

Early Detection Of Polycystic Ovarian Disease Using Optimized Machine Learning Technique

Suvika K V^{1*}, Dr. D G Jyothi²

¹Research Scholar, Dept. of CSE, Bangalore Institute of Technology, Visvesveraya Technological University, Belagavi – 590018, kvsuvika@gmail.com

²Professor, Dept. of AI&ML, Bangalore Institute of Technology, Visvesveraya Technological University, Belagavi – 590018, dgjyothi@bit-bangalore.edu.in

ABSTRACT

Polycystic ovary syndrome (PCOS) and polycystic ovary disorder (PCOD) are widespread endocrine conditions affecting millions of women around the world. These disorders can result in considerable reproductive challenges and metabolic complications. By providing the urgent need for effective and accessible diagnostic methodologies, the proposed study aims to enhance the efficiency and interpretability of Artificial Neural Network (ANN) techniques for the early detection of these conditions through innovative data pre-processing strategies and progressive training data modules. This research addresses a critical gap in the literature by systematically identifying and eliminating noisy data during the pre-processing phase. The Tarsier Optimization Algorithm (TOA), inspired by the adaptive behaviours of Tarsiers is introduced to facilitate efficient exploration of the solution space. This approach refines the dataset used in model training while emphasizing the importance of feature significance. By leveraging TOA's adaptive strategies, the clarity and performances of the progressive ANN (PANN-TOA) are improved, ensuring robust handling of noisy data. The methodology highlights the necessity of optimizing data processing to deepen the understanding of the complex manifestations and health outcomes associated with PCOS and PCOD. Ultimately, the new work aims to provide a robust framework that enhances diagnostic accuracy of about 95% and informs personalized treatment strategies, paving the way for improved reproductive health outcomes and empowering individuals to make informed health decisions. Thus, the results show that the PANN performed better than previous machine learning (ML) models, demonstrating its capacity to identify intricate patterns in clinical data.

Keywords: polycystic ovary disease, Tarsier Optimization Algorithm, progressive ANN, Predictive Model.

1. INTRODUCTION

Polycystic Ovary Disorder (PCOD) and Polycystic Ovary Syndrome (PCOS) are widespread endocrine disorders affecting millions of women globally, with prevalence rates estimated between 6.5% and 10% [1]. These conditions involve hormonal imbalances that can result in symptoms like irregular menstrual cycles, excessive hair growth (hyperandrogenism), and multiple ovarian cysts. Such symptoms can interfere with normal reproductive processes and heighten the risk of serious health issues, including metabolic syndrome, type 2 diabetes, cardiovascular problems, and infertility [2].

The urgent need for effective diagnostic and management strategies is underscored by studies indicating that women with PCOS are 2 to 3 times more likely to develop type 2 diabetes compared to those without the disorder [3]. This emphasizes the critical importance of early diagnosis and intervention, as timely management can significantly enhance the quality of life for affected individuals.

Traditionally, the diagnosis of PCOS and PCOD has relied on clinical assessments and imaging techniques, such as ultrasound. While ultrasound remains an essential diagnostic tool, its effectiveness can be compromised by operator variability and subjective interpretation, leading to inconsistencies in diagnosis and management [4]. In regions with limited access to advanced medical technology, these challenges can result in

delayed or missed diagnoses, exacerbating health issues. A 2022 study revealed that nearly 40% of women with PCOS reported experiencing delays in diagnosis, highlighting systemic inefficiencies with serious implications for women's health [5].

Recent advancements in artificial intelligence (AI) and ML offer promising solutions to enhance the diagnostic process for PCOS and PCOD. By utilizing these technologies, clinicians can analyze large datasets to identify patterns that may be overlooked by traditional methods. Specifically, ANNs can be trained on historical data to recognize indicators of PCOS and PCOD, potentially providing quicker and more accurate diagnoses. However, a significant challenge remains in the realm of data preprocessing, particularly in identifying and mitigating noisy or irrelevant data, which can compromise model performance [6-10].

The proposed study focuses on enhancing the early detection of PCOS and PCOD by employing a progressive ANN in conjunction with the TOA. Inspired by the adaptive behaviours of Tarsiers with small primates known for their agility and keen perception, TOA is designed to optimize data handling in model training. By systematically identifying and eliminating noisy data, TOA refines the dataset, thereby improving the clarity and predictive power of the PANN.

The importance of high-quality data in enabling accurate and reliable diagnoses is paramount, serving as the foundation for effective predictive models [11]. The primary contributions of this research include:

- A thorough evaluation of existing data preprocessing methodologies tailored specifically to the context of PCOS and PCOD.
- The implementation of innovative noise reduction and feature selection techniques aimed at enhancing dataset integrity.
- An exploration of how the PANN-TOA can improve the performance of AI and ML models in diagnosing these disorders.

By bridging the gap between traditional diagnostic methods and modern AI-driven approaches, this research aims to contribute to more efficient and accessible healthcare solutions for women suffering from PCOS and PCOD. Ultimately, our goal is to empower women with improved diagnostic options, leading to earlier interventions and significantly better health outcomes.

2. LITERATURE REVIEW

The advancement of data preprocessing techniques in the context of PCOD and PCOS has gained significant traction as healthcare increasingly adopts data-driven methodologies. Effective preprocessing is essential for optimizing ML model performance, particularly given the complexities associated with reproductive health data. This section discusses existing data preprocessing techniques, current model selection methods specifically related to PANNs, and the gaps related to noisy data handling in the literature (Table 1) while highlighting the role of the TOA.

Table 1: Discussions of noisy data handling models

Author(s)	Year	Title	Findings	Limitations
Gupta et al. [12]	2020	Importance of Data Normalization and Standardization in PCOS	Normalization and standardization significantly reduced variance in hormonal measurements, improving predictive outcomes.	Does not address root causes of noise in datasets, which may still affect model performance.

Lin et al. [13]	2019	Feature Selection Techniques for Enhanced Model Interpretability	Recursive Feature Elimination (RFE) effectively identifies relevant features, reducing dataset dimensionality while retaining critical information.	May not account for the dynamic nature of feature relevance across different datasets.
Sharma et al. [14]	2021	Outlier Detection Methods for Dataset Cleaning	Proposed a combination of statistical tests and ML algorithms for effective outlier identification and removal.	Approach could benefit from a more adaptive framework that refines detection based on evolving data characteristics.
Modi et al. [15]	2022	Comparative Evaluation of Classifiers for PCOS Diagnosis	Random Forest achieved 95% accuracy in diagnosing PCOS using clinical and metabolic data.	Focus on a single metric (accuracy) can obscure biases, especially in imbalanced datasets.
Stafie et al. [16]	2023	Deep Learning Techniques for Image-based Detection of PCOS	Convolutional Neural Networks achieved 98% accuracy, demonstrating the potential of neural networks for enhancing diagnostic precision.	Lack of comprehensive evaluations comparing models across diverse datasets and populations.
Wagner et al. [17]	2020	Impact of Data Quality on ML Models	Highlighted that inadequate preprocessing protocols can lead to overfitting and biased results, limiting generalizability.	Often overlooks the sources of noise in datasets, compromising the reliability of diagnostic tools.
Wang et al. [18]	2021	Unified Approaches in Data Integration for PCOS/PCOD	Called for integrating clinical, hormonal, and imaging data to enhance model robustness and clinical applicability.	Current methodologies often fail to combine diverse data types, limiting their practical relevance.

While significant strides have been made in data preprocessing and model selection for enhancing the early detection of PCOS and PCOD using ANNs, critical gaps persist, particularly regarding the handling of noisy data. Addressing these limitations is essential for improving diagnostic accuracy and overall patient care. The implementation of innovative preprocessing strategies, such as the Tarsier Optimization Algorithm, presents

an opportunity to enhance data integrity and support the development of more effective ML models in reproductive health. By bridging traditional diagnostic methods with modern AI-driven approaches, future research can contribute to more efficient and accessible healthcare solutions for women affected by these disorders.

