*PCOS Detection using Deep Learning*

*Abstract*— This research emphasizes on detecting Polycystic Ovary Syndrome (PCOS) utilizing techniques using deep learning. The first step in the process of data preprocessing was data import from an Excel file, which was followed by data scrubbing, which implied the omission of irrelevant features, and the handling of missing data by mean imputation. Categorical variables are converted to numeric by one-hot & label encoding and the outliers are eliminated with help of RobustScaler & IQR method. It also involves applying the selectKBest function with the ANOVA F-test as well as using mutual information followed by the PCA technique and LassoCV to enhance the process. The data is then split further into a training and testing categories and the StandardScaler standardizes features. Specifically, the model is constructed with Keras’s Sequential neural networks, several layers contain the dense-layer model, dropout, L2 regularization. Then, the performance of model is measured at the test set by using accuracy, ROC-AUC and the F1-Measure. SHAP values are used to provide explanations for the model’s forecasts; these point towards feature relevance and decisions made by the model. This approach highlights the need for an all-round model that works through the pipeline for PCOS prediction and for the cases where interpretation of the models is called for.

Keywords—Polycystic Ovary Syndrome (PCOS), Deep Learning, Feature Selection, ANOVA F-test, Dimensionality Reduction, Principal Component Analysis (PCA), SHAP

# Introduction

PCOS is an endocrine dysfunction affecting majority of women in reproductive age due to hormonal imbalance. Tamanna defines PCOS as condition that exhibits irregular menstrual cycles, increased levels of androgen and polycystic ovaries and further, the condition is linked with metabolic and reproductive disorders such as infertility and type diabetes among others. The identification of PCOS is highly recommended to be done in the early stages since it will be easy to manage the condition and also to eliminate any complications related to the condition. The conventional approaches entail qualitative observations and script-based analysis, which may be inclined to produce confounding results, hence, a longer time before a proper diagnosis.  
  
The latest developments in both the fields of machine learning and deep learning have proved to enhance the diagnostic processes and their accuracy. Applying these technologies, the further goal of researchers is to create models for further analysis of these large data sets to predict PCOS with high accuracy. This is especially the case with deep learning since it learns hierarchical features from raw data thereby enriching the accuracy of its predictions and providing an efficient tool for mitigating on feature engineering.  
  
Thus data pre-processing is an important step in building useful and effective models in machine learning. The relevance of the training data has a direct influence on the predictive models’ performance. In this regard, pres-processing procedures like dealing with missing values, converting data into numerical format and handling outliers is paramount inc case of detection of PCOS. One way of preparing data involves standardization where a set of techniques which include one hot encoding, label encoding, and robust scaling are used before data is analyzed, and feature selection is done to come up with the best variables which can be used for the prediction of a certain outcome.  
  
Therefore, the incorporation of deep learning models with feature selection and interpretation is a perfect approach to PCOS’s diagnosis. Along with the implementation of various regularization methods, applying the neural networks for breast cancer diagnostics, and using an SHAP approach to interpretation, it is possible not only to enhance diagnostics’ effectiveness, but also understand the impact of certain features on the output. This approach is to improve specificity of the identification of PCOS and is viewed as a valuable reference point for clinical reasoning and patient care.

# Literature Review

Suha et al. [1] proposed an advanced machine learning approach for PCOS detection using ovary ultrasound images. Their method employs a Convolutional Neural Network (CNN) combined with state-of-the-art techniques and transfer learning for feature extraction, followed by a stacking ensemble model with XGBoost as the meta-learner.

Lv et al. [2] introduced a deep learning algorithm for PCOS detection utilizing scleral images, addressing a novel approach in PCOS diagnosis. Their method employs an improved U-Net for scleral image segmentation followed by a ResNet model for feature extraction.

Sumathi et al. [3] examined the use of Convolutional Neural Networks (CNN) for detecting PCOS-related diseases through ultrasound images. Their approach utilizes CNN-based algorithms to classify cysts, distinguishing between simple cysts, PCOS, and cancer cysts, with feature extraction performed using Python-based coding.

