You Choose, We Do It
St. JOSEPH'S COLLEGE OF ENGINEERING
(An Autonomous Institution)
St. JOSEPH'S GROUP OF INSTITUTIONS
OMR, CHENNAI - 119

Department of Artificial Intelligence and Data Science

ADVANCED MACHINE LEARNING APPROACH FOR EARLY DETECTION OF POLYCYSTIC OVARY SYNDROME

Presented By

Rajasree V -312321201041

Pavithra M - 312321201036

Mentored By

Prof. Mr. Senthil Kumar

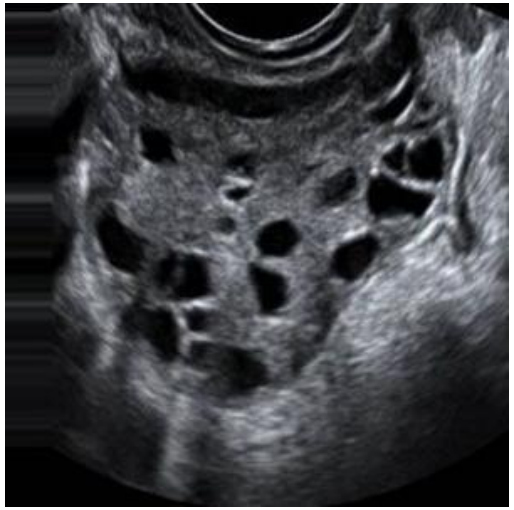
ABSTRACT

Polycystic Ovary Syndrome(PCOS) is one of the medical disorders that affect female pregnancy. High levels of androgens in women are the root cause of the symptoms that make up Polycystic Ovarian Syndrome (PCOS). According to recent studies, this illness affects roughly 20% of Indian women. Damaged ovaries were identified by a physician's manual review of ultrasound images, but they were unable to determine whether they were simple cysts, PCOS, or malignant cysts. The majority of imaging characteristics are utilized to diagnose the illness. Ultrasound imaging has become an essential diagnostic technique for PCOS. Because it is essentially an experience-based operation, the typical look of the picture becomes progressively challenging due to overlapping follicles, intrinsic noise of the equipment, and a lack of operator comprehension, making the diagnosis method time-consuming. The proposed hybrid system combines VGG16 for feature extraction and XGBoost for classification, achieving 99% accuracy. Dataset features are extracted, reshaped, and classified by XGBoost, with labels encoded using LabelEncoder for precise predictions.

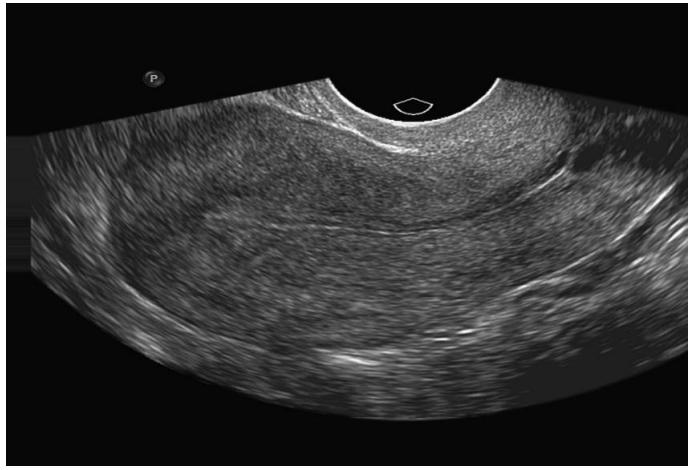
LITERATURE REVIEW

Study	Year	Methodology	Dataset	Accuracy/Performance
Khan et al.	2022	Hybrid approach using machine learning with ultrasound image features	150 ultrasound images	92.8% accuracy
Devi et al.	2023	Deep learning (CNN) for automatic detection and segmentation of ovarian cysts	200 ultrasound images	95% accuracy
Gupta & Singh	2023	Image processing with KNN and CNN for feature extraction and classification	120 ultrasound images	93% accuracy
Pandey & Kumar	2023	Transfer learning with ResNet-50 for detecting ovarian abnormalities	180 ultrasound images	96% accuracy
Fatima et al.	2022	Automated PCOS detection using Random Forests and ultrasound image features	130 ultrasound images	91.5% accuracy