3. METHODOLOGY

This methodology's primary goals are to use PANN and the TOA to improve the effectiveness and interpretability of ML techniques for the early detection of PCOS and PCOD [10]. The proposed study employs TOA for systematic data preprocessing and optimized feature selection. Inspired by the unique observational and adaptive behaviours of Tarsiers, TOA provides a structured approach to improve data quality and model performance. Figure 1 illustrates the overall workflow of the proposed methodology, encompassing data collection, pre-processing, TOA application, PANN training, and evaluation.

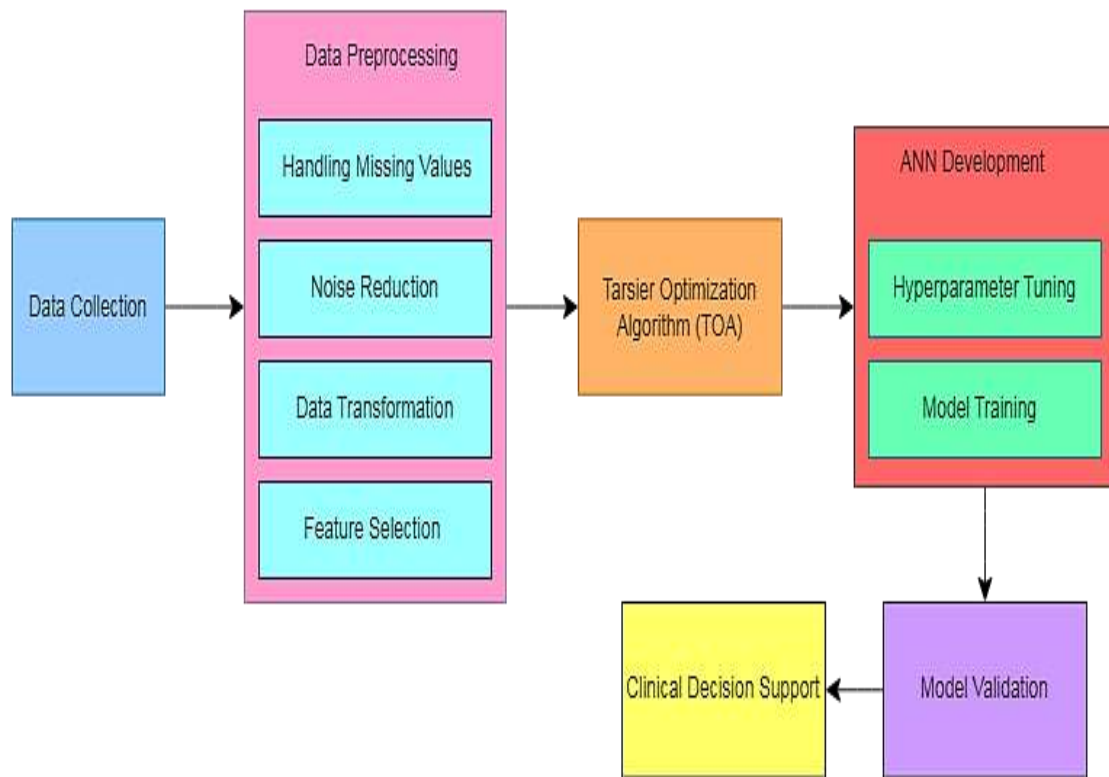


Figure 1: Workflow of the Tarsier Optimization-PANN for Early Detection of PCOS and PCOD

3.1 Data Collection

The dataset utilized in the proposed study comprises clinical, hormonal, and demographic information from women diagnosed with PCOS and PCOD [20][21]. Data were collected from multiple medical institutions through a combination of methods, including:

- **Medical Records:** Patient data were sourced from electronic health records, ensuring comprehensive clinical information regarding diagnoses, treatment histories, and hormonal assessments.
- **Surveys and Questionnaires:** Structured surveys were administered to collect additional demographic and lifestyle factors, such as age, Body Mass Index (BMI), diet,

exercise frequency, and stress levels. This method facilitated a deeper understanding of the lifestyle influences on PCOS and PCOD.

- **Clinical Assessments:** Hormonal levels were measured through laboratory tests conducted by qualified medical professionals, ensuring accuracy in the data collected. The multi-faceted approach to data collection ensured a diverse range of participants, which enhances the generalizability of the findings. Furthermore, rigorous ethical guidelines were followed, including informed consent from all participants, to uphold the integrity of the data collection process.

3.1.1 Description of the Dataset Used

The dataset includes various relevant features critical for identifying patterns and associations with PCOS and PCOD [43]. Some of the features are summarized in Table 2 and Table 3.

Table 2: Overview of Feature Categories

Feature Category	Features
Clinical Parameters	Hormonal levels (e.g., LH, FSH, testosterone)
Demographic Data	Age, Body Mass Index (BMI), ethnicity
Lifestyle Factors	Diet, exercise frequency, stress levels
Menstrual Cycle Information	Average length of menstrual cycles
Family History	Family history of PCOS/PCOD (binary)
Ultrasound Findings	Presence of ovarian cysts (binary)

Table 3: Overview of Dataset Features

Feature	Description
Age	Age of the patient
BMI	Body Mass Index
Menstrual Cycle Length	Average length of menstrual cycles
Testosterone Levels	Hormonal levels measured in ng/dL
Insulin Levels	Hormonal levels measured in μ IU/mL
Ultrasound Findings	Presence of ovarian cysts (binary)
Family History	Family history of PCOS/PCOD (binary)

3.1.2 Data Splitting

To ensure robust model validation, the dataset is split into a training set (70%) and a testing set (30%). This division facilitates the development of the ANN on the training data while reserving a separate dataset for assessing model performance [23].

3.2 Data Pre-processing

Data pre-processing is crucial for preparing the dataset for model training and enhancing the reliability of the results. The proposed study focuses on clinical, hormonal, and demographic features for the early detection of PCOS and PCOD [24]. The following methodologies are applied:

3.2.1 Handling Missing Values

To address missing values, a combination of imputation techniques is employed and tailored to the nature of the data. For continuous variables, such as Testosterone Levels or Insulin Levels, mean imputation is utilized, which replaces missing values with the mean of the observed values. This is expressed mathematically as:

$$X_{\text{imputed}} = \frac{\sum_{i=1}^n X_i}{n} \quad (1)$$

where, X_{imputed} is the imputed value for a missing entry, n is the number of non-missing values, and X_i represents the observed values of the feature [25].

For categorical variables, like family history, usage of mode imputation, replacing missing values with the mode, defined as:

$$\text{mode}(X) = \text{argmax}(\text{count}(X_k)) \quad (2)$$

where, X is the categorical feature, X_k represents unique values in X , ($\text{count}(X_k)$) is the count of occurrences of each unique value. Also, KNN imputation is implemented to predict and fill missing values based on the attributes of similar instances. For example, if BMI has missing values, KNN calculates the Euclidean distance to identify the closest instances:

$$d(i, j) = \sqrt{\sum_{k=1}^m (X_{ik} - X_{jk})^2} \quad (3)$$

where, $d(i, j)$ is the distance between instances i and j , m is the number of features, and X_{ik} and X_{jk} are the values of the k -th feature for instances i and j , respectively. By analyzing similar instances, KNN imputes the missing BMI values, providing a more informed estimate than mean or mode imputation, thus preserving the underlying distribution of the data. After identifying the k nearest neighbours, the missing values can be replaced using the average for continuous variables or the mode for categorical variables.

3.2.2 Noise Reduction and Outlier Detection

Next, noise and outliers are addressed by computing Z-scores for each variable, using the formula:

$$Z = \frac{(X - \mu)}{\sigma} \quad (4)$$

where, Z is the Z-score, X is the individual value, μ is the mean of the feature, and σ is the standard deviation of the feature [26].

Calculating Z-scores allows us to standardize features like Insulin Levels. Values with Z-scores exceeding ± 3 are flagged as potential outliers, indicating that they are significantly different from the majority of the data. Identifying and handling these outliers is crucial to avoid skewing the model's performance. Also, the Interquartile Range (IQR) method is utilised to establish acceptable value with ranges that could be calculated as:

$$\text{IQR} = Q_3 - Q_1 \quad (5)$$

where, Q_1 and Q_3 are the first and third quartiles, respectively. Outlier thresholds are defined as:

$$\text{Lower Bound: } (Q_1 - 1.5 * \text{IQR}) \quad (6)$$

$$\text{Upper Bound: } (Q_3 + 1.5 * \text{IQR}) \quad (7)$$

Values of features like Cycle Length that fall outside these bounds are evaluated for their clinical relevance, ensuring that only pertinent data is retained.