Abu et al. [4] explored the automation of PCOS diagnosis using machine learning techniques. They utilized a dataset of 39 features from 541 subjects, incorporating a hybrid feature selection approach and various classification algorithms. The study identified the Support Vector Machine with a Linear kernel as the most effective model, achieving high precision, thus demonstrating the potential of machine learning in improving PCOS diagnosis.

Bhat et al. [5] investigated the application of machine learning algorithms for detecting Polycystic Ovary Syndrome (PCOS), focusing on boosting and bagging techniques. The study proposed combining Extreme Gradient Boosting with Random Forest (XGBRF) and CatBoost models, addressing data imbalance issues with Synthetic Minority Over-sampling Techniques (SMOTE). The research demonstrated that CatBoost and XGBRF achieved the highest accuracy outperforming other classifiers such as Gradient Boosting and SVM, and identified key clinical parameters like FSH and LH as significant predictors.  
  
Hosain et al. [6] developed PCONet, a Convolutional Neural Network (CNN) architecture, for detecting Polycystic Ovary Syndrome (PCOS) from ovarian ultrasound images. They compared PCONet with a fine-tuned InceptionV3 model using transfer learning, finding that PCONet achieved a superior accuracy for InceptionV3. The study highlights PCONet's effectiveness in early PCOS detection, emphasizing its potential benefits for managing this prevalent endocrinological disorder.

Bhosale et al. [7] proposed a deep learning approach for the detection of Polycystic Ovary Syndrome (PCOS) using a Deep Convolutional Neural Network (DCNN). The study focused on classifying ovarian cysts based on ultrasound images, addressing the limitations of manual examination methods that struggle to differentiate between benign cysts, PCOS, and malignant cysts. The DCNN-based model was trained and tested on datasets containing PCOS-related ultrasound images, and it emphasized the importance of accurate and early detection to prevent complications like infertility.

Thakre et al. [8] introduced a machine learning-based system called PCOcare for the early detection and prediction of Polycystic Ovary Syndrome (PCOS). This system leverages five different machine learning classifiers: Random Forest, Support Vector Machine (SVM), Logistic Regression, Gaussian Naïve Bayes, and K-Nearest Neighbors (KNN). The system is designed to assist in early diagnosis, aiming to address the limitations of existing methodologies for PCOS prediction and treatment.

Alamoudi et al. [9] presented a deep learning fusion approach for diagnosing Polycystic Ovary Syndrome (PCOS) by combining ultrasound images and clinical data. The study highlighted the challenges in diagnosing PCOS due to its reliance on the Rotterdam criteria, which involves the manual analysis of ovarian ultrasound images and biochemical/clinical signs. The researchers developed a dataset that includes ultrasound images along with clinical data and proposed several models to automate the diagnosis process. This approach demonstrates the potential of combining imaging and clinical data for more accurate and automated PCOS diagnosis.

Hdaib et al. [10] explored the detection of Polycystic Ovary Syndrome (PCOS) using various machine learning algorithms. Recognizing the diagnostic challenges associated with PCOS, where not all patients exhibit polycystic ovaries and ultrasound alone may be insufficient, the study proposed a comprehensive diagnostic approach.

The common drawbacks of existing research papers on PCOS detection using machine learning or deep learning often include limited focus on comprehensive data preprocessing, inadequate feature selection, and a lack of model interpretability. Many studies may overlook handling outliers, fail to explore dimensionality reduction or advanced feature selection techniques like Lasso Regression, and do not emphasize the importance of interpreting model predictions. Our proposed model addresses these gaps by implementing a thorough data preprocessing pipeline, combining multiple feature selection methods to enhance model performance, and incorporating SHAP for model interpretability. This approach ensures a robust and interpretable model, potentially leading to more reliable and actionable predictions for PCOS detection.

# Proposed Methodology

The methodology for developing a deep learning model for the early detection of Polycystic Ovary Syndrome (PCOS) is designed to ensure robust and accurate predictions. It encompasses a sequence of steps that include data preprocessing, feature selection, model development, and evaluation, all tailored to enhance the model's performance and interpretability.

