are required, among which most of them are invasive. In consequence, over the years researchers have been trying to figure out the features which can be easily accessible and diagnose PCOS appropriately. In the following section, we have reviewed some relevant papers, in which researchers have explored approachable diagnosis of PCOS.

2.1 Reviews

2.1.1 Accessible Polycystic Ovarian Syndrome Diagnosis Using Machine Learning

Akanksha Tanwar et al. [21] in their work aims to promote early diagnosis of PCOS in women who are at a greater risk of being diagnosed. They have also aimed to make pre-diagnosis of PCOS easily accessible to every woman via a web application.

- **Methods:** They have used a Kaggle dataset containing 541 data from 10 different hospitals in Kerala, India. The dataset has 44 different features. For selecting the top five features they have used 3 different feature selection techniques. These are Chi-Square Method, Extra Tree Classifier, and Correlation Matrix. They have applied the Random Forest Classifier model which has given an accuracy of around 92.59%. The web application takes the input for the following five features: Skin Darkening(Y/N), Hair Growth(Y/N), Weight Gain(Y/N), Period Cycle (R/I), and Fast Food(Y/N). The user can give input as 1/0 for all the five parameters and click on predict. Here 1 stands for YES and 0 stands for NO. If the user has a significant chance of having PCOS, the

app will display “You have significant chances of having PCOS. Please check with your doctor as soon as possible. And if the user has comparatively lesser chances of suffering from PCOS, it will show- “You do not have significant chances of having PCOS. In the future, they are hoping to extend some more facilities in their web application such as nearest gynecologists, best practices, online consultations, government health schemes etc.

- **Strength:** Easy accessible features and User-friendly web application.

2.1.2 Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms

Bharati et al. [12] focus on the prediction of PCOS using machine learning classifier in their paper. The main contributions are, using a feature selection model to select the most important attributes for the given dataset, then applying machine learning algorithms on the important features of the PCOS dataset, and finally comparing algorithms in terms of accuracy and recall.

- **Methods:** They have used a dataset with 43 attributes of 541 women collected from the Kaggle repository. Out of these 541 instances, 364 are for normal and the remaining 177 are for PCOS-affected patients. They have used the univariate feature selection method for identifying the important features. Using holdout and cross-validation methods they have divided the dataset into training and testing portions. They have applied a few classifiers such as gradient boosting, random forest, logistic regression, and hybrid random forest and logistic regression (RFLR). Results show that the first 10 highest-ranked attributes are quite good at predicting PCOS disease. Among these classifiers, RFLR exhibits the best testing accuracy of 91.01% and recall value of 90% using 40-fold cross-validation. Attribute ranking is computed and it is found that the most important attribute is the ratio of Follicle-stimulating hormone (FSH) and Luteinizing hormone (LH).
- **Strength:** Holdout and cross-validation methods are applied to separate the training and testing datasets.

2.1.3 i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques

Amsy and Raj et al. [14] propose a system for the early detection and prediction of PCOS from optimal and minimal but promising clinical and metabolic parameters. They have developed a system that automates PCOS detection based on a minimal set of potential markers.

- **Methods:** They have considered a total of 541 samples which were collected from various clinics and hospitals in and around the district of Thrissur. There are 23 features, including the reports on transvaginal Ultrasound scans, hormone profile, and lifestyle of the patient with impressions on physical fitness. They have selected the most contributing feature set by terminating the redundant data set through the implementation of Principal Component Analysis (PCA). The important features for the final design of the system have been chosen with the help of a significance study individually based on the independent sample test, Pearson, and Spearman's rho correlation of parameters with the help of SPSS software by IBM. Around 12 features were identified in SPSS. According to their result, AMH turns out to be a very promising feature for detecting PCOS and infertility. They have applied the simple linear algorithms LR and LDA, Non-linear methods are KNN, CART, RFC, NB, and SVM. They have reset the random number seed in each run to the data split. By analyzing the confusion matrix of each model, they have selected the best model with better performance. Among all models, Random Forest Classifiers have given the best performance with an accuracy of 89%.
- **Strength:** Dataset Collection and Real-world Accuracy.

2.1.4 An Efficient Decision Tree Establishment and Performance Analysis with Different Machine Learning Approaches on Polycystic Ovary Syndrome

Aroni Saha Prapty et al. [18] mainly focus on improving the performance of machine learning approaches. They have developed a very useful decision tree using the Random Forest and top PCOS-responsible features are chosen.

- **Methods:** In their proposed method, they have used a total of 542 data where 177 people are affected by PCOS and the rest of the patients are normal. For the confirmation of having PCOS, it contains 31 individual features. But they have focused on around 7-8 features. Data preprocessing is pursued by dividing the data into a

training set and testing set for model creation, where 70% of data have been used as training data and the rest of them have been used as testing data. They have applied KNN, SVM, Naive classifier, and Random Forest methods and have made a comparison among these. To make the comparison they have considered accuracy, precision, recall, and f1 score as parameters. After analyzing the performances Random forests can be concluded with the best classifiers among others based on this overall performance with an accuracy of 93.5%. So Random forest is used for further analysis. They are well pleased with their chosen attributes as they have applied Principal Component Analysis (PCA) and the top 5 principal components of PCA match with their selection.

- **Strength:** Better accuracy is achieved using the Random Forest method where Shannon Entropy and Information Gain are the keys.

2.1.5 A Classification of Polycystic Ovary Syndrome Based on Follicle Detection of Ultrasound Images

Bedy Purnama et al. [19] delineated an application to classify Polycystic Ovary Syndrome based on follicle detection using USG images in the paper. This application is designed to help physicians to detect PCO follicles.

- **Methods:** The process to detect PCO follicles, consists of 5 stages. Stages are medical ultrasound image, preprocessing, segmentation, feature extraction, and classification. They have considered 80 images as a dataset, where 60 normal ovary images and 20 PCO ovary images. They have followed some stages at preprocessing. The first one is to obtain ROI. For noise elimination, they have used Gaussian-based noise filtering 5x5 windows. Using equalization histogram they have increased contrast level. Then the resulting images of the equalization histogram are inverted with a negative transform. So, the follicles are detected as white objects. For black background and white foreground, binarization is performed. At the segmentation stage, the background of the image is distinguished from the desired object. Here Canny method is used for edge detection. Then the labeled follicles have become new images to be processed. For feature extraction, the Gabor wavelet method is applied to produce a filter that is adjustable to the configuration of frequency and orientation utilized. After applying the Gabor wavelet they have gotten 2 groups of dataset. 1) data A, 40 images consist of 26 normal images and 14 PCOS-indicated images, with Mean texture feature, obtained 275 follicle images, 2) data B, 40 images consist of 34 normal images and 6 PCOS-indicated images, with Mean, Entropy, Kurtosis, Skewness, and Variance texture features, obtained 339 follicle images. They have applied the Neural Network-Learning Vector Quantization (LVQ) method, Support Vector Machine (SVM)

method, and K-Nearest Neighbor (K-NN) method for classification. Among these best accuracy is gained from SVM - RBF Kernel on $C=40$. It shows that dataset A reaches 82.55% while dataset B obtained from KNN-euclidean distance classification on $K=5$ reaches 78.81%.

- **Strength:** Use of Ultrasound images for the diagnosis of PCOS.

2.1.6 Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence

Hela Elmannai et al. [16] set their main goal to provide model explanations to ensure efficiency, effectiveness, and trust in the developed model through local and global explanations.

- **Methods:** The main contributions of this paper are 1) to solve class imbalance, a combination of SMOTE (Synthetic Minority Oversampling Techniques) and ENN (Edited Nearest Neighbour). 2) To optimize ML algorithms and enhance accuracy, Bayesian Optimization with cross-validation. 3) To reduce data dimensionality, applied feature selection (FS). 4) Proposed stacking ML and compared it with different ML models using evaluation methods. 5) Using global and local explainability terms, increased the model trust by clearly explaining the final prediction. They have used the PCOS dataset from Kaggle, which includes 541 instances and 41 attributes where 178 instances of the positive class and 363 instances of the negative class. To enhance the performance of machine learning models, they have re-sampled data using SMO-TEENN. To increase learning accuracy, the optimal feature subset is determined by feature selection (FS). FS is categorized into 3 main types. The dataset was split into two sets using a stratified sampling method, where 80% training set and 20% testing sets and a ratio of 70% training set and 30% testing set. Here Bayesian Optimization (BO) is applied to optimize different ML models. They have used different ML models, logistic regression (LR), random forest (RF), decision tree (DT), naive Bayes (NB), support vector machine (SVM), k-nearest neighbor (KNN), Xgboost, and the AdaBoost algorithm. To improve the model's overall performance, they have built The stacking ensemble model by combining decisions from several models with several ML in the base learner level and RF meta learner. After analyzing the results it is shown that Stacking ML with REF feature selection recorded the highest performance at 100 compared to other models, with the highest percentages of different evaluation Metrics at ACC, PRE, REC, and F1 at 98.87, 98, 98.87, and 98.89, respectively.

- **Strength:** Applied SMOTE (Synthetic Minority Oversampling Techniques) and ENN (Edited Nearest Neighbour) to solve class imbalance.

2.1.7 An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image

Sayma Suha et al. [20] propose an extended machine learning classification technique for PCOS prediction using ovary USG images as the main objective of this paper. Here they have used ultrasonography images of the ovary.