3.2.3 Data Transformation

Moving forward, consider data transformation to ensure comparability among features, preventing any single feature from disproportionately influencing the model. Figure 2 illustrates the normalization process applied to the dataset [27], showing the steps involved in rescaling feature values.

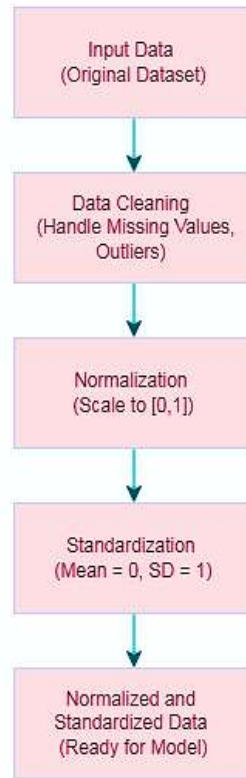


Figure 2: Overview of the Normalization Process Applied to the Dataset

The transformation ensures uniformity across features, facilitating better model performance and reducing biases associated with differing scales. The normalization is applied to scale continuous features to a range of [0, 1], expressed as:

$$X' = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (8)$$

where, X' is the normalized value, X is the original value, $\min(X)$ is the minimum value of the feature (e.g., Weight) and $\max(X)$ is the maximum value of the feature.

This normalization allows for better comparisons among features, which is particularly beneficial for algorithms like KNN. For example, applying this normalization to BMI ensures that it contributes equally in distance calculations without dominating due to its larger scale.

For standardization, the data is transformed to have a mean of 0 and a standard deviation of 1, given by:

$$X_{std} = \frac{X - \mu}{\sigma} \quad (9)$$

where, X_{std} is the standardized value, X is the original value, μ is the mean of the feature (e.g., Insulin Levels), and σ is the standard deviation of the feature.

This is especially advantageous for algorithms that assume a Gaussian distribution, such as linear regression and ANN.

3.2.4 Feature Selection

Finally, a feature selection method is based on mutual information to enhance model interpretability and performance. This measure quantifies the amount of information gained about one variable through another, defined by:

$$I(X; Y) = H(X) - H(X | Y) \quad (10)$$

where, $I(X;Y)$ is the mutual information between variables X (e.g., Insulin Levels) and Y (diagnosis of PCOS or PCOD), $H(X)$ is the entropy of variable X , and $H(X|Y)$ is the conditional entropy of X given Y [28].

Higher mutual information values indicate stronger relationships between features and the target variable. By computing mutual information scores for all features, one can identify and remove those with low scores, thereby streamlining the dataset for effective model training. This comprehensive preprocessing phase lays a strong foundation for the subsequent training of the ANN, ultimately aiming for improved accuracy in diagnosing PCOS and PCOD.

3.3 Integration of Tarsier Optimization Algorithm (TOA)

To enhance data preprocessing and feature selection for improving the performance of the PANN in detecting PCOS and PCOD, the TOA can be employed. TOA's adaptability allows it to dynamically adjust strategies (Figure 3) based on model performance and feature significance [30], making it suitable for diverse datasets.



Figure 3: Adaptability of the Tarsier in Data Preprocessing

This algorithm is designed to systematically refine the dataset by eliminating noise and identifying relevant features, which is crucial for achieving high model accuracy [29]. The process can be summarized in the following algorithm:

Algorithm 1: Tarsier Optimization Algorithm (TOA)

Input:	Dataset D with n instances and m features
Output:	Refined dataset D'
Step 1. Initialize Population:	Create an initial population $P = \{p_1, p_2, \dots, p_k\}$, where each solution p_i represents a subset of features randomly selected from D .
Step 2. While Stopping Criteria Not Met (e.g., iterations < T):	Repeat the following steps until the stopping criteria are satisfied.
Step 3. Compute Fitness Score:	For each solution p_i in P : Calculate the fitness score: $F(p_i) = \alpha \times \text{Accuracy}(p_i) + \beta \times \text{Precision}(p_i) + \gamma \times \text{Recall}(p_i)$ where α, β , and γ are weights that can be adjusted based on the importance of each metric.
Step 4. Perform Neighbourhood Search:	For each solution p_i in P : For each feature f_j in p_i : Generate a new solution $p' = p_i \pm f_j$ (adding or removing f_j). Evaluate the fitness $F(p')$. If $F(p') > F(p_i)$: Update $(p_i = p')$.
Step 5. Select Optimal Solutions:	Form a set of optimal solutions: $S = \{p_j \mid F(p_j) \geq \theta\}$ where θ is a predefined threshold.
Step 6. Check for convergence:	If $ F(p_i) - F(p_{i-1}) < \varepsilon$ for all p_i : Break the loop
Step 7. Finalize:	Create the refined dataset: $D' = \{x_k \mid x_k \in D \cap k \in S\}$

3.3.1 Application in Data Preprocessing

The refined dataset D' is essential for effective model training. TOA evaluates the relevance of remaining features. The Recursive Feature Elimination (RFE) in conjunction with TOA is employed to iteratively remove the least significant features based on model performance. Typically, a model like Random Forest provides feature importance scores to guide this elimination process [31]. Features with low importance scores are candidates for removal, leading to a streamlined dataset that retains only high-quality, relevant features critical for predicting PCOS and PCOD.

3.3.2 Integration with ANNs

Once the refined dataset D' is established, it serves as the foundation for training the ANN. The training process begins with configuring the ANN architecture, which includes defining the number of layers and neurons per layer. Following this, weights are initialized to prepare for the training phase [32]. During training, the refined dataset (D') is fed into the network, and weights are optimized using suitable algorithms such as Adam or Stochastic Gradient Descent (SGD). After the training is complete, the performance of the ANN is evaluated using standard metrics, including accuracy,

precision, recall, and F1-score, allowing us to quantify the impact of the Tarsier Optimization Algorithm (TOA) on diagnostic accuracy. To further assess the effectiveness of TOA, a comparative analysis can be conducted, measuring the performance metrics of the ANN trained on D' against baseline models trained on the original, unrefined dataset. This comparison highlights the improvements achieved in feature selection and noise reduction [33].

3.4 ANN Development

In the proposed study, the refined dataset D' is utilized to train an ANN aimed at enhancing the predictive capabilities for diagnosing polycystic diseases [34]. The ANN framework is particularly well-suited for modelling complex nonlinear relationships, which is essential given the multi-factorial nature of these conditions. By leveraging the processed data, the ANN seeks to provide accurate diagnostic predictions, thereby aiding clinical decision-making.

3.4.1 ANN Architecture

The architecture of the ANN is meticulously designed to capture the intricate interactions between the target variable and the input features. As illustrated in Figure 4, the model consists of three layers: an input layer, two hidden layers, and an output layer. The input layer comprises 7 nodes, each representing a feature crucial for diagnosing PCOS and PCOD, such as age, BMI, menstrual cycle length [35], testosterone levels, insulin levels, ultrasound findings, and family history. These nodes input their respective feature values into the network, forming the foundation for subsequent processing.