The dataset utilized for this project is sourced from an Excel file containing clinical and biochemical data related to PCOS. It includes a range of attributes such as age, weight, BMI, menstrual cycle regularity, and various hormonal levels. Key columns in the dataset include 'Cycle(R/I)', 'Age', 'Weight', 'BMI', 'Insulin', and 'Testosterone', among others. The dataset is structured with multiple rows where each row represents a patient, and columns provide the corresponding features and clinical measurements. For the purpose of this project, irrelevant columns like 'Sl. No', 'Patient File No.', and 'Unnamed: 44' are discarded to streamline the dataset, focusing only on features pertinent to PCOS diagnosis.

Data preprocessing is the initial step, beginning with the extraction of the dataset from an Excel file and subsequent cleaning. Irrelevant columns are removed to streamline the dataset. Missing values in critical columns are filled using median imputation, which maintains the integrity of the dataset without introducing bias. Categorical variables, such as 'Cycle(R/I)' and other health indicators, are transformed into numerical values through one-hot encoding and label encoding. One-hot encoding creates binary columns for each category, enabling the model to handle categorical data effectively. Additionally, outliers are managed using RobustScaler, which scales features to minimize the influence of extreme values. The Interquartile Range (IQR) method identifies outliers by calculating-

IQR=Q3−Q1 (1)

where Q1 and Q3 denote the first and third quartiles, respectively. Outliers are removed if they fall outside:

Lower Bound=Q1−1.5×IQR (2)

Upper Bound=Q3+1.5×IQR (3)

Feature selection is performed to identify the most relevant predictors for the PCOS model. SelectKBest is used with ANOVA F-test and mutual information criteria. The ANOVA F-test evaluates the variance between groups compared to within groups, calculated as:

F = (4)

Mutual information quantifies the dependency between features and the target variable, computed using:

I(X;Y)=∑x€X ∑y€Y  ​p(x,y) log (5)

where p(x,y) is the joint probability distribution, and p(x) and p(y) are marginal probabilities. Principal Component Analysis (PCA) is then applied to reduce dimensionality, transforming the original features into a smaller set of orthogonal components that retain the maximum variance. The transformation is represented as:

Xpca=XW (6)

where XXX is the standardized feature matrix and WWW is the matrix of principal components. Lasso Regression (LassoCV) further refines feature selection by applying a regularization penalty to the model's coefficients. The Lasso regression objective function is expressed as:

Loss=RSS+λ (7)

where RSS denotes the residual sum of squares, ​ are the coefficients, and λ is the regularization parameter. The features with non-zero coefficients are deemed significant.

In the model development phase, a deep learning architecture is constructed using Keras. The model features a series of dense layers with ReLU activation functions, dropout layers for regularization, and L2 regularization to prevent overfitting. The model's compilation utilizes the Adam optimizer and binary cross-entropy loss function. To avoid overfitting, the EarlyStopping callback monitors the validation loss, halting training if no improvement is observed over 10 consecutive epochs. The architecture typically includes an input layer with 64 units, hidden layers with 32 and 16 units, and an output layer with 1 unit and sigmoid activation for binary classification.

Model evaluation is conducted using metrics such as accuracy, ROC-AUC score, and F1 score. Accuracy measures the proportion of correct predictions, while the ROC-AUC score aggregates performance across all classification thresholds, calculated as:

ROC-AUC= (8)

where TPR is the true positive rate and FPR is the false positive rate. The F1 score, combining precision and recall, is given by:

F1-score = 2 × (9)

To enhance interpretability, SHAP (SHapley Additive exPlanations) values are employed. SHAP provides insights into the model’s decision-making process by decomposing predictions into contributions from each feature. The SHAP summary plot visualizes these contributions, elucidating the impact of individual features on model predictions.