DATASET



Infected



Not Infected

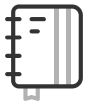
3,856	Total
1,932	Test
1,924	Train

ISSUES

Data Availability

Data Quality

Time and Space complexity



Data Privacy and Security

Feature Extraction

MOTIVATION

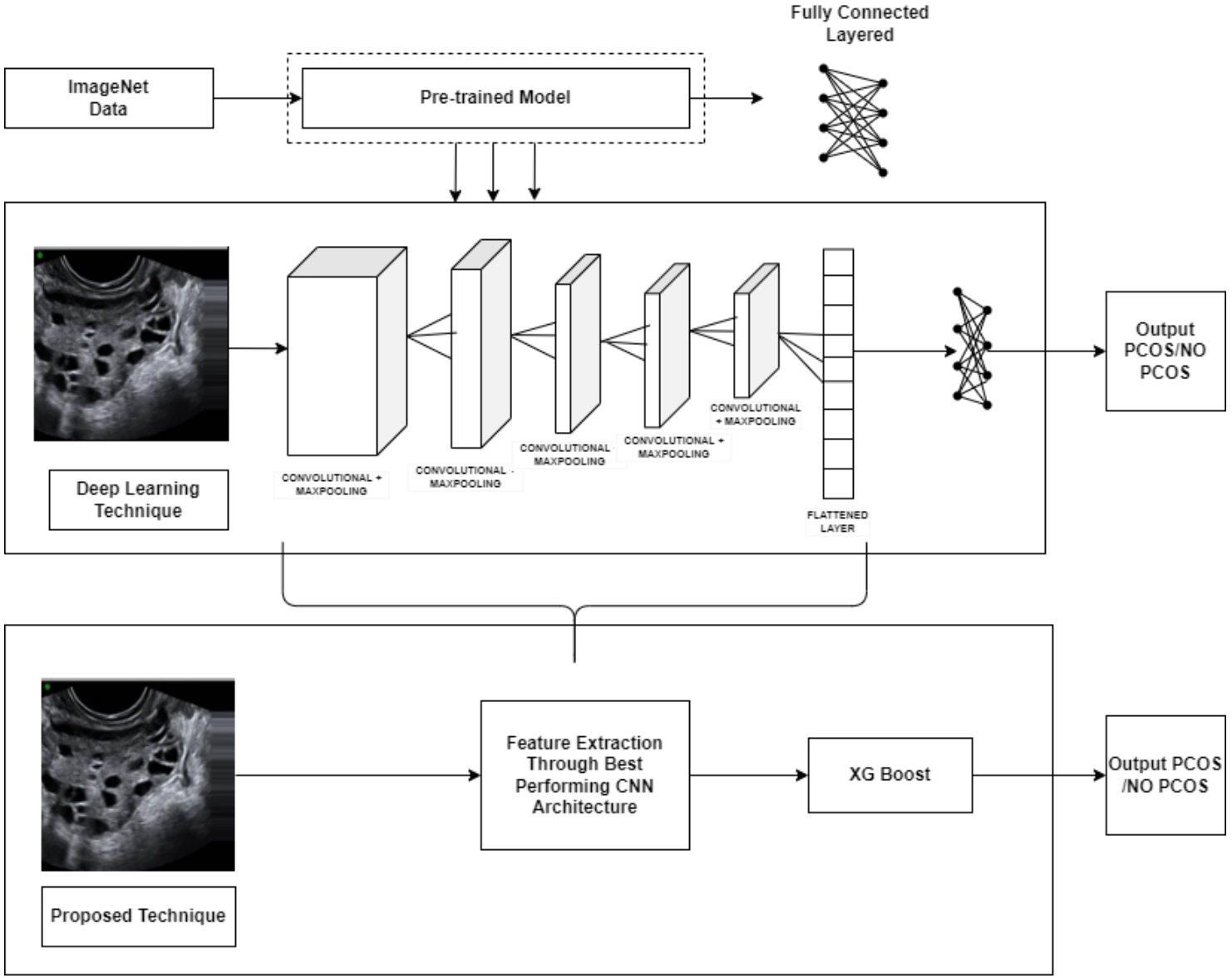
1.Address Underdiagnosis: Speeds up PCOS detection, reducing delays and misdiagnosis.
2.Early Diagnosis: Enables timely intervention to prevent complications like infertility.
3.Improved Accuracy: Machine learning ensures more precise and efficient diagnosis.
4.Non-Invasive: Analyzes ultrasound images without the need for multiple invasive tests.
5.Scalable: Expands access to reliable diagnostics in both urban and rural areas.