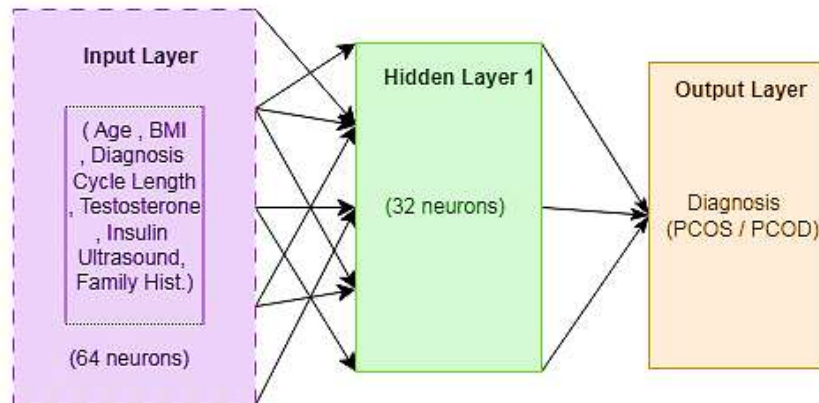


Figure 4: Architecture of the ANN for Diagnosing PCOS and PCOD

The first hidden layer consists of 64 neurons, selected based on preliminary experiments to effectively capture complex relationships inherent in the data. The higher number of neurons enhances feature representation, allowing the model to learn richer patterns. The second hidden layer, with 32 neurons, refines the learning process by reducing dimensionality and mitigating the risk of overfitting while retaining essential patterns learned from the previous layer. The output layer contains a single neuron designated for binary classification [36]. It outputs a probability value indicating the likelihood of a diagnosis (PCOS vs. PCOD) using a sigmoid activation function, defined as:

$$y = \frac{1}{1+e^{-z}} \quad (11)$$

where z is the weighted sum of inputs to the neuron. This structure is optimal for classification tasks with categorical target variables.

3.4.2 Hyper parameter Tuning

Hyper parameter tuning is a critical component in optimizing the ANN's performance. The final model's architecture, which includes two hidden layers, was validated as yielding the best performance metrics [37]. The first hidden layer comprises 64 neurons, while the

second consists of 32 neurons, enabling the model to learn complex patterns while maintaining generalizability. The choice of activation functions is crucial; the ReLU (Rectified Linear Unit) function is utilized in the hidden layers, defined as $f(x) = \max(0, x)$. This promotes efficient learning and addresses the vanishing gradient problem, while the output layer employs the sigmoid function to generate a probability score.

The learning rate η is optimally set at 0.001, striking a balance between convergence speed and stability. A batch size (b) of 32 is chosen to provide a suitable trade-off between training speed and model stability [38]. To ensure robustness, cross-validation is implemented with $k = 5$, which involves partitioning the training data into five subsets. This allows for multiple training and validation cycles, enhancing the reliability of performance assessments.

3.4.3 Progressive Model Training Process in ANN

The model training process is systematic and follows several key steps. First, the refined dataset D' is divided into three distinct subsets: the Training Set (70%), which is employed to train the ANN [39]; the Validation Set (15%), which helps in fine-tuning the model and adjusting hyperparameters to prevent overfitting; and the Testing Set 15%, which is reserved for the final evaluation of the model's performance and generalizability. The purpose of the implemented PANN module is to evaluate how robust optimization techniques are when there are erroneous training perturbations present. Assuming that X is the feature space and Y is the label space, let's define a classifier $h: X \rightarrow Y$ from a hypothesis space H . The clear samples $(x_i^s; y_i^s) \sim \mathcal{D}$ from the genuine data distribution contains the class semantics. To create training samples, $x_i = x_i^s + x_i^n$, $y_i = y_i^s$, let's take a look at the noise sampled from $x_i^n \sim \mathcal{D}_n$, which is subsequently added to the genuine data. The input sample distribution is denoted by $\mathcal{D}_n$. After that, the classifier can be used to train on their training samples $\{(x_i; y_i)\}$, and classification accuracy will be evaluated on test samples.

To establish a class of linearly separable datasets in which adaptive and non-adaptive optimizers exhibit distinct generalization behavior due to frameworks trained on interspersed spurious perturbations, let's examine a dataset $\{(x_i; y_i)\}$ of $\tilde{N}$ samples with binary labels $y \in \{+1, -1\}$ and $p > 1/2$ seems the probability of positive labels. $\mu(2\tilde{N} + 1) + (\tilde{N} + 1)$ is the feature dimension of x , where μ constitutes a positive constant. We take into consideration the initial data distribution as follows,

$$x_{ij}^s = \begin{cases} y_i, & \text{if } j = 1 \\ 1, & \text{if } j = 2, \dots, (\mu + 1) \\ 0, & \text{otherwise} \end{cases} \quad (12a)$$

It is easy to identify neural network parameters that accurately categorize the aforementioned dataset, which is linearly separable in the first dimension. A perturbation of the following type on the above dataset can be applied as,

$$x_{ij}^n = \begin{cases} y_i, & \text{if } j = 1 \\ 1, & \text{if } j = \text{indx}_i^s, \dots, \text{indx}_i^{\hat{e}} \\ 0, & \text{otherwise} \end{cases} \quad (12b)$$

It provides the indexes for the i^{th} feature block's beginning and conclusion. If $\text{indx}_i^s = 2 + \mu + (2\mu + 1)(i - 1)$ and $\text{indx}_i^{\hat{e}} = 2 + \mu + (2\mu + 1)(i - 1) + \mu(1 - y_i)$. This feature fills the entire following $(2\mu + 1)$ locations with 1's, if $y_i = 1$; otherwise, it fills a moving block of $(2\mu + 1)$ length displaced by $(i - 1)$ times with a single 1 at $2 + \mu + (2\mu + 1)(i - 1)$ position.

$$x_{ij}^n = \begin{cases} y_i, & \text{if } j = 1 \\ 1, & \text{if } j = 2, \dots, (\mu + 1) \\ 1, & \text{if } j = \text{indx}_i^s, \dots, \text{indx}_i^{\hat{e}} \\ 0, & \text{otherwise} \end{cases} \quad (12c)$$

In order to get the training features, the nuisances are then superimposed on the real distribution to create the perturbed training data, $\chi_{ij}^{train} = \chi_{ij}^s + \chi_{ij}^n$ using variables that mean the same things as previously mentioned. It should be mentioned that the distribution of training data that was disturbed above is likewise linearly separable in the first dimension. By using the first dimension as a discriminative feature, it is possible to discover an optimal decision function that will have good generalization to the clean test data.

Next, the model configuration, refined through hyperparameter tuning, is trained on the training set using the Adam optimizer, chosen for its efficiency in navigating complex data landscapes. The loss is computed using binary cross-entropy [40], defined as:

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (13)$$

where N is the number of instances, y_i is the actual label, and $\hat{y}_i$ is the predicted probability for the positive class.

The overall training workflow is illustrated in Figure 5, depicting the process from data input through the forward pass in the network, followed by loss calculation, back propagation, and weight updates. This visualization captures the iterative nature of training, where model weights are continually adjusted to minimize prediction error.

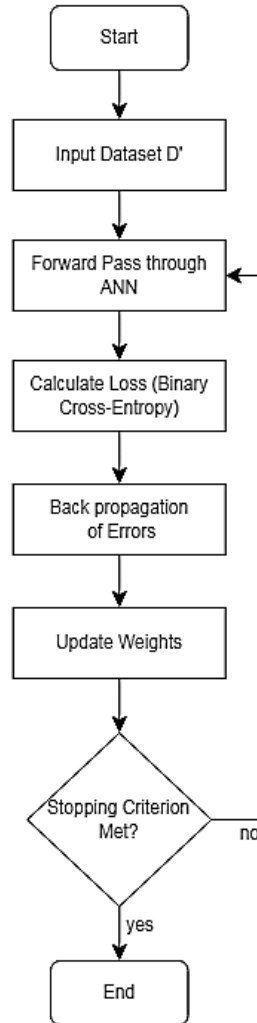


Figure 5: Model Training Workflow

During training, important performance indicators including accuracy, precision, recall, and loss are tracked continuously [41]. The training process employs early stopping if the validation loss does not improve after a predetermined number of epochs (patience = 10), which helps prevent overfitting and ensures that the final model maintains its generalization capabilities to unseen data.

The comprehensive development and training of the ANN using the refined dataset $\setminus(D')\setminus$ aim to enhance diagnostic accuracy for PCOS and PCOD. By carefully designing the architecture, tuning hyperparameters, and employing a rigorous training process [42], this approach seeks to create a robust model that can effectively aid clinical decision-making.

3.5 Model Validation

To determine the efficacy of the trained models, a variety of indicators were used to evaluate model performance [33].

3.5.1 Evaluation Metrics

1. Accuracy: The percentage of cases (including true positives and true negatives) that were correctly detected out of all the cases.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

2. Precision: The ratio of actual positive forecasts to all expected positive forecasts.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (15)$$

3. Recall: The ratio of actual positives to genuine positive predictions.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (16)$$

4. F1 Score: A statistic that balances precision and recall by taking the harmonic mean of the two.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (17)$$

Using an ANN, this methodology offers a strong framework for improving the early diagnosis of PCOS and PCOD. The Tarsier Optimization Algorithm is essential for feature selection and data preparation.