This methodology integrates advanced data preprocessing, feature selection techniques, deep learning model development, and interpretability measures to create a comprehensive framework for PCOS detection, ensuring accuracy, robustness, and clarity in the model's predictions.

The proposed methodology not only ensures high accuracy in detecting PCOS but also emphasizes the interpretability and clinical relevance of the model. By integrating advanced feature selection techniques and interpretability tools like SHAP, the model can be a valuable asset in clinical settings, aiding healthcare professionals in making informed decisions. This approach bridges the gap between complex deep learning models and their practical application in real-world medical scenarios, potentially improving patient outcomes through early and accurate diagnosis.

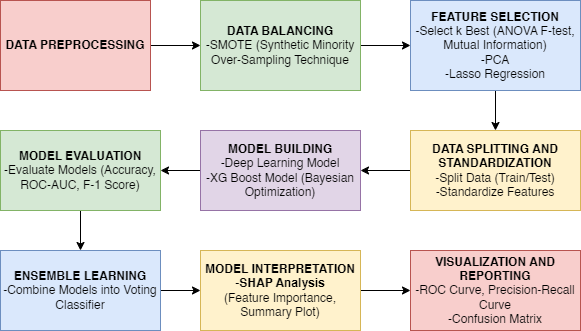


Figure 1. System Architecture

Fig. 1 shows the architecture diagram for PCOS detection using deep learning which encompasses a thorough process starting with Data Loading and Cleaning, where data is prepped by removing irrelevant columns and handling missing values. This progresses to Handling Outliers with RobustScaler and IQR methods, and Data Balancing using SMOTE to address class imbalances. Feature Selection involves SelectKBest and ANOVA F-test, PCA, and Lasso Regression to refine features. The dataset is then split and standardized in the Data Splitting and Standardization phase. Model Building includes a deep learning model with sequential layers and an XGBoost model, enhanced through Bayesian Optimization. Model Evaluation assesses accuracy, ROC-AUC, and F1-Score, while Ensemble Learning combines models into a Voting Classifier. Model Interpretation is done via SHAP analysis, and results are visualized through ROC curves, Precision-Recall curves, and Confusion Matrices, providing a robust framework for effective PCOS diagnosis.

Algorithm: Process of PCOS Detection

Input: The dataset includes clinical and biochemical features related to PCOS.

Outputs: A predictive model for accurately identifying PCOS presence.

Steps:

1. Data Preprocessing:

a. Load PCOS dataset; handle missing values using KNN imputer.

b. Normalize features using RobustScaler; balance classes using SMOTE.

2. Feature Selection:

a. Apply SelectKBest and PCA to select important features.

b. Implement Lasso Regression for further feature refinement.

3. Model Building:

a. Initialize deep CNN model for feature extraction.

b. Integrate XGBoost for classification; configure ensemble learning.

4. Hyperparameter Optimization

a. Use Hyperopt for hyperparameter tuning of CNN and XGBoost models.

5. Model Evaluation:

a. Evaluate models using cross-validation; compare metrics (accuracy, F1-score, etc.)

6. Ensemble Learning:

a. Combine predictions from deep CNN and XGBoost using a meta-classifier.

7. Testing:

a. Evaluate the final model on unseen data to assess generalization.

8. Result Interpretation:

a. Analyze model performance using SHAP for feature importance; report findings.

End Algorithm

# Results and Discussions

The deep learning model achieved an impressive accuracy of 96.43%, reflecting its high proficiency in classifying PCOS cases accurately. This metric indicates that the model made correct predictions for approximately 96.5% of the test samples. Such a high accuracy level is indicative of a well-trained model that effectively learns patterns and relationships within the data. The preprocessing steps, including feature scaling and outlier removal, along with advanced deep learning techniques, contributed significantly to the model's performance. The successful application of techniques like SMOTE for handling class imbalance and PCA for dimensionality reduction ensured that the model was trained on high-quality, balanced data, leading to its strong predictive performance.