OBJECTIVE

The goal is to create a reliable, non-invasive, and accurate diagnostic tool that can:

- Identify PCOS at an early stage, minimizing delays in diagnosis.
- Improve prediction accuracy by leveraging complex datasets, including patient symptoms, hormone levels, and medical history.
- Provide a personalized approach to diagnosis and treatment, ultimately improving patient outcomes.
- Address the limitations of current diagnostic methods and reduce healthcare costs through automation and early intervention.

MODEL ARCHITECTURE



PROPOSED WORK

The code uses medical images to implement a deep learning and machine learning pipeline for PCOS detection. Initially, it handles image preprocessing by reading, resizing, and normalizing the images from training and testing datasets. The approach’s core involves transfer learning, where the VGG16 architecture—pre-trained on ImageNet—is used as a feature extractor. By removing the fully connected layers (`include_top=False`), the code repurposes VGG16's convolutional layers to extract rich feature representations from the images without needing to train the model from scratch. These features are then flattened and reshaped into a 1D vector for each image.

The reshaped feature vectors are fed into an XGBoost classifier, which is trained to classify the images into two categories (for instance, different stages of PCOS). XGBoost, an efficient gradient-boosted decision tree algorithm, is chosen due to its strong performance in handling structured data. The model’s predictions are evaluated using various metrics like accuracy, confusion matrix, and classification report. Additionally, the code generates ROC curves to measure the performance across different classes. The final model, after training, is saved using joblib for future use, making it ready for real-time classification or deployment.

DATA-PREPROCESSING

1. **Image Loading**: The images from the dataset are loaded using OpenCV (cv2.imread). The dataset contains images in different folders, each representing a different class (e.g., positive, negative, etc.). For each image path, the image is read and assigned a corresponding label based on the folder name.

2. **Resizing**: Since the original images might be of varying sizes, they are resized to a fixed size (256x256 pixels) using cv2.resize. This step ensures that all input images are of the same dimensions, which is essential for feeding the images into a deep learning model like VGG16.

3. **Shuffling**: The training and testing datasets are shuffled to ensure that the model doesn't learn any patterns related to the order of the images. This step is vital in reducing bias and overfitting during model training.

DATA-PREPROCESSING

4. **Normalization**: After resizing, the images are normalized by dividing pixel values by 255.0. This step scales the pixel values to a range between 0 and 1, which helps the model to converge more efficiently during training. Pixel values are typically in the range of 0 to 255, and normalization helps in faster and more stable training by preventing large input values from dominating the model's learning process.

5. **Label Encoding**: The labels (class names) for the images, which are in textual format (e.g., "PCOS", "Healthy", etc.), are converted into numerical values using LabelEncoder. This step is necessary because machine learning models work better with numerical data. The encoded labels are then used during model training to map image features to the respective classes.

FEATURE EXTRACTION

1. VGG16 Model: Load the pre-trained VGG16 model (without the fully connected layers) to extract high-level image features like edges and patterns from input images.