- **Methods:** A total of 594 images among which 123 have been acquired from various open sources from the internet and the rest of them are collected from two diagnostic centers and three hospitals in Bangladesh including Combined Military Hospital (CMH). To differentiate between PCOS and non-PCOS ovaries, they have trained and tested integrating ensemble ML models with Convolutional Neural Network(CNN) architecture. The reduced set of features has been explored from the best performing CNN model and then utilized as the input training data for the stacked machine learning classifiers. Images of the dataset are converted to grayscale colorspace using an OpenCV python function ‘COLOR.BGR2GRAY’ and are resized to 224X224 size. They have used Google Colaboratory as a platform for implementation and integrated mainly the python scikit-learn and TensorFlow packages. To train the machine learning models four types of techniques have been conducted. They have applied around ten different machine learning models, including Naive Bayes model, Decision Tree model, Support Vector Machine(SVM), K Nearest Neighbour(KNN) model, etc. After the analysis, they revealed the proposed hybrid strategy of employing the “VGG16” pre-trained model for transfer learning in the CNN architecture for feature extraction and then the “XGBoost” machine learning model as the meta-learner of stacking ensemble model for image classification yields the maximum accuracy of 99.89% and also have relatively shortest execution time to detect PCOS from ultrasound images.
- **Strength:** Use of XGBoost machine learning model as the meta-learner of stacking ensemble model for image classification.

2.1.8 Classification of polycystic ovary based on ultrasound images using competitive neural network

Kang Adiwijaya et al. [15] focuses on developing an automated system for classifying polycystic ovaries (PCO) using ultrasound images. The system employs Gabor Wavelet for fea-

ture extraction and a Competitive Neural Network (CNN) for classification. The CNN is chosen for its ability to classify data based on specific ultrasound characteristics, combining Hemming Net and The Max Net. The system achieved a maximum accuracy of 80.84% and a processing time of 60.64 seconds with 32 feature vectors and weight and bias values of 0.03 and 0.002, respectively.

- **Methods:** The research methodology includes several key steps: preprocessing ultrasound images, segmentation to isolate follicles, feature extraction using Gabor Wavelet, and classification using CNN. Preprocessing involves converting images to grayscale, histogram equalization, binarization, morphology, image inversion, and data cleaning. Segmentation is performed using edge detection and cropping. The Gabor Wavelet method is used for feature extraction, and the CNN, which operates on a winner-takes-all system, is classified into non-PCOS or PCOS categories.
- **Strength:** Significantly improves the efficiency of PCO detection, reducing the time and manual effort required by gynecologists to count and size follicles in ultrasound images. The use of CNN for classification, due to its combination of Hemming Net and The Max Net, allows for accurate classification based on the specific characteristics of ultrasound data.

2.1.9 Polycystic Ovarian Syndrome Detection Using Deep Learning

Shubham Bhosale et al. [13] explores the use of Deep Convolutional Neural Networks (DCNN) to classify and detect Polycystic Ovarian Syndrome (PCOS) using ultrasound images. The study aims to improve the accuracy and efficiency of PCOS diagnosis through automated image classification, addressing the challenges posed by doctors' manual inspection of ultrasound scans. The methodology involves DCNN based image classification for feature extraction and accuracy measurement on a dataset of PCOS related illnesses.

- **Methods:** The methodology section outlines the use of machine learning classifiers to predict PCOS in a dataset containing 43 attributes of 541 women, with 364 healthy cases and 177 PCOS sufferers. Python programming language and tools like Anaconda, Scikit-learn, Jupiter Notebook, and Spyder are employed for machine learning. The process includes developing an automated system for cyst detection from ultrasound images, evaluating various filtering techniques for speckle noise removal, enhancing the contrast of de-noised images, segmenting the cyst accurately from the ultrasound image background, optimizing features for classification, and comparing the performance of different classifiers to use the most accurate one. The performance of the proposed system is validated against existing methods

- **Strength:** The study introduces an innovative approach to PCOS detection using DCNN, which significantly enhances the accuracy and speed of diagnosis compared to manual methods. By employing deep learning techniques for image classification, the research addresses the challenges of manual cyst detection accuracy affected by overlapping follicles, inherent noise of the equipment, and lack of operator understanding.

2.2 Summary

Table 2.1: Literature Review Summary

Title	Dataset	Method	Result
Accessible Polycystic Ovarian Syndrome Diagnosis Using Machine Learning	PCOS data from Kaggle	Random Forest Classifier	92.59%
Diagnosis of Polycystic Ovary Syndrome Using Machine Learning Algorithms	PCOS data from Kaggle	Gradient Boosting	91%
		Random Forest	
		Logistic Regression	
		Hybrid Random Forest	
i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques	PCOS data from Kaggle & Manual	LR	89%
		LDA	
		KNN	
		CART	
		RBC	
		NV	
		SVM	
An Efficient Decision Tree Establishment and Performance Analysis with Different Machine Learning Approaches on Polycystic Ovary Syndrome	PCOS Data 542 data 31 features	KNN	93.5%
		SVM	
		Naive classifier	
		RF	
A Classification of PCOS Based on Follicle Detection of Ultrasound Images	80 Images	LVQ	82%
		KNN	
		SVM	
Polycystic Ovary Syndrome Detection Machine Learning Model Based on Optimized Feature Selection and Explainable Artificial Intelligence	PCOS data from Kaggle	LR	98.87%
		RF	
		DT	
		KNN	
		NB	
		SVM	
		Xgboost	
An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image	Manually Collected	Adaboost	99.89%
		NB	
		SVM	
		DT	
		Xgboost	
		Adaboost	
		Catboost	

Title	Dataset	Models	Accuracy
Classification of polycystic ovary based on ultrasound images using competitive neural network	Manually Collected	CNN	80.84%
Polycystic Ovarian Syndrome Detection Using Deep Learning	From kaggle	DCNN	

2.3 Research Gap

After analyzing these studies, some research gaps can be pointed out as follows.

1. Although early PCOS detection is widely progressed by using machine learning on extracted important features, many of the body features are invasive and patients have to go through surgeries or heavy medical tests. Very little research has been done to correctly predict PCOS using non-invasive body features which are known to patients and can be measured without involving any kind of equipment that breaks the skin.
2. Deep learning models for predicting PCOS have seen relatively fewer research efforts compared to machine learning approaches.
3. Predict PCOS with better accuracy and with less non-invasive features.
4. There is less accessible tool predict PCOS easily at home

Chapter 3

Background Studies

This background research analyzes different machine-learning models, deep-learning models, and feature extraction algorithms. Moreover, it emphasizes various evaluation metrics. This study focuses on applying various feature extraction methods such as Chi-square and Extra tree classifiers to choose the most important features. It explores different machine learning models and deep learning, especially classification models. Random Forest Classification, SVM, GradientBoost Classification, and AdaBoost Classification are some of the most widely used ml models for predictions, which are also analyzed in this study. Long-Short Term Memory(LSTM) and Multilayer Perceptron(MLP) models are some of the widely used deep learning models for prediction. This background study emphasizes each model's strengths and benefits to the field of classification and prediction. Moreover, it details the evaluation metrics that assess the performance of Machine Learning models and Deep Learning models.

3.1 Feature Selection Methods

The choice of feature extraction method depends on the nature of the data and the specific requirements of the machine learning task. It often involves experimentation to determine which method or combination of methods works best for a particular problem.

3.1.1 Chi-Square

Chi-Square [2], a statistical test which is used to compare the outcomes of data that can be organized into categories. To examine the existence of any difference between the observed values and expected value, Chi-Square Method is used. This difference between two events are computed based on a formula:

$$\chi^2 = \sum_a^b \frac{(O_i - E_i)^2}{E_i}$$

Here, χ is chi, O is the observed outcome, and E is the expected outcome. Lower independence of the target variable on the feature it is testing against is indicated by a lower Chi-Square value.

3.1.2 Extra tree classifier

Extremely Randomized Tree Classifier (extra-trees classifier) implements a meta estimator that fits a number of randomized decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Based on the Gini impurity the importance of features are calculated.

3.2 Training Models

The choice of a specific model has been determined on the basis of the characteristics of the data, and the desired outcome. Model selection often involves experimentation and evaluation to determine which model performs best for a given problem.

3.2.1 Machine Learning Models

Some of the common and widely used machine learning models which are effective and efficient to work with medical data are Support Vector Machine, Random Forest, Ada-Boosting and Gradient Boosting models. These models perform better than other traditional machine learning models to capture the inter-relation between multiple health features and medical diseases.

3.2.1.1 SVC

One well-liked supervised learning technique for classification problems is Support Vector Classifier (SVC) [10]. It determines the ideal hyperplane in the feature space that maximally divides the various classes. Data is mapped onto a higher-dimensional space using SVC, determining which decision boundary would best divide the classes. It works well for handling high-dimensional data and is capable of managing classification issues that are both linear and non-linear using kernel functions.

SVC is primarily used for binary classification, where the goal is to classify data into one of two classes. However, SVC can also be extended to handle multiclass classification problems using techniques like One-vs-One or One-vs-Rest, which break down the problem into multiple binary classification tasks. SVC operates by finding a hyperplane in the feature space that maximally divides the data points of different classes. The data points nearest to the decision boundary (hyperplane) are known as support vectors. These support vectors are utilized to mathematically define the hyperplane and are essential in establishing the decision boundary.