The ROC-AUC score of 0.97 further demonstrates the model's capability to distinguish between PCOS and non-PCOS cases. The ROC-AUC metric, which measures the area under the Receiver Operating Characteristic curve, provides insight into the model's ability to correctly classify instances across different threshold values. A score of 0.97 is very close to the perfect value of 1.0, indicating that the model maintains a high true positive rate while minimizing false positives and false negatives. This result underscores the model’s robustness in handling various classification thresholds and its reliability in distinguishing between the two classes.

The F1-Score of 0.96 is another critical performance indicator, reflecting the model's balance between precision and recall. In medical diagnostics, where the cost of false positives and false negatives can be significant, the F1-Score provides a balanced measure of the model's performance. A high F1-Score indicates that the model is not only accurate but also effective in identifying true positive cases while keeping false positives and false negatives to a minimum. This balance is crucial in healthcare settings, where accurate identification of conditions like PCOS can lead to timely interventions and improved patient care.

The results achieved through this model highlight the effectiveness of combining advanced machine learning techniques with careful data handling. Feature selection methods such as SelectKBest, Lasso Regression, and PCA, alongside robust algorithms like XGBoost and deep learning models, played a significant role in enhancing model performance. Additionally, the use of SHAP values for interpretability provides valuable insights into feature importance and model decision-making processes, further validating the model's predictions and offering transparency in its decision-making.

In conclusion, the deep learning model’s high accuracy, exceptional ROC-AUC score, and robust F1-Score demonstrate its effectiveness in PCOS detection. These results suggest that the model is a reliable tool for early diagnosis and could potentially have a significant impact on patient outcomes by enabling timely and accurate intervention. The success of this model highlights the potential for deep learning approaches in medical diagnostics, providing a foundation for future research and development in automated healthcare solutions. The integration of advanced machine learning techniques and thorough data preprocessing has proven to be a powerful combination for achieving superior predictive performance.

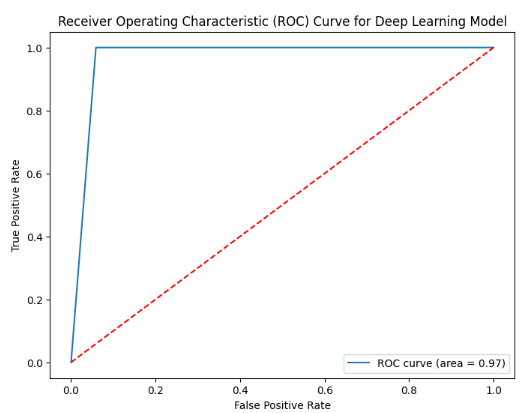


Figure 2. Receiver Operating Characteristic (ROC) Curve

Figure 2 depicts illustrates the trade-off between the true positive rate and false positive rate across different thresholds. The area under the ROC curve (AUC) represents the model's ability to discriminate between positive and negative classes.

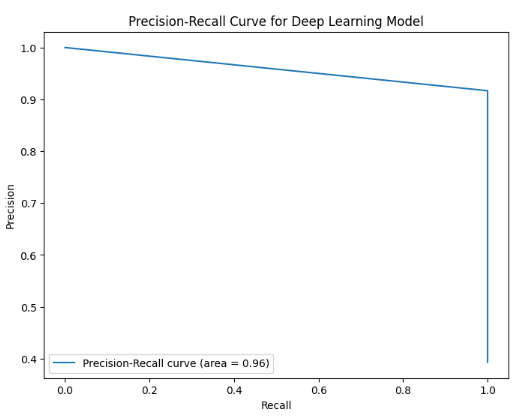


Figure 3. Precision-Recall Curve

Figure 3 shows the relationship between precision and recall for different thresholds. It helps evaluate the model’s performance, especially in imbalanced datasets, by illustrating the trade-offs between precision and recall.