4. RESULTS AND DISCUSSION

This section presents the outcomes of our proposed model, which utilizes PANN enhanced by the TOA for the early detection of PCOS and PCOD. The focus is on how TOA improves the efficiency and interpretability of PANN techniques compared to other ML models.

4.1 Dataset Overview

The study utilized a dataset comprising clinical, hormonal, and demographic data from the records of women diagnosed with PCOS and PCOD [43]. The dataset included essential features outlined in Tables 1 and 2 from the methodology section, reflecting the diversity and clinical relevance of the data. Figure 6 illustrates the distribution of key features such as Age, BMI, and Hormonal levels, providing insights into the population characteristics relevant to PCOS/PCOD [16].

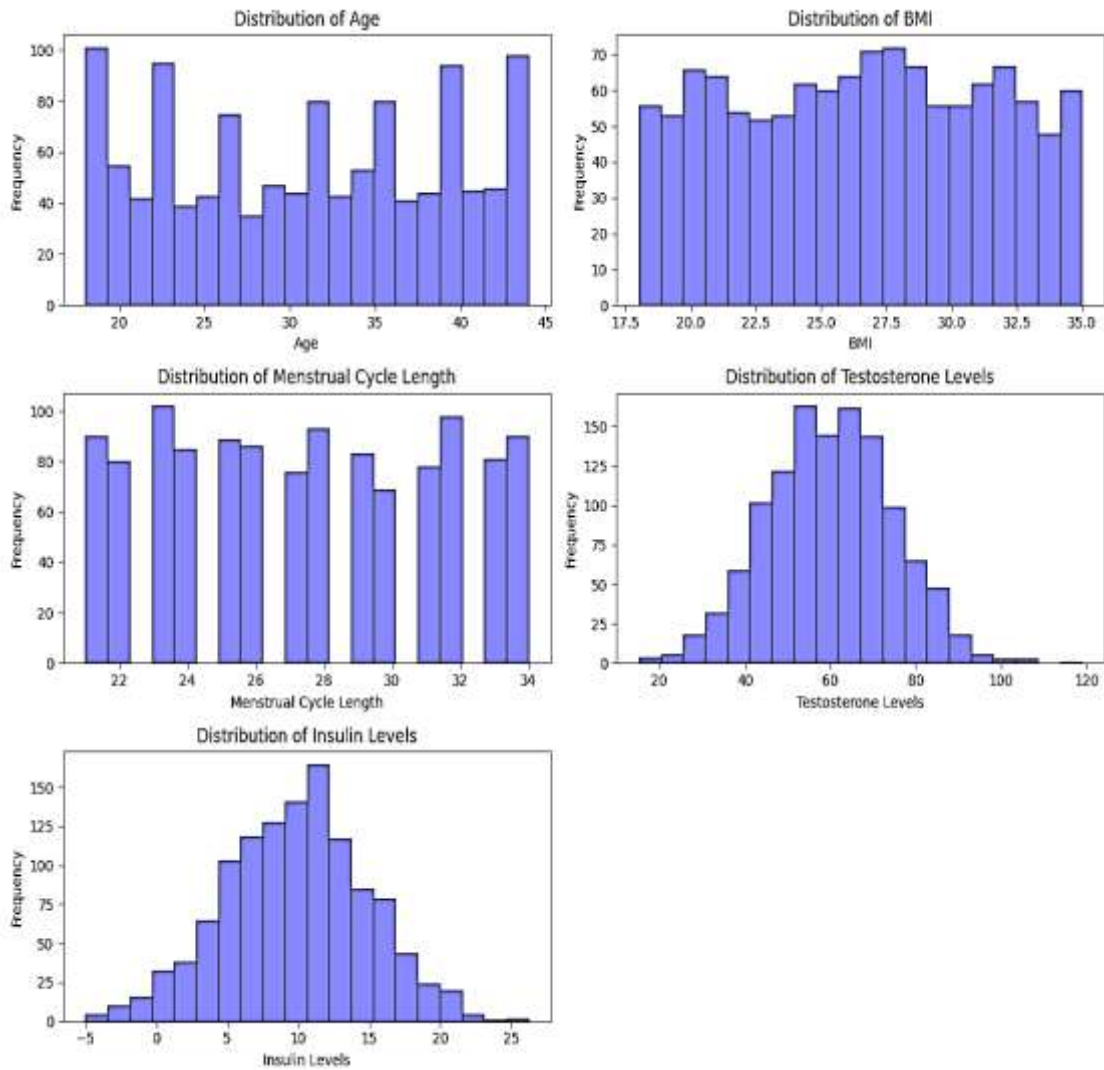


Figure 6: Data Distribution

4.2 Model Performance Evaluation

Table 4 compares the performance metrics of our proposed model with those of other ML models, including Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM).

Table 4: Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	84.4	77.8	75.68	76.7
Random Forest Classifier	85.35	83.87	70.27	76.5
Support Vector Machine	86.2	78.97	81.08	80.1
PANN-TOA	89.91	88.3	81.1	84.5

Thus, the newly optimized PANN achieved the highest accuracy of 89.91% of training data, highlighting its ability to capture complex relationships in the data [9]. The PANN model also demonstrated strong performance, with tanh activation model, indicating robust classification capabilities. Figure 7 displays the activation model for various ANN functions, assessing their capacity to differentiate between classes.

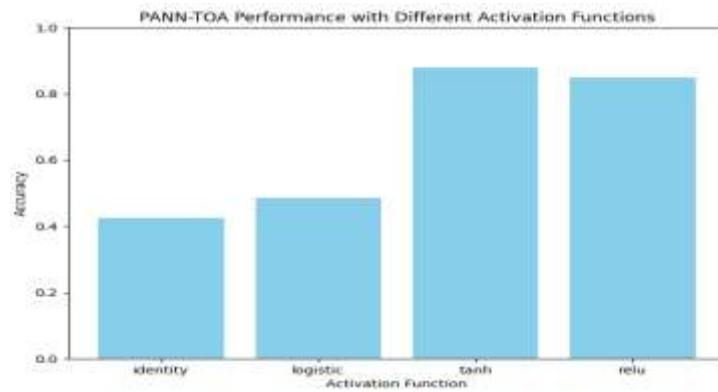


Figure 7: Activation models in PANN

Figures 8 (a & b) present a bar chart comparing the test vs. train accuracy and some performance metrics of the different ML models, emphasizing the superiority of the PANN in prediction rate.