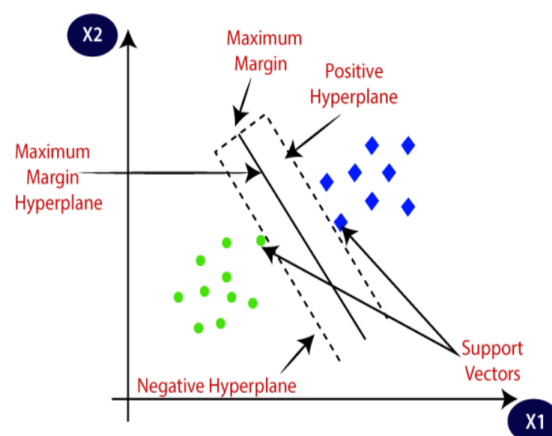


Figure 3.1: SVC model methodology

In this figure, two different categories are classified using a decision boundary or hyperplane. [9]

3.2.1.2 Random Forest Classifier

Random Forest [8] is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The number of features in each subset is equal to the square root of the number of total features in the dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

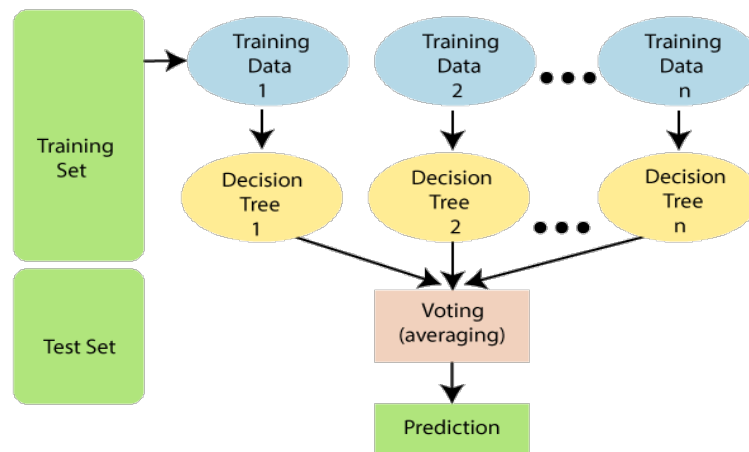


Figure 3.2: Random Forest Algorithm [8]

3.2.1.3 AdaBoost

AdaBoost algorithm [1], short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

What this algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points with higher weights are given more importance in the next model. It will keep training models until and unless a lower error is received.

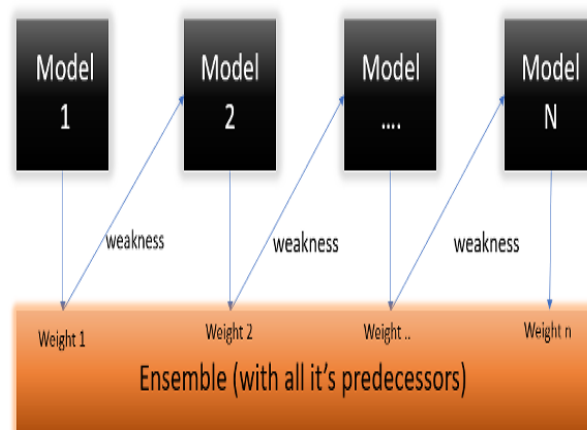


Figure 3.3: Adaboost model methodology

The basic ideas behind AdaBoost algorithm are, it combines a lot of 'Weak learners' to make classifications. The weak learners are almost always stumps, a decision tree node with two leaves. Some stumps get more say in the classification than others. Each stump is made by taking the previous stump's mistakes into account.

3.2.1.4 Gradient Boosting

Gradient Boosting [4] is a powerful boosting algorithm that combines several weak learning models to produce a powerful predicting model. In which each new model is trained to minimize the loss function such as mean squared error of the previous model using gradient descent. The loss function's purpose is to calculate how well the model predicts, given the available data. Depending on the particular issue at hand, this may change.

In each iteration, the algorithm calculates the gradient of the loss function with respect to the predictions of the current ensemble and then trains a new weak model to minimize this gradient. The predictions of the new model are then added to the ensemble, and the process is repeated until a stopping criterion is met.

Grading boosting systems can readily overfit on a training data set. Overfitting can be prevented by using various restrictions or regularization techniques that improve algorithm performance. Certain constraints can prevent overfitting depending on the decision tree's topology. Besides restricting the number of observations each split, the number of observations trained on, the depth of the tree, and the number of leaves or nodes in the tree, the gradient can be controlled. Also the contributions of the trees can be blocked or slowed down using a method known as shrinkage since the forecasts of each tree are added together.

3.2.2 Deep Learning Models

In recent years, deep learning models have gained significant attention and shown promising results in various fields, such as healthcare, text processing, computer vision, etc. Some commonly used architectures in deep learning are the Sequential model with LSTM and MLP. These models have distinct characteristics and are suitable for different types of tasks.

3.2.2.1 Long-Short Term Memory(LSTM)

Long Short-Term Memory [17], an improved version of recurrent neural network that allows information to persist designed by Hochreiter and Schmidhuber. There are three types of gates in an LSTM: the input gate, the forget gate, and the output gate. The memory cell is controlled by these three gates.

- The information that is added to the memory cell is controlled by the input gate.
- What information is removed from the memory cell is managed by the forget gate.
- Likewise, the output gate controls the information that the memory cell outputs.

This helps information to be retained or discarded selectively by LSTM networks as it passes through the network. The structure of LSTM is shown below.

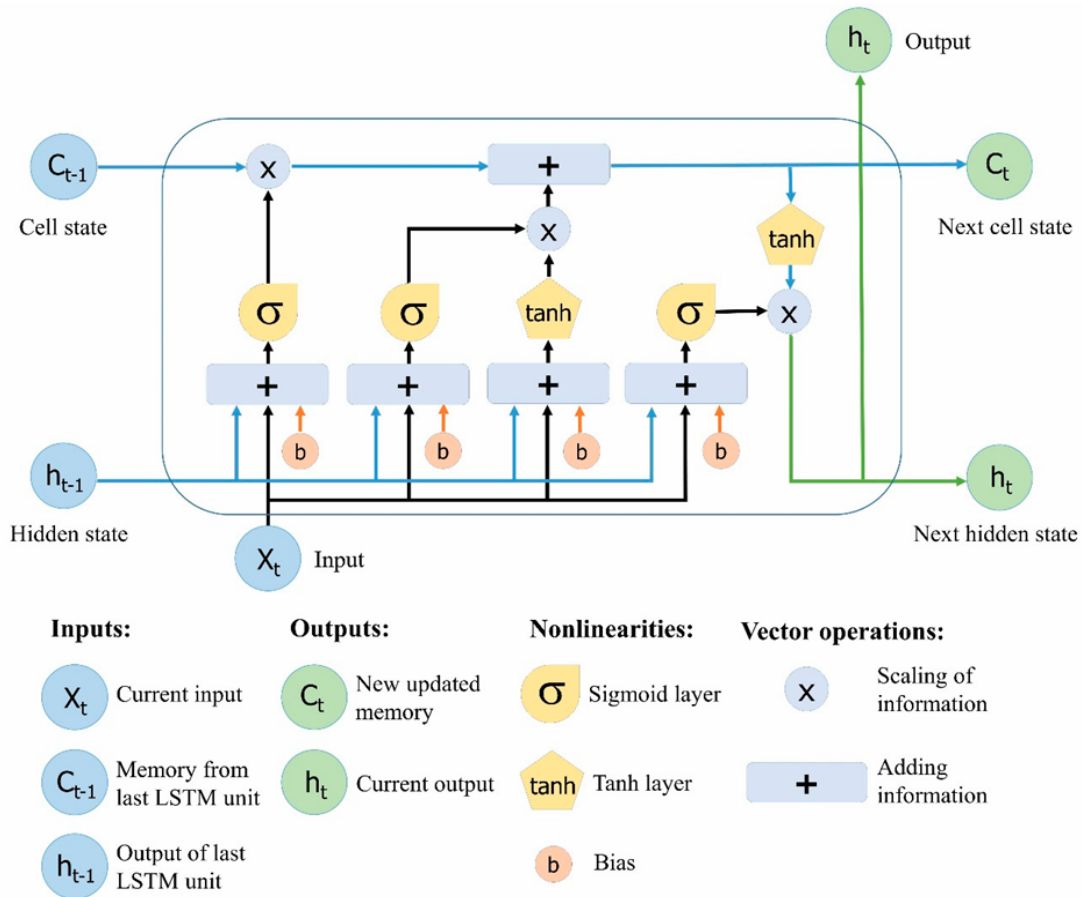


Figure 3.4: LSTM Sturcture [17]

3.2.2.2 Multilayer Perceptron(MLP)

A Multilayer Perceptron (MLP) [11] is one of the simplest and most common neural network architectures used in machine learning. It is a feedforward artificial neural network consisting of multiple layers of interconnected neurons, including an input layer, one or more hidden layers, and an output layer. In the context of Deep Learning, a Perceptron is usually referred to as a neuron, and a Multi- Layer Perceptron structure is referred to as a Neural Network.

MLPs are capable of learning complex and non-linear relationships in data, especially when they have multiple hidden layers and non-linear activation functions.