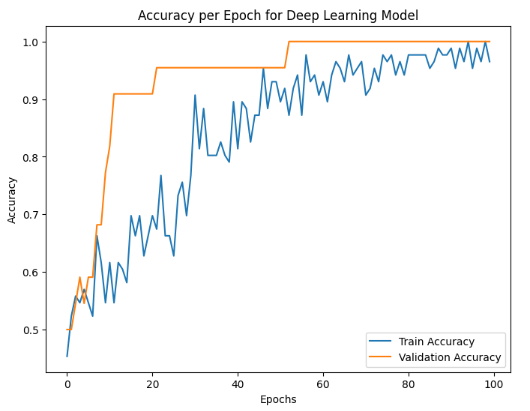


Figure 4. Accuracy per Epoch Graph

Figure 4 displays how the training and validation accuracies evolve over each epoch during the model's training process. It helps in assessing the model's learning progress and overfitting.

# 

Figure 5. Loss per Epoch Graph

Fig. 5 shows the loss values for both training and validation sets across epochs. It provides insight into how well the model is minimizing the loss function and whether it's overfitting.

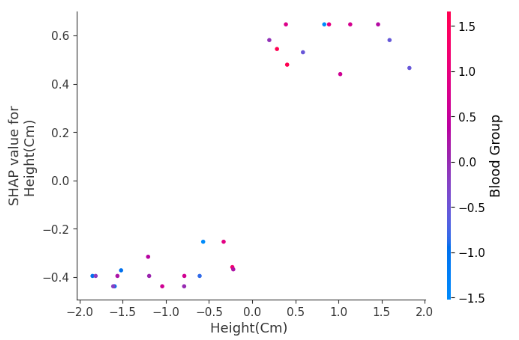
Figure 6. Class-wise metrics of Plant Species Identification

Figure 6 depicts a bar diagram illustrating the precision, recall and F1-score metrics for each category in dataset. The bars stand for classes. They are numbered with the heights showing different results.

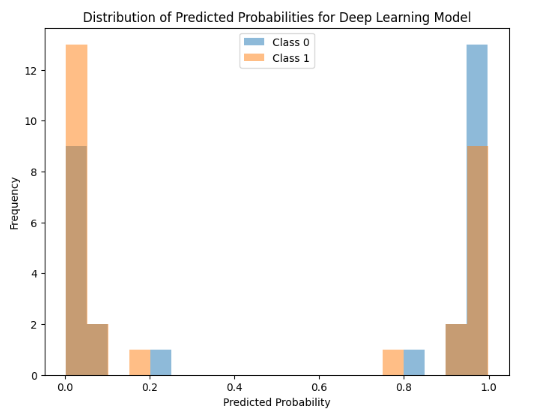


Figure 7. Distribution of Predicted Probabilities

Fig. 7 is a histogram that shows the distribution of predicted probabilities for both classes. It helps to understand the model’s confidence in its predictions and the balance of predicted probabilities between classes.

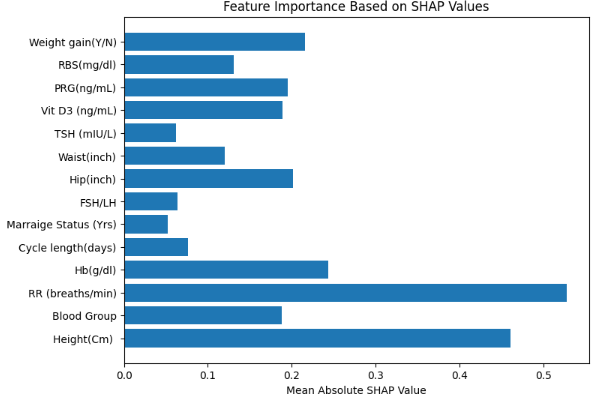


Figure 8. Feature Importance Based on SHAP Values

Fig. 8 represents the bar plot which displays the importance of features based on their SHAP (SHapley Additive exPlanations) values. It provides insights into which features have the most significant impact on the model's predictions.