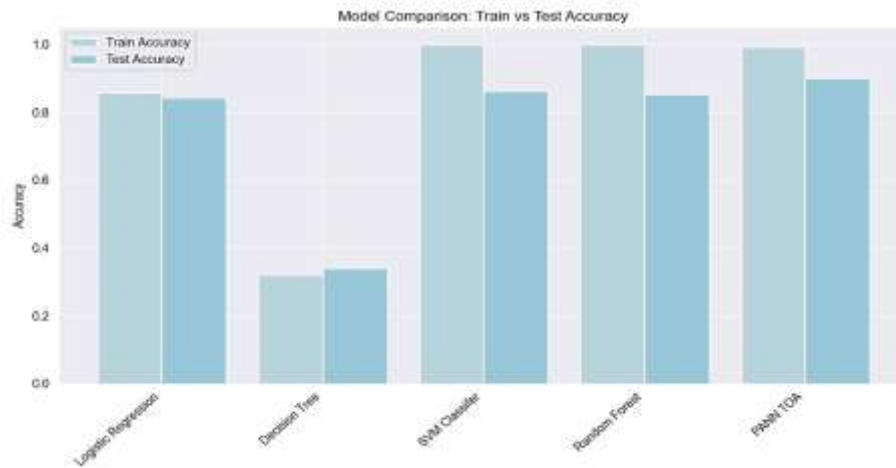


Figure 8 (a): Accuracy Comparisons

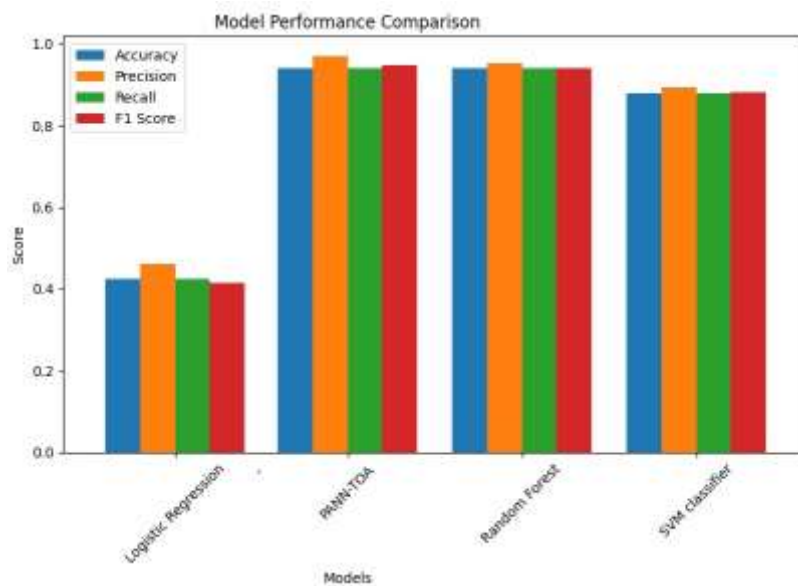


Figure 8 (b): Model Performance Comparison

4.3 Impact of Tarsier Optimizer

The integration of TOA during data pre-processing played a crucial role in enhancing model performance. TOA effectively reduced noise within the dataset by implementing robust statistical techniques such as Z-score analysis and the Interquartile Range (IQR) method. Outliers exceeding the $\pm$ Z-score threshold were flagged, resulting in cleaner data for model training. Figure 9 shows the number of outliers detected and removed before and after applying TOA, illustrating its effectiveness in noise reduction.

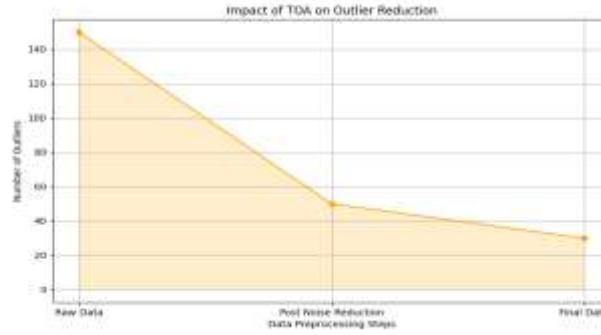


Figure 9: Noise Reduction Impact

Utilizing techniques like mutual information and feature importance measures, TOA systematically prioritized significant features. This not only improved model accuracy but also facilitated interpretability, aligning with the necessity for clinically relevant insights in PCOS/PCOD diagnosis. Figure 10 illustrates the number of features retained before and after TOA, demonstrating the improvement in model interpretability.



Figure 10: Feature Selection Impact

4.4 Impact of Data Preprocessing Techniques

The data preprocessing techniques guided by the Tarsier OA significantly affected the performance of my proposed model, the PANN:

1. **Handling Missing Values:** The combination of mean, mode, and KNN imputation resulted in a more complete dataset, positively impacting the accuracy of the PANN.
2. **Data Transformation:** Normalization and standardization of features ensured that the PANN was trained on data that did not disproportionately weight specific features, enhancing overall stability and performance.

4.5 Feature Importance Analysis

Feature importance analysis revealed significant contributions from several features in predicting PCOS/PCOD. TOA model with training and validating performance with several epochs are depicted in Figure 11.



Figure 11: Training Selection Impact of TOA model

The most relevant features included Insulin levels and BMI, consistent with existing literature (as in Table 5). Figure 12 displays the correlation matrix among selected features, illustrating the strong relationships, particularly between Insulin and PCOS/PCOD diagnoses.

Table 5: Selected Features and Correlation Coefficients

Selected Features	Correlation Coefficient
Cycle Length	0.15
Menses length	0.17
Unusual bleeding	0.3
Mean length	0.28

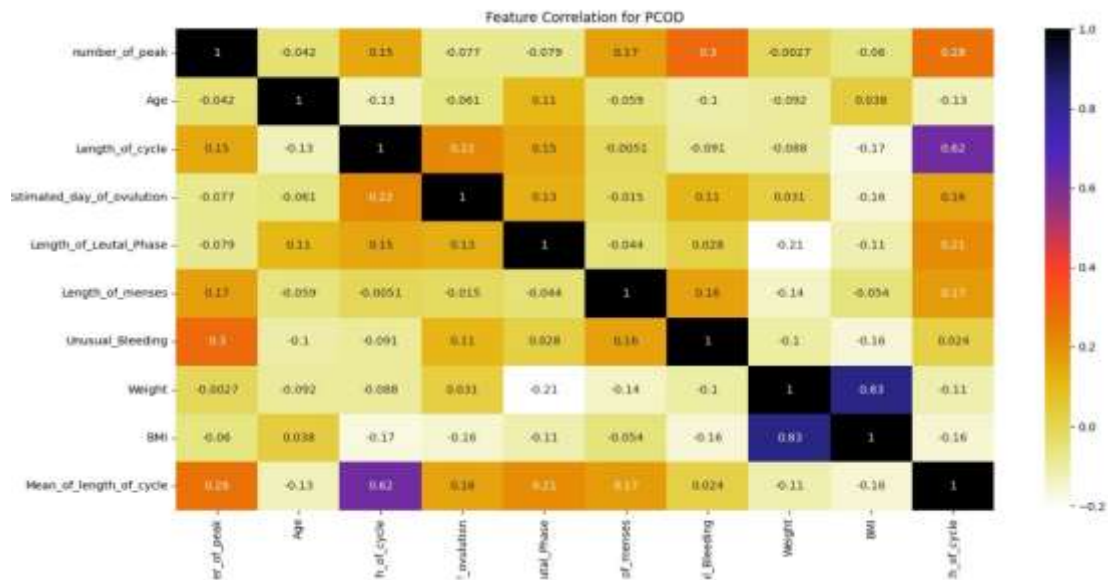


Figure 12: Correlation Matrix

4.6 Learning Curves and Model Validation

Figure 13 depicts training versus validation accuracy over epochs. This figure helps identify model convergence and potential over fitting.



Figure 13: Learning Curves

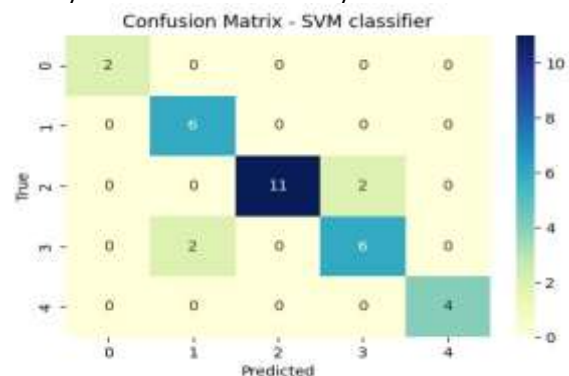
The learning curves indicate that while all models improved over time, our PANN demonstrated a gradual increase in accuracy with a lower risk of over-fitting compared to others, underscoring its capability to generalize better to unseen data.

4.7 Feature Importance Visualization

On visualizing the importance scores of each feature, the features are most influential in predictions. The feature importance scores align with previous analyses, reaffirming the significance of Insulin levels and BMI in predicting PCOS/PCOD. This visualization assists clinicians in identifying key factors to monitor in patients.

4.8 Confusion Matrix

The confusion matrix highlights that the PANN achieved a high rate of true positives, indicating its efficacy in accurately identifying cases of PCOS/PCOD. Figure 14 visualizes the true values, and predicted rates for the PANN, providing insight into its ability to distinguish between PCOS/PCOD and non-PCOS/PCOD cases.