A Multi-Layer Perceptron (MLP) is a neural network model that learns to map inputs to outputs by adjusting its weights and biases through forward propagation and backpropagation. The input layer receives input data, which is then processed through hidden layers where

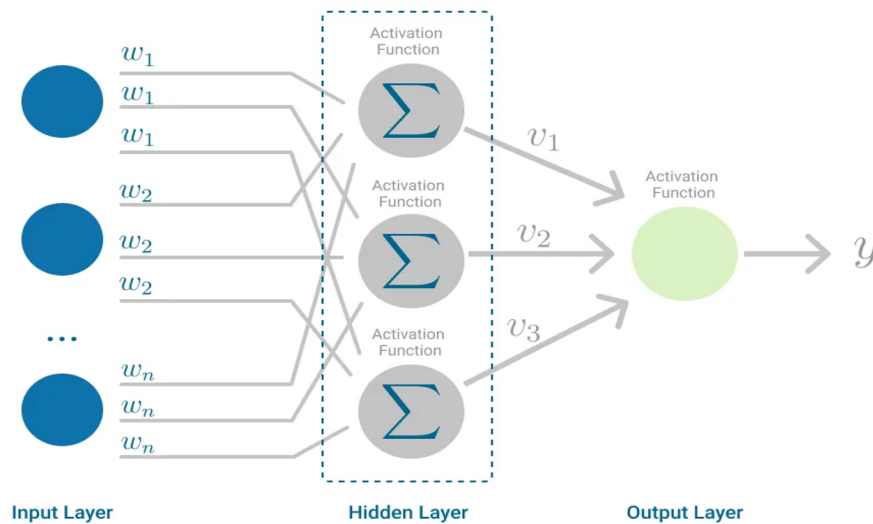


Figure 3.5: Multilayer Perceptron [5]

weighted connections and biases are applied, followed by activation functions to introduce non-linearity. Forward propagation computes the output layer's values, and errors are calculated by comparing predictions to actual data. Backpropagation adjusts weights and biases to minimize errors using optimization algorithms like stochastic gradient descent. This iterative process continues until the network converges, enabling accurate predictions.

3.3 Evaluation metrics

Evaluation metrics [3] is used to rigorously assess the performance of Machine Learning models. Sometimes it is called an Evaluation Framework. It forms the main analytical framework for an evaluation.

3.3.1 Accuracy

The ratio of accurately anticipated occurrences to the total number of instances is known as accuracy. For instance, the accuracy of a model is 0.9, or 90%, if it predicts 90 out of 100 instances accurately. The number of true positives and true negatives divided by the total number of instance yields the accuracy calculation:

$$Accuracy = \frac{CorrectPredictions}{AllPredictions}$$

It can also be written like this with the help of a confusion metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Here TP = True Positive : model correctly predicts the positive class TN = True Negative : model correctly predicts the negative class FP = False Positive : model incorrectly predicts the positive class. FN = False Negative : model incorrectly predicts the negative class. In short, it can be said that accuracy gives an overview of the model's performance.

3.3.2 Precision

The factor known as precision measures the rate with which a machine learning model accurately predicts the positive class.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

The precision can be expressed as a percentage or on a 0–1 scale. The more accurate the data, the better. Achieving a perfect precision of 1.0 indicates that the model consistently predicts the target class appropriately:

3.3.3 Recall

A model's recall is its capacity to identify every relevant case in a particular set of data.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Recall can be expressed as a percentage or on a 0–1 scale. Perhaps even better, the recall should be higher.

3.3.4 F1-Score

By taking the average of the harmonics of a classifier's precision and recall, the F1-score aggregates these two metrics into a single figure. The primary focus is to compare two classifiers' performances.

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Chapter 4

Dataset

A PCOS dataset [7] from Kaggle have been used which is collected from 10 different hospitals across Kerala, India. There are 41 features and 541 instances in this dataset shown in fig. 4.1. Such features include pimples, hair growth, cycles, vitamin d3, etc. It is a supervised dataset.

Patient File	Age (yrs)	Weight (Kg)	Height(Cm)	BMI	Blood Group	Pulse rate(bp)	RR (breath)	Hb(g/dl)	...	Follicle No. (R)	Avg. F size	Avg. F size	Endometrium	PCOS (Y/N)
1	28	44.6	152	19.3	15	78	22	10.48	...	3	18	18	8.5	0
2	36	65	161.5	24.9	15	74	20	11.7	...	5	15	14	3.7	0
3	33	68.8	165	25.3	11	72	18	11.8	...	15	18	20	10	1
4	37	65	148	29.7	13	72	20	12	...	2	15	14	7.5	0
5	25	52	161	20.1	11	72	18	10	...	4	16	14	7	0
6	36	74.1	165	27.2	15	78	28	11.2	...	6	16	20	8	0
7	34	64	156	26.3	11	72	18	10.9	...	6	15	16	6.8	0
8	33	58.5	159	23.1	13	72	20	11	...	6	15	18	7.1	0
9	32	40	158	16	11	72	18	11.8	...	7	17	17	4.2	0
10	36	52	150	23.1	15	80	20	10	...	1	14	17	2.5	0
11	20	71	163	26.7	15	80	20	10	...	15	17	20	6	0
12	26	49	160	19.1	13	72	20	9.5	...	2	18	19	7.8	0
13	25	74	152	32	17	72	18	11.7	...	8	20	21	8	1
14	38	50	152	21.6	13	74	20	12.1	...	3	18	17	5.6	0
15	34	57.3	162	21.8	13	74	22	11.7	...	1	19	21	5.5	0

Figure 4.1: Dataset

The classes in the dataset are not evenly distributed. They are distributed in these way :

Table 4.1: Class distribution

Class	Number of instances
Positive Class	177
Negative Class	364
Total	541

Chapter 5

Project Management

This section provides insight into the plan, organization, and execution of this study over time. It also reflects detailed methodology with respect to the timelines.

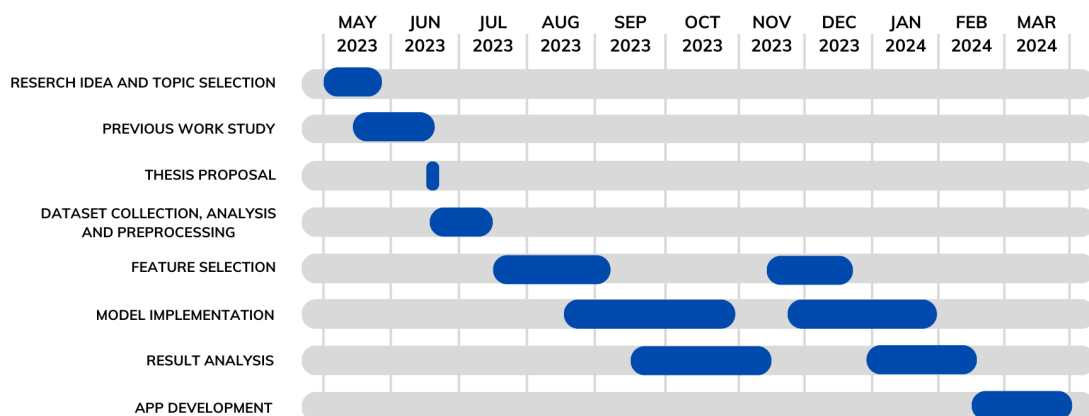


Figure 5.1: A Gantt Chart displaying the schedule, milestones, and dependencies throughout the timeline

The gantt chart in fig. 5.1 shows a timeline that runs from May 2023 to March 2024 where a sequence of tasks is plotted against the given timeline. The tasks listed on the left side include the stages of our research: idea selection, previous work study, idea proposing, dataset collection and analysis, top and easily accessible feature selections, model implementations and deploying app. With some tasks overlapping at the same time, indicating concurrent activity, the blue highlights indicate the time frame taken to complete tasks.

Chapter 6

Methodology

The methodology of this study includes machine learning and deep learning models to select important features and predict Polycystic Ovary Syndrome(PCOS) using only easily accessible non-invasive health features of patients. The whole methodology is executed into several significant steps. Following the initial steps of data collection from Kaggle, the study employs meticulous data pre-processing techniques to ensure the dataset's quality and relevance. This involves handling missing values and standardizing or normalizing data as needed. Then in the dataset feature extraction is performed. Chi-square and extra tree classifier feature extraction method is used to extract important features. This process helps filter out noise and irrelevant information, focusing solely on the factors that significantly contribute to the prediction of PCOS. Notably, the study places emphasis on non-invasive features, aligning with the goal of developing a predictive model that relies solely on easily accessible and non-intrusive body features of patients. Having identified the key features, the dataset is then partitioned into training and testing sets. This division is fundamental for evaluating the model's performance on unseen data, providing insights into its generalization capabilities. The study employs a range of machine learning and deep learning models to train on the designated training dataset. Among them Random forest classifier, Gradient Boosting, AdaBoost and Support Vector classifier from machine learning model, along with LSTM and MLP from deep learning model outperformed the rest of the models. The evaluation process involves assessing various performance metrics such as accuracy, precision, recall, and F1 score to comprehensively gauge the effectiveness of each model.