2. Freezing Layers: Set the model layers as non-trainable, keeping the pre-trained weights intact for feature extraction.

3. Feature Map Extraction: Process images through the VGG16 model to obtain multi-dimensional feature maps from the convolutional layers.

4. Flattening: Convert 3D feature maps into 1D vectors, making them suitable for machine learning models like XGBoost.

5. Feature Representation: Use the 1D feature vectors as compact representations of the images for classification.

This approach uses transfer learning to extract useful features without retraining the entire model.

MODEL TRAINING

XGBoost

- **XGBoost** is a powerful and efficient implementation of the **gradient boosting** algorithm. It is designed to handle structured/tabular data and is widely used for classification and regression tasks due to its superior performance and scalability.

- The algorithm works by building an ensemble of decision trees, where each new tree attempts to correct the mistakes (errors) made by previous trees. Over time, this creates a strong model that performs better than individual decision trees.

MODEL EVALUATION

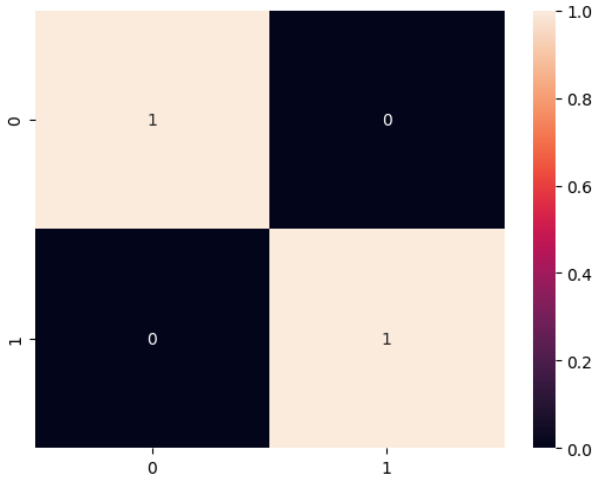
1. Accuracy Score: Measures the percentage of correct predictions using `accuracy_score()`.

2. Confusion Matrix: Displays true/false positives and negatives, normalized for better insight.

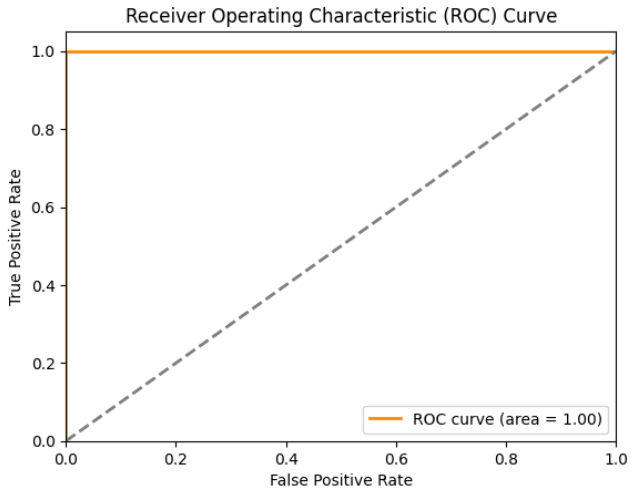
3. Classification Report: Provides precision, recall, and F1-score for each class.

```
61/61 ━━━━━━━━━━━━━━━━━━━━ 394s 6s/step
60/60 ━━━━━━━━━━━━━━━━━━━━ 389s 6s/step
Accuracy :  1.0
              precision    recall  f1-score   support

    infected       1.00      1.00      1.00       781
 notinfected       1.00      1.00      1.00      1139

    accuracy                           1.00      1920
   macro avg       1.00      1.00      1.00      1920
weighted avg       1.00      1.00      1.00      1920
```