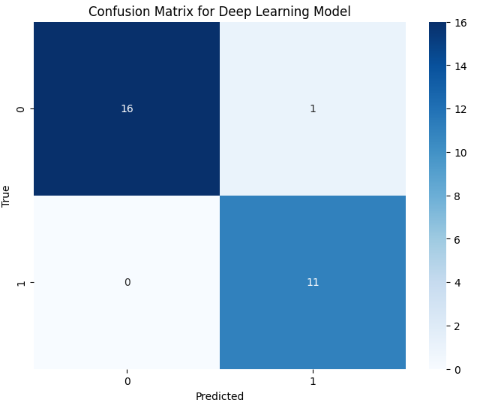


Figure 9. Confusion Matrix

Figure 9 visualizes the performance of the deep learning model by showing the number of true positives, false positives, true negatives, and false negatives. It helps to understand the model's classification accuracy and error distribution.

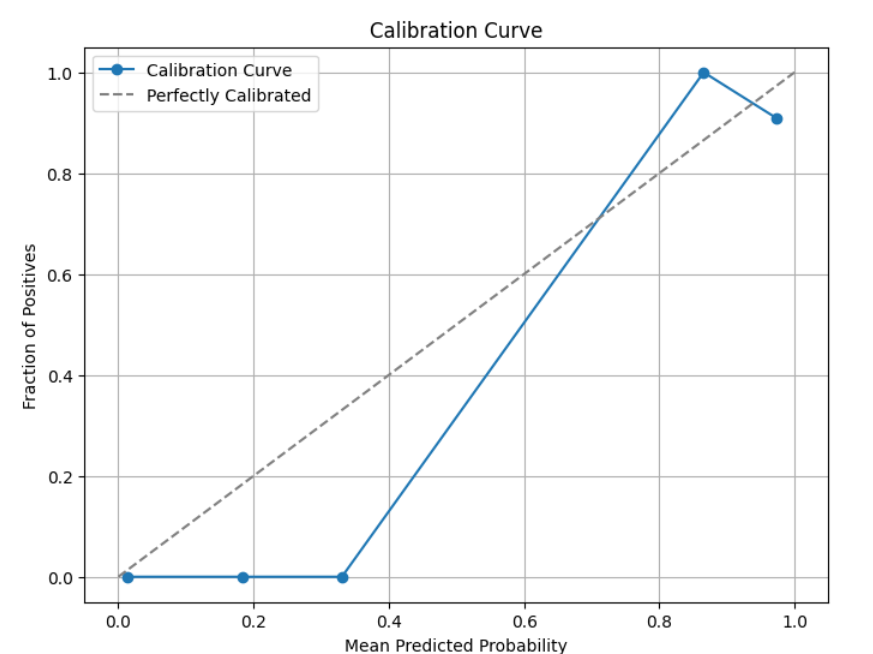


Figure 10. Calibration Curve

Figure 10 visualizes how well the predicted probabilities align with actual outcomes. It shows the mean predicted probability versus the fraction of positives, helping to evaluate the reliability of probability estimates.

# Conclusion

The deep learning model developed for PCOS detection effectively integrates robust data preprocessing, feature selection, and advanced neural network techniques to achieve reliable predictive performance. By leveraging methods such as outlier handling, feature reduction, and regularization, the model is capable of accurately identifying key patterns associated with PCOS. The use of SHAP for interpretability further enhances the model's transparency, providing valuable insights into feature importance and decision-making processes. Overall, this approach demonstrates the potential of deep learning to offer a comprehensive and actionable tool for early PCOS diagnosis, paving the way for improved healthcare outcomes.

# Future Enhancements

Future enhancements for the PCOS prediction model could include exploring more advanced neural network architectures, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to capture potential patterns or sequences in the data that may be overlooked by the current model. Additionally, integrating more diverse data sources, such as genetic or lifestyle information, could provide a richer dataset and potentially improve prediction accuracy. Implementing hyperparameter optimization techniques, such as grid search or Bayesian optimization, could fine-tune model parameters for better performance. Expanding the interpretability of the model with techniques like LIME (Local Interpretable Model-agnostic Explanations) alongside SHAP could offer deeper insights into the model’s behavior. Finally, exploring ensemble methods by combining multiple models could enhance predictive performance and robustness.