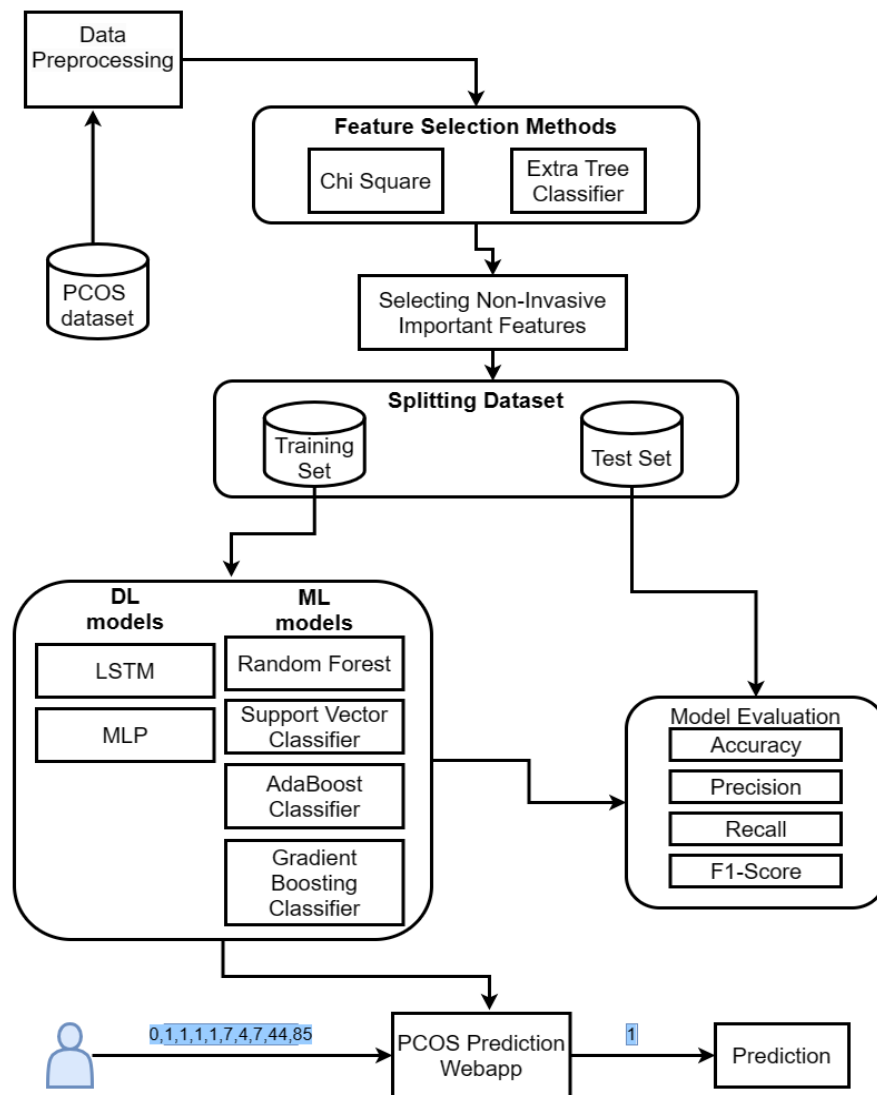


Figure 6.1: Methodology of A Non-invasive Approach of PCOS Prediction using Machine Learning

6.1 Data Pre-processing

Data preprocessing, a process of changing the raw data into a suitable data for machine learning model. The following mentioned steps are performed to preprocess the dataset: The Unnecessary Columns (Serial Number and Patient File Number) are removed and checked for any duplicate data, if it exists then removed. The 'NaN' values are filled with the mean value of the respective feature. The dataset is split into features and a target label, which indicates whether or not PCOS has been detected (0,1). The features are scaled.

6.2 Feature Selection

6.2.1 Chi-Square

After applying Chi-Square on the dataset, the score of each feature is acquired. With this score, the importance of each feature is visualized. These are the top 20 features with their corresponding score.

Table 6.1: Top 20 Features by Chi-Square

Rank	Features	Score	Rank	Features	Score
1	Vit D3 (ng/ml)	9477.648	11	Hair Growth (Y/N)	84.854
2	I beta-HCG (mIU/mL)	6950.525	12	Weight Gain (Y/N)	65.554
3	LH (mIU/mL)	2558.471	13	Weight (Kg)	49.466
4	FSH (mIU/mL)	1601.145	14	Fast Food (Y/N)	37.721
5	II beta-HCG (mIU/mL)	949.362	15	Cycle (R/I)	27.681
6	Follicle No. (R)	672.789	16	PRG (ng/mL)	24.638
7	Follicle No. (L)	573.647	17	Pimples(Y/N)	22.587
8	AMH (ng/mL)	232.516	18	Marriage Status (Yrs)	22.181
9	FSH/LH	96.831	19	BMI	14.568
10	Skin Darkening (Y/N)	84.870	20	Age (yrs)	14.284

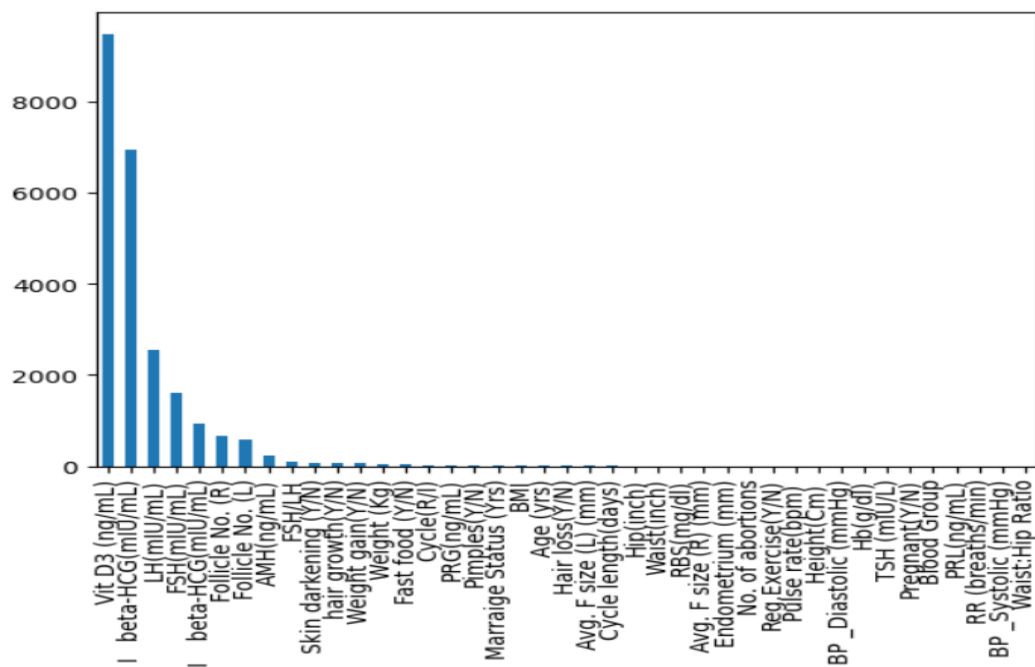


Figure 6.2: Ranking of features based on Chi-square method

All the features are plotted in ascending order based on the score obtained from the Chi-Square method. There X-axis represents the lists of the features from the dataset and Y-axis represents the feature importance scores.

These are the top 10 easily accessible non invasive health features that can be measured without any medical test and surgery:

- | | |
|-------------------|-------------------|
| 1. Skin Darkening | 6. Cycle |
| 2. Hair Growth | 7. Pimples |
| 3. Weight Gain | 8. Marital Status |
| 4. Weight | 9. BMI |
| 5. Fast Food | 10. Age(yrs) |

6.2.2 Extra Tree Classifier

Applying Extra Tree Classifier, the score of each feature is acquired. With this score, the importance of each feature is visualized. These are the top 20 features with their corresponding score:

Table 6.2: Top 20 Features by Extra Tree Classifier

Rank	Features	Score	Rank	Features	Score
1	Follicle No. (R)	0.111	11	Hip(inch)	0.0165
2	Follicle No. (L)	0.098	12	Marriage Status (Yrs)	0.0161
3	Skin Darkening (Y/N)	0.090	13	Avg. F size (L) (mm)	0.015
4	Hair Growth (Y/N)	0.084	14	Weight (Kg)	0.015
5	Weight Gain (Y/N)	0.071	15	Age (yrs)	0.0149
6	Cycle (R/I)	0.048	16	Avg. F size (R) (mm)	0.0148
7	Fast Food (Y/N)	0.0462	17	LH(mIU/mL)	0.0146
8	Pimples(Y/N)	0.025	18	Reg.Exercise(Y/N)	0.0145
9	Cycle length(days)	0.022	19	Waist(inch)	0.0141
10	AMH(ng/mL)	0.018	20	Height(Cm)	0.0130

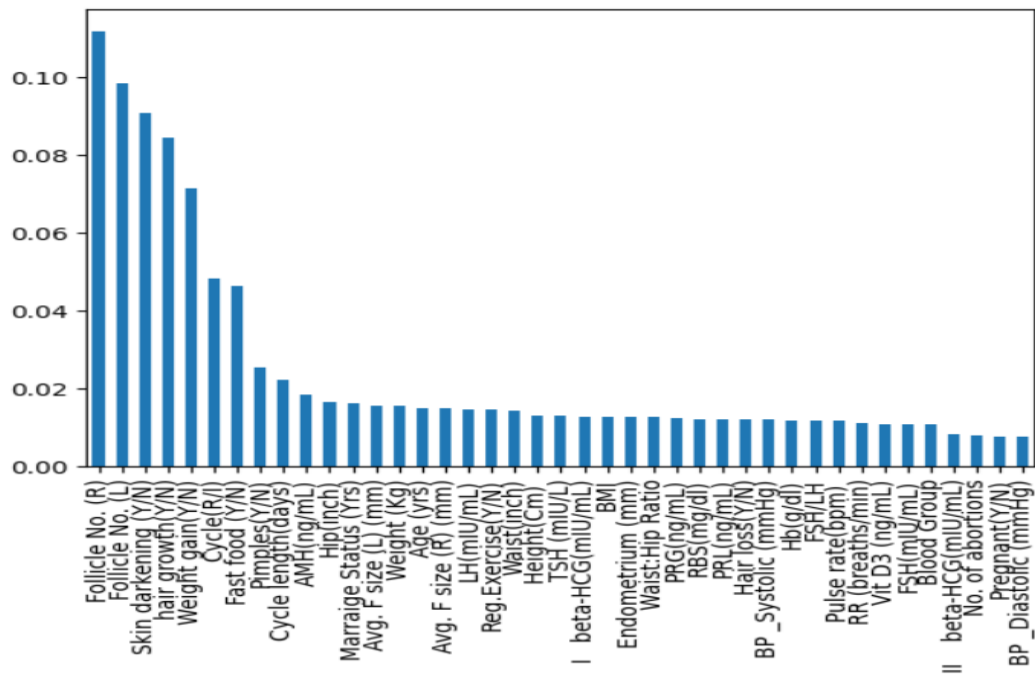


Figure 6.3: Ranking of features based on Extra tree classifier method

All the features are plotted in ascending order based on the score obtained from the extra tree classifier method. There X-axis represents the lists of the features from the dataset and Y-axis represents the feature importance scores in fig. 6.3.