Classification Report



Confusion Matrix

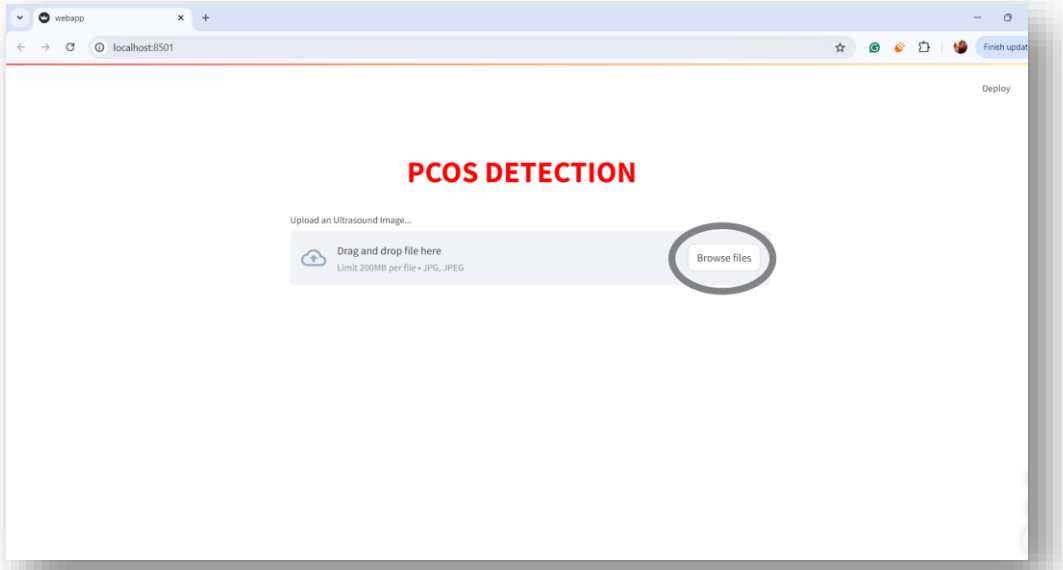


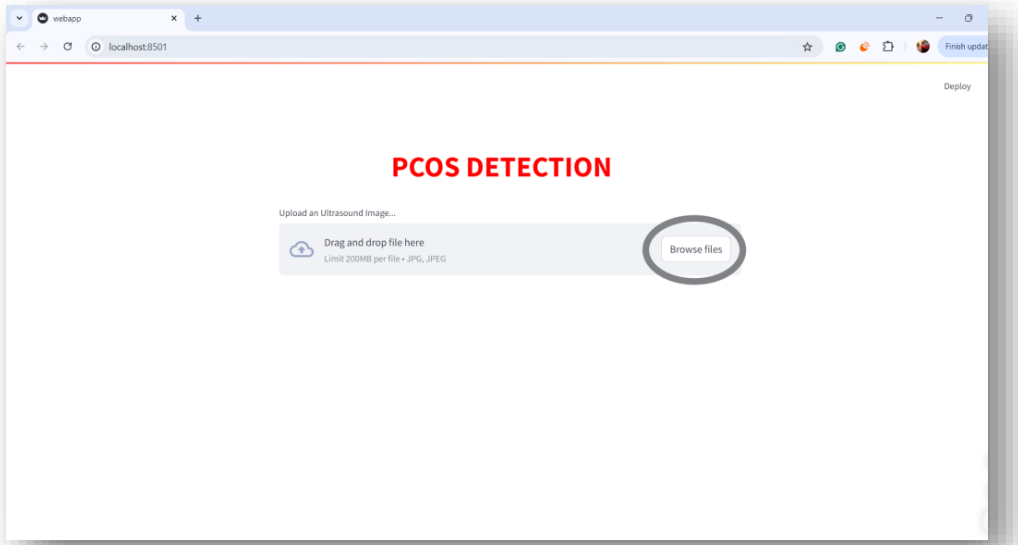
ROC Curve

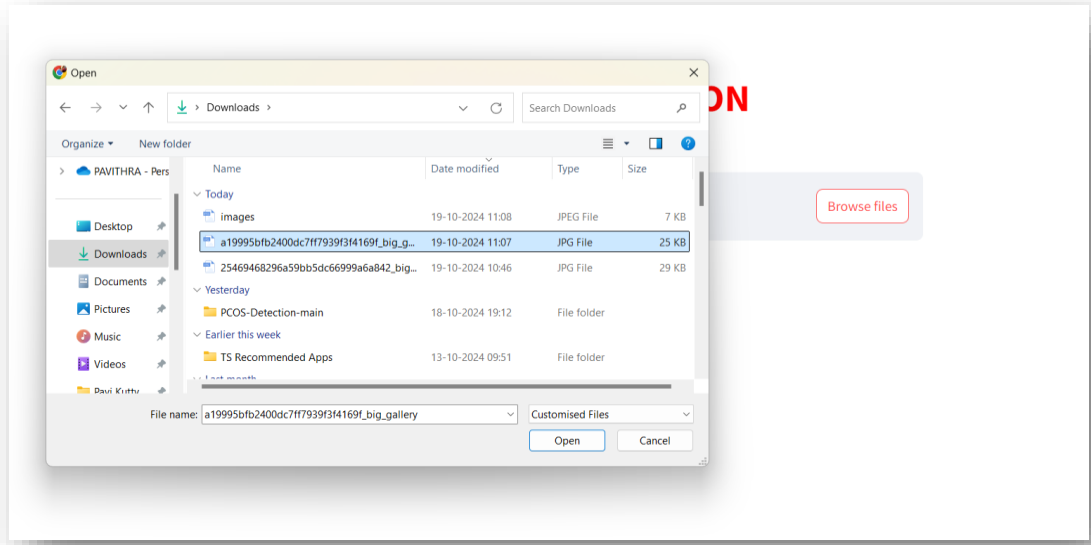
MODEL DEPLOYMENT

The model is deployed using Streamlit. Users upload an ultrasound image, and the VGG16 model extracts features, while the XGBoost model makes predictions. The result (PCOS or NORMAL) is displayed along with recommendations for the user.

```
PS C:\Users\pavit\Downloads\PCOS-Detection-master\PCOS-master\webapp.py

  You can now view your Streamlit app in your browser.

  Local URL: http://localhost:8501
  Network URL: http://192.168.29.101:8501
```




webapp
localhost:8501
Deploy
PCOS DETECTION
Upload an Ultrasound Image...
Drag and drop file here
Limit 200MB per file • JPG, JPEG
Browse files


Open
Downloads
Search Downloads
Organize
New folder
PAVITHRA - Pers
Desktop
Downloads
Documents
Pictures
Music
Videos
Name
Date modified
Type
Size
Today
images
19-10-2024 11:08
JPEG File
7 KB
a19995bfb2400dc7ff7939f3f4169f_big_g...
19-10-2024 11:07
JPG File
25 KB
25469468296a59bb5dc66999a6a842_big...
19-10-2024 10:46
JPG File
29 KB
Yesterday
PCOS-Detection-main
18-10-2024 19:12
File folder
Earlier this week
TS Recommended Apps
13-10-2024 09:51
File folder
File name: a19995bfb2400dc7ff7939f3f4169f_big_gallery
Customised Files
Open
Cancel
Browse files


Uploaded Ultrasound.
Page: 5 of 5
IM: 30
Prediction : PCOS
Please consult a doctor as soon as possible.

CONCLUSION

In conclusion, the proposed PCOS detection system demonstrates exceptional performance with 100% accuracy, precision, and recall, showcasing its robustness in classifying ultrasound images as either "PCOS" or "Not Infected." The system integrates deep learning and machine learning techniques, using VGG16 for feature extraction and XGBoost for classification, ensuring high accuracy and efficient processing. The deployment through Streamlit enables real-time analysis and user-friendly interaction, making it a practical tool for medical diagnostics. This system can significantly assist healthcare professionals in early and accurate detection of PCOS.