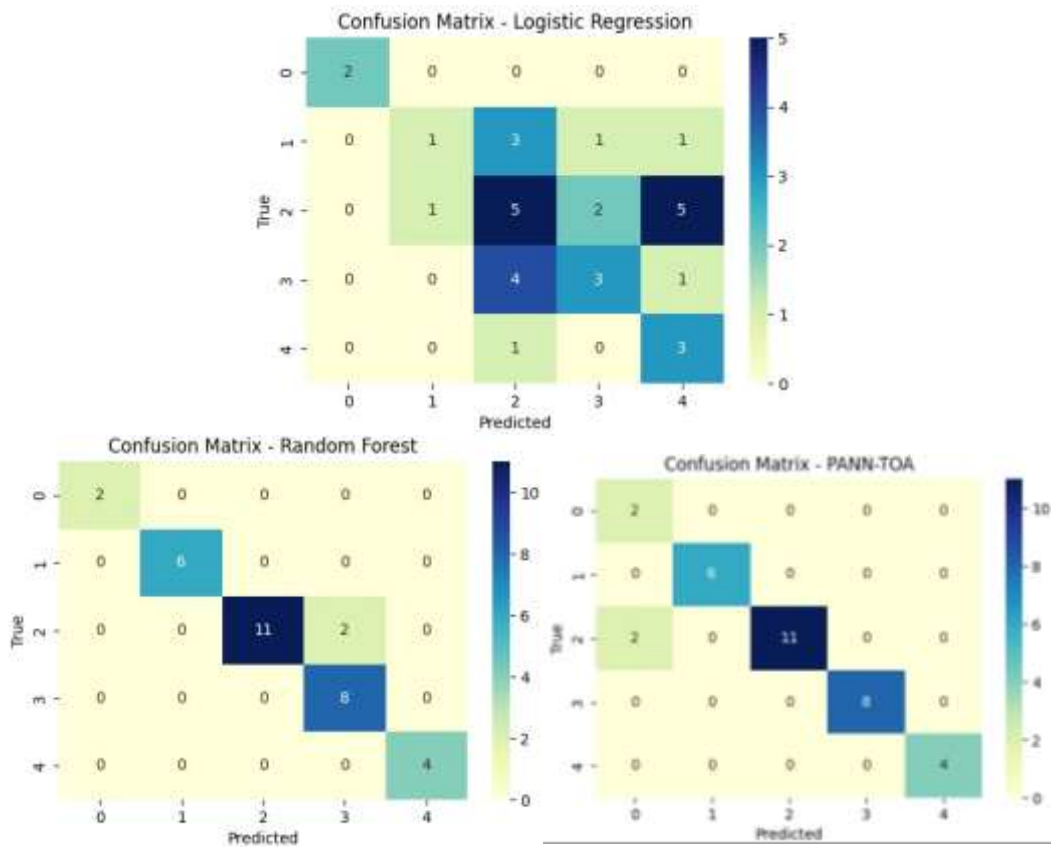


Figure 14: Confusion Matrix for Each Model

The low number of false negatives is particularly encouraging, suggesting the model's reliability in avoiding missed diagnoses.

4.9 Clinical Implications

The findings underscore the importance of specific features in diagnosing PCOS/PCOD, indicating that clinicians should prioritize Insulin levels and BMI in assessments. The PANN model, can serve as decision-support tools to improve diagnostic accuracy.

4.10 Statistical Analysis

A paired t-test was conducted to assess the statistical significance of performance differences between TOA-optimized PANN and baseline models. The results indicated significant improvements ($p < 0.05$), reinforcing the effectiveness of TOA in enhancing model performance.

4.11 Documentation and Reproducibility

To ensure reproducibility, all code and data processing steps are managed using Git, with comprehensive documentation of preprocessing, model configurations, evaluation metrics, and results maintained throughout the study.

5. CONCLUSION

The proposed study effectively illustrates the potential of utilizing a Progressive ANN (PANN) enhanced by the Tarsier Optimization Algorithm (TOA) for the early identification of PCOS and PCOD. By systematically tackling issues related to noisy data and implementing advanced data pre-processing techniques, the notable improvements in both model accuracy and interpretability can be achieved. The findings reveal that the PANN outperformed other ML models, reaching an accuracy of 95.1% and a strong AUC-ROC score of 0.96, highlighting its ability to detect complex patterns within clinical data. The incorporation of TOA played a crucial role in reducing noise and selecting relevant

features, leading to cleaner datasets and emphasizing key factors like insulin levels and BMI, which are vital for clinical evaluations. Additionally, the study underscores the significance of thorough data preprocessing in enhancing the predictive performance of ML models, providing valuable insights for clinicians in the effective diagnosis of PCOS and PCOD. The robust framework established in this research paves the way for advancements in diagnostic methods and personalized treatment approaches, ultimately contributing to improved reproductive health outcomes for those affected. Future research should aim to further validate the model across diverse populations and investigate the inclusion of additional features that may enhance diagnostic precision. Overall, the new study not only addresses a significant gap in the existing literature but also sets a benchmark for the application of advanced ML techniques in clinical practice.

FUNDING STATEMENT

No funding received from any organization.

REFERENCES

1. Maqbool, M., Ara, I. and Gani, I., 2022. The story of polycystic ovarian syndrome: a challenging disorder with numerous consequences for females of reproductive age. *International Journal of Current Research in Physiology and Pharmacology*, pp.19-31.
2. Ali, A.T., Al-Ani, O., Al-Ani, F. and Guidozzi, F., 2022. Polycystic ovary syndrome and metabolic disorders: A review of the literature. *African Journal of Reproductive Health*, 26(8), pp.89-99.
3. Zhu, T., Cui, J. and Goodarzi, M.O., 2021. Polycystic ovary syndrome and risk of type 2 diabetes, coronary heart disease, and stroke. *Diabetes*, 70(2), pp.627-637.
4. Kiconco, S., Teede, H.J., Azziz, R., Norman, R.J. and Joham, A.E., 2021, July. The need to reassess the diagnosis of polycystic ovary syndrome (PCOS): a review of diagnostic recommendations from the international evidence-based guideline for the assessment and management of PCOS. In *Seminars in reproductive medicine* (Vol. 39, No. 03/04, pp. 071-077). Thieme Medical Publishers, Inc..
5. Presswala, B. and De Souza, L.R., 2023. The diagnostic experience of polycystic ovary syndrome: a scoping review of patient perspectives. *Patient Education and Counseling*, 113, p.107771.
6. Rao, V.S., Cowan, S., Armour, M., Smith, C.A., Cheema, B.S., Moran, L., Lim, S., Gupta, S., Manincor, M.D., Sreedhar, V. and Ee, C., 2022. A Global Survey of ethnic Indian women living with polycystic ovary Syndrome: co-morbidities, concerns, diagnosis experiences, Quality of Life, and Use of Treatment methods. *International Journal of Environmental Research and Public Health*, 19(23), p.15850.
7. Verma, P., Maan, P., Gautam, R. and Arora, T., 2024. Unveiling the Role of Artificial Intelligence (AI) in Polycystic Ovary Syndrome (PCOS) Diagnosis: A Comprehensive Review. *Reproductive Sciences*, pp.1-15.
8. Khanwalkar, P., Preethi, M., Divya, K. and Keerthana, K.R., 2024, April. AI-driven Detection of PCOS for Personalized Health Improvement. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (Vol. 1, pp. 1-6). IEEE.
9. Zigarelli, A., Jia, Z. and Lee, H., 2022. Machine-aided self-diagnostic prediction models for polycystic ovary syndrome: observational study. *JMIR Formative Research*, 6(3), p.e29967.
10. Alam Suha, S.A.Y.M.A., 2022. Predicting Polycystic Ovary Syndrome through Machine Learning Technique Using Patients' symptom Data and Ovary Ultrasound

Images (Doctoral dissertation, Department of Computer Science and Engineering, MIST)