# Limitations

##### Despite its strengths, the proposed model has several limitations. The reliance on a single dataset, even with comprehensive preprocessing and feature selection, may limit the model's generalizability to other populations or datasets. The model's interpretability, while enhanced by SHAP, may still be challenging to fully understand and explain to non-technical stakeholders. Additionally, the feature selection process, although thorough, may inadvertently exclude potentially relevant features or interactions that could improve model performance. The computational cost and complexity of training deep learning models may also pose practical challenges, particularly with larger datasets. Lastly, the model's performance is contingent upon the quality and representativeness of the input data, and any biases in the data could affect the model's accuracy and fairness.

##### References

1. Suha, S. A., & Islam, M. N. (2022). An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image. *Scientific Reports, 12*, 17123. <https://doi.org/10.1038/s41598-022-21724-0>
2. Lv, W., Song, Y., Fu, R., Lin, X., Su, Y., Jin, X., Yang, H., Shan, X., Du, W., Huang, Q., Zhong, H., Jiang, K., Zhang, Z., Wang, L., & Huang, G. (2022). Deep learning algorithm for automated detection of polycystic ovary syndrome using scleral images. *Frontiers in Endocrinology, 12*, 789878. <https://doi.org/10.3389/fendo.2021.789878>
3. Sumathi, M., Chitra, P., Sakthi Prabha, R., & Srilatha, K. (2021). Study and detection of PCOS related diseases using CNN. *IOP Conference Series: Materials Science and Engineering, 1070*, 012062. <https://doi.org/10.1088/1757-899X/1070/1/012062>
4. Abu Adla, Y. A., Saad, R. A., Raydan, D. G., Nasreddine, J., Charaf, M.-Z. J., & Diab, M. O. (2021). Automated detection of polycystic ovary syndrome using machine learning techniques. *IOP Conference Series: Materials Science and Engineering, 1070*, 012062. <https://doi.org/10.1109/ICABME53305.2021.9604905>
5. Bhat, S. A. (2021). Detection of polycystic ovary syndrome using machine learning algorithms (MSc Research Project). *National College of Ireland*. <https://doi.org/10.1109/ICABME53305.2021.9604905>
6. Hosain, A. K. M. S., Mehedi, M. H. K., & Kabir, I. E. (2022). PCONet: A convolutional neural network architecture to detect polycystic ovary syndrome (PCOS) from ovarian ultrasound images. *Proceedings of the 8th International Conference on Engineering and Emerging Technologies (ICEET)*, 27-28 October 2022, Kuala Lumpur, Malaysia. <https://arxiv.org/abs/2210.00407>
7. Bhosale, S., Joshi, L., & Shivsharan, A. (2022). PCOS (Polycystic Ovarian Syndrome) detection using deep learning. *International Research Journal of Modernization in Engineering, Technology and Science*, 4(1). <https://www.irjmets.com>
8. Thakre, V., Vedpathak, S., Thakre, K., & Sonawani, S. (2020). PCOcare: PCOS detection and prediction using machine learning algorithms. *Bioscience Biotechnology Research Communications*, 13(14), 240-244. <http://dx.doi.org/10.21786/bbrc/13.14/56>
9. Alamoudi, A., Alsaif, H. S., Khan, I. U., AlDandan, O., Aslam, N., Alqahtani, N., AlGadeeb, M., & Al Bahrani, R. (2023). A deep learning fusion approach to diagnose polycystic ovary syndrome (PCOS). *Applied Computational Intelligence and Soft Computing*, 2023, Article ID 9686697, 15 pages. <https://doi.org/10.1155/2023/9686697>
10. Hdaib, D., Almajali, N., Alquran, H., Mustafa, W. A., Al-Azzawi, W., & Alkhayyat, A. (2022). Detection of Polycystic Ovary Syndrome (PCOS) Using Machine Learning Algorithms. *2022 5th International Conference on Engineering Technology and its Applications (IICETA)*, 1-6. <https://doi.org/10.1109/IICETA54559.2022.9888677>