The top 10 easily accessible non invasive health features that can be measured without any medical test and surgery:

- | | |
|-------------------|--------------------|
| 1. Hair Growth | 6. Pimples |
| 2. Skin Darkening | 7. Cycle length |
| 3. Weight Gain | 8. Hip (inch) |
| 4. Cycle | 9. Marriage Status |
| 5. Fast Food | 10. Weight |

6.3 Used Model

6.3.1 Machine Learning Models

6.3.1.1 SVC Model

SVC works by identifying a hyperplane in the feature space that maximally separates the data points of different classes. Support vectors are the data points closest to the decision boundary (hyperplane). These support vectors play a crucial role in determining the decision boundary and are used to define the hyperplane mathematically.

A Linear Support Vector Machine (SVM) model using the SVC implementation from the `sklearn.svm` module is trained to handle this binary classification problem. The important non-invasive features by using chi-square and extra tree classifier are extracted at first. Then the SVC model is trained on the feature vectors of the training set and tested on the test set with the optimal parameter achieved by Grid Search.

The set of optimal parameters are mentioned below individually for both of the feature selection method.

Table 6.3: Optimal parameter for SVC

Parameter Name	Parameter Value	
	Chi-Square	Extra Tree Classifier
c	1	0.1
gamma	1	1
kernel	linear	linear

6.3.1.2 Random Forest Classification

In a random forest classification, multiple decision trees are generated by taking random subsets of the data and features. Each decision tree provides a result on classification of the data. Predictions are made by calculating the result for each decision tree, then taking the most popular result. Multiple decision trees are generated by the subsets of top extracted features and spilled train data and results are generated for each decision tree. Averaging values of all the results is the predicted class. The model is trained with the option parameter which is achieved by Grid search.

The set of optimal parameters are mentioned below individually for both of the feature selection method.

Table 6.4: Optimal parameters of Random Forest

Parameter Name	Parameter Value	
	Chi-Square	Extra Tree Classifier
max_depth	None	8
max_features	0.2	0.2
max_samples	0.5	0.75
n_estimators	100	50
random_state	10	10

6.3.1.3 Gradient Boosting Classification

Gradient Boosting Classification is an ensemble model where multiple weak models are trained to get the best output. It's a classification model where Gradient descent is used to train each new model to minimize the loss function. We imported the sklearn.ensemble library where we have the GB classification model. In the next step, we trained the GB Classifier model on our training dataset. To train the model, we have used grid search. Then the accuracy is evaluated with the optimal parameters which is achieved by grid search.

The set of optimal parameters are mentioned below individually for both of the feature selection method.

Table 6.5: Optimal parameter of Gradient Boosting Classification

Parameter Name	Parameter Value	
	Chi-Square	Extra Tree Classifier
learning_rate	0.1	0.075
max_depth	2	2
max_features	2	2
n_estimators	40	40

6.3.1.4 AdaBoost Classification

AdaBoost Classification is also an ensemble model where it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. To run the model sklearn.ensemble library is imported where AdaBoostClassifier model is utilized. In the next step, the AdaBoost Classifier model is trained with the training dataset. The model is trained using the optimal parameters obtained from grid search.

The set of optimal parameters are mentioned below individually for both of the feature selection methods.

Table 6.6: Optimal parameter of AdaBoost Classification

Parameter Name	Parameter Value	
	Chi-Square	Extra Tree Classifier
learning_rate	0.075	0.05
n_estimators	300	100
random_state	20	20

6.3.2 Deep Learning Models

6.3.2.1 Long-Short Term Memory(LSTM)

The Long Short-Term Memory (LSTM) Deep Learning Model was constructed to address the classification task. The model architecture consisted of an LSTM layer with 64 units, followed by a dense layer with a sigmoid activation function for binary classification. The model was compiled using the Adam optimizer and binary cross-entropy loss function. The validation loss and accuracy curves were plotted to analyze how the model was learning with each epoch. Training of the model was conducted by fitting it to the training dataset of 10 features by using optimal number of epoch and batch size. Subsequently, the model's accuracy was evaluated using the test dataset. Moreover, a comparison study has been done between Chi-Square method and Extra Tree Classifier method by running the built model for two separate dataframes containing both Chi-square selected 10 features and Extra tree classifier selected 10 features. Optimal hyperparameters are mentioned in table 6.7

Table 6.7: Optimal parameter of LSTM model

Parameter Name	Parameter Value	
	Chi-Square	Extra Tree Classifier
epoch	10	8
batch_size	32	32
units	64	64
activation	sigmoid	sigmoid
optimizer	adam	adam
loss	binary_crossentropy	binary_crossentropy

6.3.2.2 Multilayer Perceptron(MLP)

The Multi-Layer Perceptron (MLP) Deep Learning Model was employed for the classification task. The model architecture comprised multiple dense layers with rectified linear unit (ReLU) activation functions, followed by a final dense layer with a sigmoid activation function for binary classification. The model was compiled using the Adam optimizer and binary cross-entropy loss function. The training process involved fitting the model to the training dataset for a total of 10 epochs with a batch size of 32. The performance of the model was evaluated using the validation dataset, and the validation loss and accuracy were plotted over the epochs to monitor the training progress. Additionally, the model's accuracy was assessed using the test dataset. Furthermore, the F1 score, precision, and recall metrics were computed to provide a comprehensive evaluation of the model's performance in terms of its ability to correctly classify instances from the test dataset. Through trial and error hyperparameters are tuned into optimal one as its mentioned in table 6.8

Table 6.8: Optimal parameter of MLP model

Parameter Name	Parameter Value
	Chi-Square & Extra Tree Classifier
epoch	10
batch_size	32
units	64
activation	relu
optimizer	adam
loss	binary crossentropy

6.4 Web-app Demostation

The PCOS Prediction web application is designed to predict the likelihood of Polycystic Ovary Syndrome (PCOS) based on 10 key features selected through an Extra Trees Classifier method. These features include weight gain, excess hair growth, acne problems, fast food consumption frequency, skin darkening, menstrual cycle length (in days), menstrual cycle regularity, years of marriage, hip size, and weight. The LSTM (Long Short-Term Memory) model, which was previously trained and saved, is employed to make predictions using the provided feature inputs.

Upon accessing the web application, users are presented with sliders for each feature, allowing them to input their relevant information easily. Once all features are selected, users can prediction by clicking the "PCOS Test Result" button. The input data is then transformed into a numpy array, reshaped to fit the model's input requirements, and fed into the LSTM

model for prediction. The prediction outcome is displayed to the user indicating whether they are prone to PCOS or not. If the predicted probability class is 1 the application flags the individual as PCOS-prone. Conversely, if the predicted probability class is 0, the application indicates that the individual is not prone to PCOS.

The PCOS Prediction web application is developed using Streamlit, a popular Python library for building interactive web applications with ease. Streamlit provides intuitive tools and components for creating data-driven applications, allowing seamless integration of machine learning models, visualizations, and user input elements.

Chapter 7

Result

7.1 Result Analysis

This section presents the results of the application of machine learning and deep learning algorithms on the PCOS dataset. Using Chi square and Extra tree classifier method, the ranking of all the 43 features are determined. For finding the optimum number of easy accessible features many trial has been attempted with different number of features. For 10 easily accessible non-invasive features many classifiers provided a better score.

With these conditions, the classification accuracy obtained by machine learning models including Random Forest, Support Vector Machine, Adaptive Boosting (AdaBoost), and Gradient Boosting is 86.4%, 84.5%, 83.8%, and 85.2% respectively using Chi-square method selected top 10 accessible features and 86%, 86.02%, 83.8%,82.3% respectively using Extra Tree classifier selected top 10 accessible features.