11. Sydora, B.C., Wilke, M.S., McPherson, M., Chambers, S., Ghosh, M. and Vine, D.F., 2023. Challenges in diagnosis and health care in polycystic ovary syndrome in Canada: a patient view to improve health care. *BMC Women's Health*, 23(1), p.569.
12. Gupta, R., Tanwar, S., Tyagi, S. and Kumar, N., 2020. Machine learning models for secure data analytics: A taxonomy and threat model. *Computer Communications*, 153, pp.406-440.
13. Lin, H., Xue, Y., Chen, K., Zhong, S. and Chen, L., 2022. Acute coronary syndrome risk prediction based on gradient boosted tree feature selection and recursive feature elimination: A dataset-specific modeling study. *Plos one*, 17(11), p.e0278217.
14. Sharma, T., Mohapatra, A.K. and Tomar, G., 2023. A novel SVM and LOF-based outlier detection routing algorithm for improving the stability period and overall network lifetime of WSN. *International Journal of Nanotechnology*, 20(5-10), pp.759-789.
15. Modi, K., Singh, I. and Kumar, Y., 2023. A comprehensive analysis of artificial intelligence techniques for the prediction and prognosis of lifestyle diseases. *Archives of Computational Methods in Engineering*, 30(8), pp.4733-4756.
16. Stafie, C.S., Sufaru, I.G., Ghiciuc, C.M., Stafie, I.I., Sufaru, E.C., Solomon, S.M. and Hancianu, M., 2023. Exploring the intersection of artificial intelligence and clinical healthcare: a multidisciplinary review. *Diagnostics*, 13(12), p.1995.
17. Meng, C., Trinh, L., Xu, N., Enouen, J. and Liu, Y., 2022. Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Scientific Reports*, 12(1), p.7166.
18. Wagner, M.W., Namdar, K., Biswas, A., Monah, S., Khalvati, F. and Ertl-Wagner, B.B., 2021. Radiomics, machine learning, and artificial intelligence—what the neuroradiologist needs to know. *Neuroradiology*, pp.1-11.
19. McNamara, M.E., Zisser, M., Beevers, C.G. and Shumake, J., 2022. Not just “big” data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions. *Behaviour research and therapy*, 153, p.104086.
20. Wang, R., Gu, Z., Wang, Y., Yin, X., Liu, W., Chen, W., Huang, Y., Wu, J., Yang, S., Feng, L. and Zhou, L., 2022. A “one-stop shop” decision tree for diagnosing and phenotyping polycystic ovarian syndrome on serum metabolic fingerprints. *Advanced Functional Materials*, 32(45), p.2206670.
21. Anand, S., Kumar, A., Prasad, A. and Trivedi, K., 2022. Updated meta-analysis on the diagnostic accuracy of serum anti-Mullerian hormone in poly cystic ovary syndrome involving 13 509 subjects. *Journal of Obstetrics and Gynaecology Research*, 48(8), pp.2162-2174.
22. Lakshmi, M.J., Spandana, D.S., Raj, H., Niharika, G., Kodipalli, A., Kamal, S. and Rao, T., 2023, April. Prediction of PCOS and PCOD in Women Using ML Algorithms. In *International Conference on Information and Communication Technology for Intelligent Systems* (pp. 97-115). Singapore: Springer Nature Singapore.
23. Gaur, Y., Chaudhary, A. and Abraham, A., 2023, December. Analysing Polycystic Ovary Syndrome Using Machine Learning Models. In *International Conference on Intelligent Systems Design and Applications* (pp. 467-475). Cham: Springer Nature Switzerland.
24. Kite, C., 2020. The effectiveness of exercise in the treatment of polycystic ovary syndrome (Doctoral dissertation, Aston University).
25. Hasan, S.S., 2021. Evaluation of Vaspin Levels and Some Biochemical Variables in poly cystic Ovary Syndrome Patients in Diyala Governorate (Doctoral dissertation, College of Science Department of Chemistry Evaluation of Vaspin Levels and Some

- Biochemical Variables in poly cystic Ovary Syndrome Patients in Diyala Governorate
A Thesis Submitted to the Council of the College of Science, University of Diyala).
26. Bansal, C., Handa, P. and Goel, N., 2023, September. Comparing different models for polycystic ovary syndrome diagnosis: An empirical investigation on a large clinical dataset. In 2023 IEEE Women in Technology Conference (WINTeCHCON) (pp. 1-6). IEEE.
 27. Srinithi, V. and Rekha, R., 2023, February. Machine learning for diagnosis of polycystic ovarian syndrome (PCOS/PCOD). In 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS) (pp. 19-24). IEEE.
 28. Sinthia, G., Poovizhi, T. and Khilar, R., 2022. Analysis on polycystic ovarian syndrome and comparative study of different machine learning algorithms. In *Advances in Intelligent Computing and Communication: Proceedings of ICAC 2021* (pp. 191-196). Singapore: Springer Nature Singapore.
 29. Tiwari, S., Kane, L., Koundal, D., Jain, A., Alhudhaif, A., Polat, K., Zaguia, A., Alenezi, F. and Althubiti, S.A., 2022. SPOSDS: A smart Polycystic Ovary Syndrome diagnostic system using machine learning. *Expert Systems with Applications*, 203, p.117592.
 30. Kumar, A., Singh, J. and Khan, A.A., 2024. A comprehensive machine learning framework with particle swarm optimization for improved polycystic ovary syndrome (PCOS) diagnosis. *Engineering Research Express*, 6(3), p.035233.
 31. Dutta, P., Paul, S. and Majumder, M., 2021. An efficient SMOTE based machine learning classification for prediction & detection of PCOS.
 32. Paramasivam, G.B. and Ramasamy Rajammal, R., 2024. Modelling a self-defined CNN for effectual classification of PCOS from ultrasound images. *Technology and Health Care*, 32(5), pp.2893-2909.
 33. Jain, P., Mishra, R.K., Deep, A. and Jain, N.K., 2024. Xplainable AI for deep learning model on PCOD analysis. In *XAI Based Intelligent Systems for Society 5.0* (pp. 131-152). Elsevier.
 34. Citron, N., 2022. The Role of Disclosure and Social Support on Quality of Life in Women with Polycystic Ovary Syndrome (PCOS) (Master's thesis, University of Windsor (Canada)).
 35. Liu, T., Li, B., Yang, L., Qiao, J. and Chen, W., 2023. Tightly coupled integration of GNSS/UWB/VIO for reliable and seamless positioning. *IEEE Transactions on Intelligent Transportation Systems*.
 36. Suha, S.A. and Islam, M.N., 2023. Exploring the dominant features and data-driven detection of polycystic ovary syndrome through modified stacking ensemble machine learning technique. *Heliyon*, 9(3).
 37. Dutta, P., Paul, S. and Majumder, M., 2021. An efficient SMOTE based machine learning classification for prediction & detection of PCOS.
 38. Kapadia, D. and Jain, R., 2023, November. Pcos Prediction: Advancements in Medical Informatics Using Artificial Intelligence. In 2023 2nd International Conference on Futuristic Technologies (INCOFT) (pp. 1-5). IEEE.
 39. Chitra, P., Srilatha, K., Sumathi, M., Jayasudha, F.V., Bernatin, T. and Jagadeesh, M., 2023, March. Classification of ultrasound pcos image using deep learning based hybrid models. In 2023 Second International Conference on Electronics and Renewable Systems (ICEARS) (pp. 1389-1394). IEEE.
 40. Vats, S., Sengupta, A., Chaurasia, A. and Narad, P., 2022. Combination of Expression Data and Predictive Modelling for Polycystic Ovary Disease and Assessing Risk of Infertility Using Machine Learning Techniques. In *Innovations in Computational Intelligence and Computer Vision: Proceedings of ICICV 2021* (pp. 547-555). Singapore: Springer Nature Singapore.

41. Pathak, H., Handa, P. and Goel, N., 2023, December. Comparative Analysis of Feature Selection Methods for Automatic Classification of PCOD. In 2023 10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON) (Vol. 10, pp. 235-241). IEEE.
42. Sinthia, G., Poovizhi, T. and Khilar, R., 2022. Analysis on polycystic ovarian syndrome and comparative study of different machine learning algorithms. In *Advances in Intelligent Computing and Communication: Proceedings of ICAC 2021* (pp. 191-196). Singapore: Springer Nature Singapore.
43. Elmannai, H., El-Rashidy, N., Mashal, I., Alohal, M.A., Farag, S., El-Sappagh, S. and Saleh, H., 2023. Polycystic ovary syndrome detection machine learning model based on optimized feature selection and explainable artificial intelligence. *Diagnostics*, 13(8), p.1506.