Table 7.1: Result analysis for top 10 easily accessible noninvasive features by ML

Serial No.	Machine Learning Models	Feature Extraction Model	Evaluation Matrices			
			Accuracy	Precision	Recall	F1 Score
1	SVC	Chi-Square	84.5%	74.4%	79.5%	76.9%
		Extra Tree Classifier	86.02%	77.7%	79.5%	78.6%
2	Random Forest	Chi-Square	86.4%	76.19%	72.7%	74.4%
		Extra Tree Classifier	86%	80%	75%	77.64%
3	Gradient Boosting	Chi-Square	85.2%	78.5%	75%	76.7%
		Extra Tree Classifier	82.3%	75%	68.18%	71.4%
4	Ada Boost	Chi-Square	83.8%	76.19%	72.7%	74.4%
		Extra Tree Classifier	83.8%	78.9%	68.18.7%	73.17%

When top 10 easily accessible non-invasive features are taken, then the highest accuracy comes from Random Forest Classifier with feature extraction Chi-Square method. The accuracy is 86.4%. Though there is slightly difference of accuracy for both feature extraction method in Random Forest Classifier.

The accuracy obtained by deep learning models including LSTM is 89.91%, 83% respectively. It can be seen that good accuracy scores are obtained by LSTM using Extra Tree Classifier feature extraction method.

Table 7.2: Result analysis for top 10 easily accessible noninvasive features by Deep Learning

Serial No.	Deep Learning Models	Feature Extraction Model	Evaluation Matrices			
			Accuracy	Precision	Recall	F1 Score
1	LSTM	Chi-Square	87.16%	78.12%	78.12%	78.12%
		Extra Tree Classifier	89.91%	86.21%	78.12%	81.97%
2	MLP	Chi-Square	83%	70.21%	78.57%	74.16%
		Extra Tree Classifier	74.26%	81.82%	21.43%	33.96%

When top 10 easily accessible non-invasive features are taken, then the highest accuracy comes from LSTM with feature extraction Extra Tree classifier method. The accuracy is 89.91%.

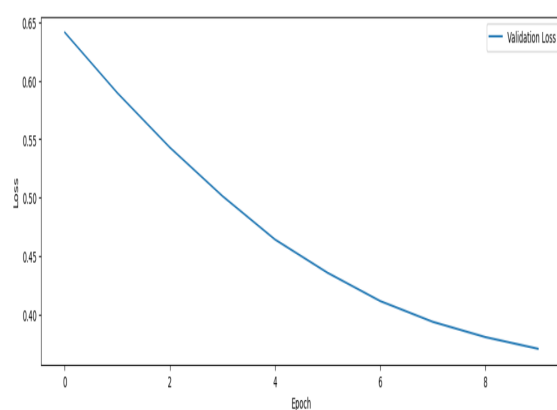
It can be concluded that the Deep Learning approach yields superior performance for this dataset, with LSTM emerging as the preferred model within this approach.

7.2 Graph Analysis

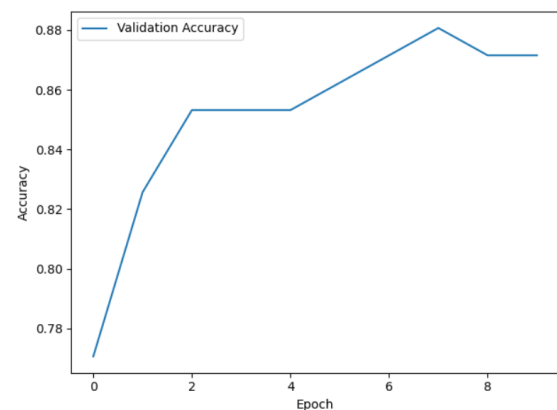
Since LSTM in deep learning models achieves higher accuracy, graph is used to demonstrate its performance by plotting the data. In this model, Extra Tree Classifier is used to extract top non invasive features. For 8 epoch and 32 batch size better accuracy is achieved.

Whereas In MLP chi-square method selected 10 features gave better result when the batch-size and epoch number was 32 and 10. In every epoch the validation loss decreases. This is the opposite in the case of validation accuracy. After reaching a certain epoch, the validation accuracy experiences a sudden decrease, prompting the selection of the desired epoch.

The validation loss and validation accuracy graph is shown below.

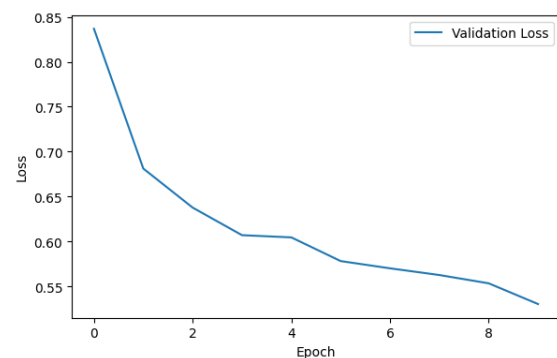


(a) Validation Loss for LSTM

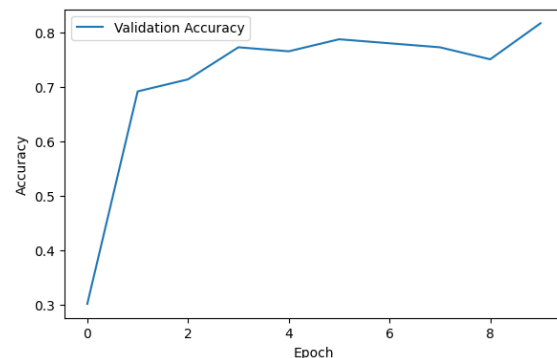


(b) Validation Accuracy for LSTM

Figure 7.1: Validation graph for LSTM



(a) Validation Loss for MLP

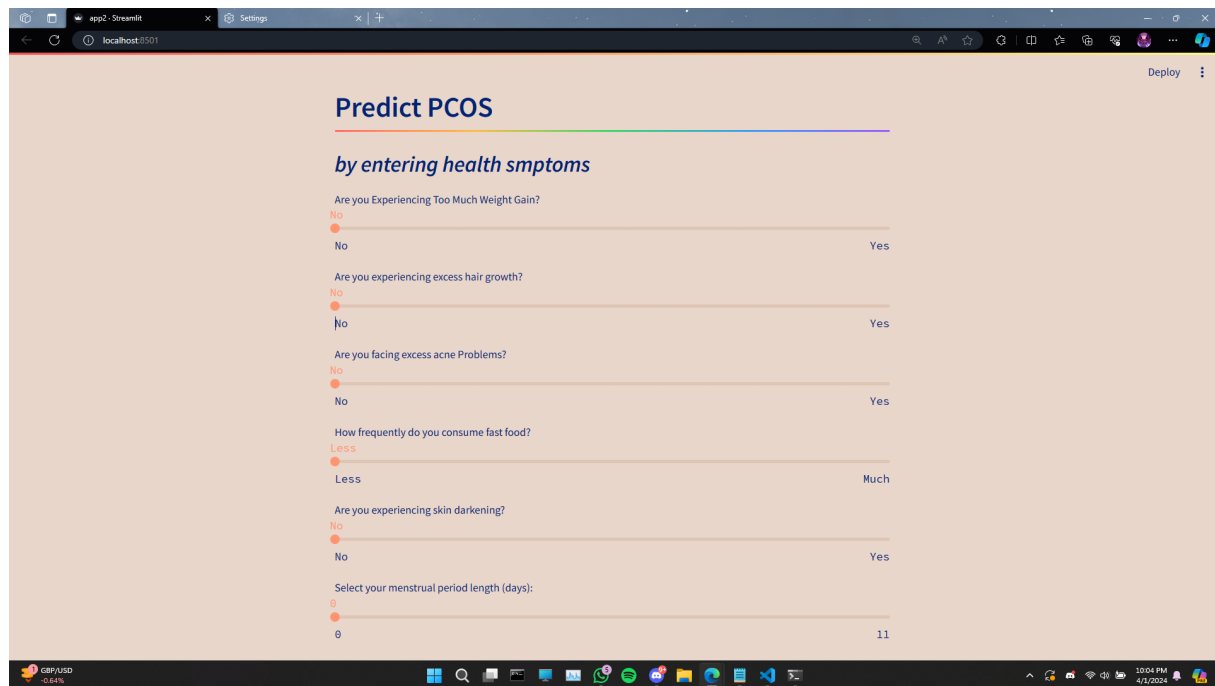


(b) Validation Accuracy for MLP

Figure 7.2: Validation graph for MLP

7.3 Model implementation on Web-app

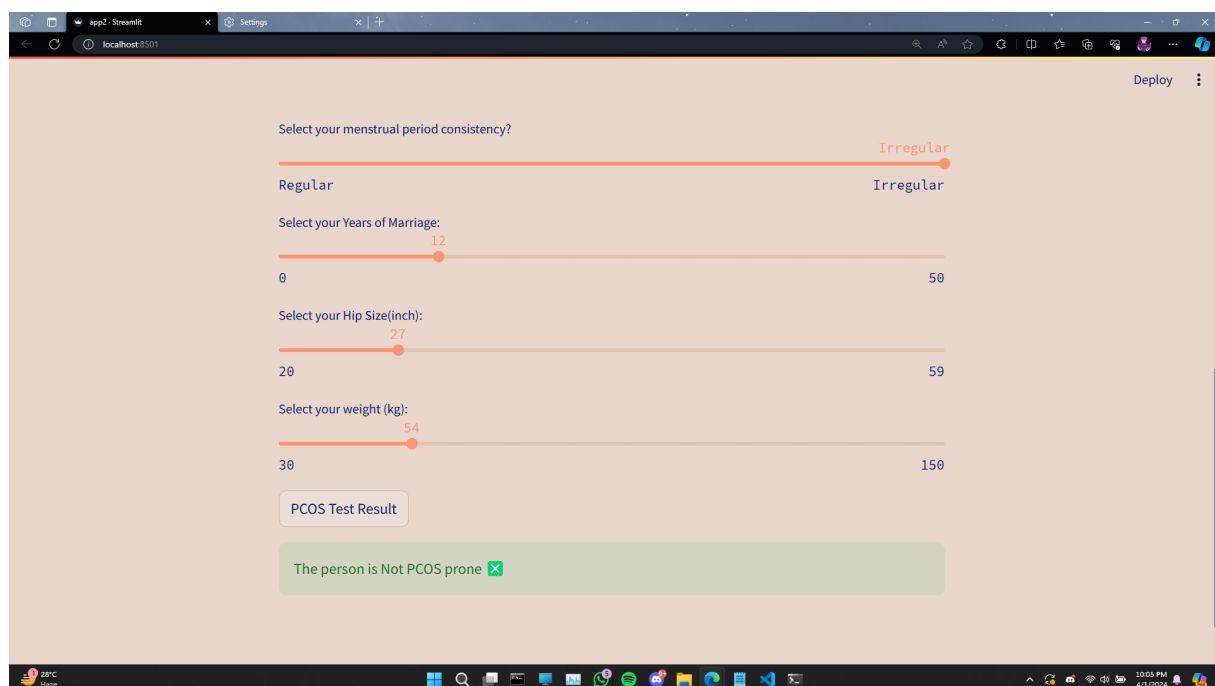
The integrated model in the app quickly analyzes the data to predict if they might have Polycystic Ovary Syndrome (PCOS) by taking user inputs on 10 features. Considering the features value, the model offers calculated predictions into the chances of having PCOS.



The screenshot shows a web browser window with the URL 'localhost:5501'. The page is titled 'Predict PCOS' and has a subtitle 'by entering health symptoms'. It contains several sliders for user input:

- 'Are you Experiencing Too Much Weight Gain?' with a slider between 'No' and 'Yes'.
- 'Are you experiencing excess hair growth?' with a slider between 'No' and 'Yes'.
- 'Are you facing excess acne Problems?' with a slider between 'No' and 'Yes'.
- 'How frequently do you consume fast food?' with a slider between 'Less' and 'Much'.
- 'Are you experiencing skin darkening?' with a slider between 'No' and 'Yes'.
- 'Select your menstrual period length (days):' with a slider between '0' and '11'.

Figure 7.3: User input of health features



The screenshot shows the same web browser window, but now with the final prediction result. The sliders are set to the following values:

- 'Select your menstrual period consistency?' is set to 'Irregular'.
- 'Select your Years of Marriage:' is set to '12'.
- 'Select your Hip Size(inch):' is set to '27'.
- 'Select your weight (kg):' is set to '54'.

Below the sliders, there is a button labeled 'PCOS Test Result'. The result is displayed in a green box: 'The person is Not PCOS prone' with a green checkmark icon.

Figure 7.4: Successful Prediction of PCOS class

The screenshot shows a web application running on a browser at localhost:5501. The application has a light beige background and a dark blue header with a 'Deploy' button. The main content area contains four sliders for inputting personal data:

- Select your menstrual period consistency:** A slider between 'Regular' and 'Irregular'. The 'Regular' end is highlighted in orange, and a red dot is positioned at the 'Regular' end.
- Select your Years of Marriage:** A slider between 0 and 50. A red dot is positioned at 7.
- Select your Hip Size(inch):** A slider between 20 and 59. A red dot is positioned at 27.
- Select your weight (kg):** A slider between 30 and 150. A red dot is positioned at 54.

Below the sliders is a button labeled 'PCOS Test Result'. Below the button is a red box containing the text 'The person is PCOS prone' followed by a red dot.

Figure 7.5: Successful Prediction of Non-PCOS class

Chapter 8

Conclusion and Future Work

8.1 Future Work

1. **Development of a Mobile App:** In the future, we will integrate the machine learning model with the Mobile App. So, any woman with medical concerns can easily input their symptoms where invasive features will not be needed.
2. **Advance Feature:** Additionally, we will try to combine an image-processing method with the existing model. This advanced feature will help to analyze the report of ultrasound test report images to extract text features like follicle no. and others.
3. **Increase of Accuracy:** We have currently achieved 89.91% accuracy from the LSTM model. We will try to increase the accuracy.
4. **Dataset Extend:** Currently we are using a dataset from Kaggle. The best try will be given to extend the dataset.

8.2 Conclusion

Polycystic Ovary Syndrome(PCOS) is an endocrinological disorder that is very prevalent in recent times. A long time of undiagnosed and untreated PCOS can lead to infertility and cancer. So it is very important to detect this illness in the early stage. The main goal of this thesis is to make the PCOS diagnosis easier and more accessible. This study offered a technique in which any concerned individual can diagnose PCOS with easily measured symptoms. The twelve most important and non-invasive features are used for predicting PCOS. This thesis proposed many machine learning and deep learning models for predicting PCOS with 10 easily accessible non-invasive features. The best model is selected after examining and comparing several models using various performance measures. The result analysis shows that classification models such as Random Forest were able to give the best prediction result for ML and LSTM for Deep Learning. It can be concluded that the Deep Learning approach yields superior performance for this dataset, with LSTM emerging as the preferred model within this approach. This PCOS prediction technique may be able to help people with PCOS start their treatment in the early stages and save them from long-term medical issues. Overall, this research study can benefit the healthcare system, especially the women's healthcare system. For Bangladeshi women who are uncomfortable with traditional diagnosis methods, this proposed method can be very helpful.

References

- [1] Adaboost-Algorithm-A-Complete-Guide-For-Beginners. <https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners>. Accessed:2023-11-22.
- [2] Chi-Square Test. <https://byjus.com/maths/chi-square-test/>. Accessed:2023-11-22.
- [3] Evaluation Matrices. <https://www.geeksforgeeks.org/metrics-for-machine-learning-model/>. Accessed:2023-11-22.
- [4] Gradient-Boosting-For-Classification. <https://blog.paperspace.com/>. Accessed:2023-11-22.
- [5] Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis. <https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code/>. Accessed:2023-03-31.
- [6] Polycystic ovary syndrome. <https://my.clevelandclinic.org/health/diseases/8316-polycystic-ovary-syndrome-pcos>.
- [7] Polycystic ovary syndrome (PCOS) Dataset From Kaggle. <https://www.kaggle.com/datasets/shreyasvedpathak/pcos-dataset>.
- [8] Random Forest Algorithm. <https://www.javatpoint.com/machine-learning-random-forest-algorithm>. Accessed:2023-11-22.
- [9] SVC Diagram. <https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm>.
- [10] The A-Z guide to Support Vector Machine. <https://www.analyticsvidhya.com/blog/2021/06/>

- [support-vector-machine-better-understanding/](#).
Accessed:2023-11-22.
- [11] What is a Multilayer Perceptron (MLP) or a Feedforward Neural Network (FNN)? <https://aiml.com/what-is-a-multilayer-perceptron-mlp/>.
Accessed:2023-03-31.
- [12] Subrato Bharati, Prajoy Podder, and M. Rubaiyat Hossain Mondal. Diagnosis of polycystic ovary syndrome using machine learning algorithms. In *2020 IEEE Region 10 Symposium (TENSYP)*, pages 1486–1489, 2020.
- [13] Shubham Bhosale, L B Joshi, and Arun Shivsharan. Pcos (polycystic ovarian syndrome) detection using deep learning. 2022.
- [14] Amsy Denny, Anita Raj, Ashi Ashok, C Maneesh Ram, and Remya George. i-hope: Detection and prediction system for polycystic ovary syndrome (pcos) using machine learning techniques. In *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, pages 673–678, 2019.
- [15] R Dewi, Kang Adiwijaya, Untari Novia Wisesty, and Jondri. Classification of polycystic ovary based on ultrasound images using competitive neural network. *Journal of Physics: Conference Series*, 971:012005, 03 2018.
- [16] Hela Elmannai, Nora El-Rashidy, Ibrahim Mashal, Manal Alohal, Sara F Abd-el Ghany, Shaker El-Sappagh, and Hager Saleh. Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence. *Diagnostics*, 13:1506, 04 2023.
- [17] Xuan Hien Le, Hung Ho, Giha Lee, and Sungho Jung. Application of long short-term memory (lstm) neural network for flood forecasting. *Water*, 11:1387, 07 2019.
- [18] Aroni Saha Prapty and Tanzim Tamanna Shitu. An efficient decision tree establishment and performance analysis with different machine learning approaches on polycystic ovary syndrome. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5, 2020.
- [19] Bedy Purnama, Untari Novia Wisesti, Adiwijaya, Fhira Nhita, Andini Gayatri, and Titik Mutiah. A classification of polycystic ovary syndrome based on follicle detection of ultrasound images. In *2015 3rd International Conference on Information and Communication Technology (ICoICT)*, pages 396–401, 2015.
- [20] Sayma Suha and Muhammad Nazrul Islam. An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. *Scientific Reports*, 12, 10 2022.

-
- [21] Akanksha Tanwar, Anima Jain, and Anamika Chauhan. Accessible polycystic ovarian syndrome diagnosis using machine learning. In *2022 3rd International Conference for Emerging Technology (INCET)*, pages 1–6, 2022.
- [22] Sagar Yeruva, Indu Gurralla, Ramya Myakala, Nimmi Agarwal, Shriya Rapolu, and Junhua Ding. Know pcos. pages 533–546, 03 2023.