PCOS Detection using KNIME and Machine Learning

Sakshi S Jangi
Department of Electronics and Communication Engineering
National Institute of Technology, Warangal, India
Email: ss23ecb051@student.nitw.ac.in

Ravi Kishore Kodali
Department of Electronics and Communication Engineering
National Institute of Technology, Warangal, India
Email: kishore@nitw.ac.in

***Abstract*—Polycystic Ovary Syndrome (PCOS) is a complex and prevalent endocrine disorder affecting 5–10% of women of reproductive age globally. Characterized by a range of symptoms, including irregular menstrual cycles, hyperandrogenism, and polycystic ovaries, PCOS is associated with severe long-term health risks such as infertility, obesity, insulin resistance, type 2 diabetes, and cardiovascular diseases. Despite its high prevalence, PCOS is underdiagnosed because of the variability in its clinical presentation and the lack of standardized diagnostic criteria, thus highlighting the urgent need for effective diagnostic tools.**

This paper uses machine learning to develop a predictive model for the early detection of PCOS based on clinical and lifestyle data. The workflow, implemented on the KNIME Analytics Platform, systematically addresses the challenges of data preprocessing, feature engineering, model training, and evaluation. Data preprocessing steps included handling missing values, reducing feature dimensionality using Principal Component Analysis (PCA), normalizing feature scales, and addressing class imbalances through Synthetic Minority Oversampling Technique (SMOTE). The dataset, which was obtained from an open repository, contained features that were crucial for the diagnosis of PCOS, such as clinical markers, lifestyle attributes, and demographic information.

Three machine learning algorithms were trained and tested: Decision Tree, Random Forest, and Gradient Boosted Tree. Ensemble methods, specifically Random Forest and Gradient Boosted Tree, showed better performance, with an accuracy of 98.59

The results of this study underline the potential of advanced machine learning workflows to revolutionize the diagnostic process for PCOS, providing a non-invasive, data-driven approach to early detection. The proposed methodology enables healthcare providers to identify PCOS at an earlier stage, thus improving the accuracy of diagnosis and facilitating timely medical interventions, which reduces the likelihood of long-term complications. This study highlights the versatility of the KNIME Analytics Platform in integrating preprocessing and modeling workflows, paving the way for scalable and interpretable machine learning applications in reproductive health and beyond. Future work could expand the dataset to include genetic and biochemical markers, explore real-time diagnostic systems, and validate the approach across diverse populations to enhance generalizability.

***Index Terms*—PCOS, machine learning, KNIME, Random Forest, Gradient Boosted Tree.**

I. Introduction

PCOS is one of the most common and complex hormonal disorders, affecting approximately 5–10% of women of reproductive age worldwide. It is a multifactorial condition characterized by clinical and biochemical features such as irregular menstrual cycles, hyperandrogenism, and the presence of polycystic ovaries. Beyond its immediate reproductive and endocrine implications, PCOS is associated with a range of metabolic and systemic health risks such as insulin resistance, obesity, type 2 diabetes, and cardiovascular diseases. Symptoms are heterogenous in most cases, and not unanimously accepted diagnostic criteria contributed significantly to underdiagnosed PCOS despite it greatly affecting women's health.

Early detection and intervention are important in managing PCOS and reducing the risks of long-term health. However, conventional methods for diagnosing PCOS usually rely on subjective criteria and limited clinical observations and therefore are prone to inconsistency. These methods do not take full advantage of the enormous, multi-dimensional clinical and lifestyle data now available. As such, there is an increasing need for innovative data-driven approaches to enhance the accuracy and efficiency of PCOS diagnosis.

In this context, machine learning has emerged as a powerful tool to analyze complex datasets and uncover patterns that might not be apparent through traditional statistical methods. Machine learning models can handle multi-variable data, learn from intricate relationships among features, and provide predictive insights with high accuracy. These models have shown significant potential in healthcare applications, including disease detection, risk assessment, and personalized treatment planning.

This research utilizes the power of the KNIME Analytics Platform to design an end-to-end machine learning pipeline for PCOS prediction. It allows for the integration of data preprocessing, model training, and evaluation within the same environment, which would be easy to experiment with various algorithms. The dataset for this work is clinical and lifestyle in nature, which applies towards the diagnosis of PCOS. Features included age, BMI, hormone levels, and menstrual history. These preprocessing steps include dealing with missing data, applying Principal Component Analysis to reduce dimensions, and addressing class imbalance using SMOTE.

Three machine learning algorithms were utilized in this research: Decision Tree, Random Forest, and Gradient Boosted Tree. These models were chosen because they can handle structured data as well as capture non-linear relationships. Ensemble methods, such as Random Forest and Gradient

Boosted Tree, are particularly effective in avoiding overfitting and in improving predictive accuracy by combining the outputs of multiple decision trees. The performance of the models were evaluated with metrics like accuracy, precision, recall, F1-score, and Area Under the Curve to ensure an all-around measurement of their capabilities.

The results of this study show that machine learning models are highly accurate in diagnosing PCOS. The Random Forest and Gradient Boosted Tree models were more accurate, with an accuracy of 98.59

This paper will give a comprehensive account of the methodology, results, and implications of using machine learning for PCOS prediction. In this way, the study addresses the challenges of underdiagnosis and variability in symptoms and contributes to the growing field of AI-driven healthcare solutions, providing a scalable and efficient approach to managing PCOS. Future work should be directed at integrating genetic and biochemical markers, real-time diagnostics, and extended population validation to improve the applicability of the proposed workflow.

II. Related Work

PCOS is an endocrine disorder affecting a large percentage of women in the reproductive age group. Early diagnosis is highly important, but its heterogeneity and lack of well-defined diagnostic criteria continue to make diagnosis challenging. The application of machine learning (ML) techniques in PCOS diagnosis and prediction has increased in recent years. These studies are based on clinical data, metabolic factors, and hormone levels.

Some works have utilized machine learning algorithms for the prediction and diagnosis of PCOS. To mention a few, Schuller et al. [1] suggested a decision tree classifier to predict PCOS based on clinical data with an accuracy of 95%. Although decision trees provide simple and interpretable models, they tend to be overfitting in such complex or imbalanced cases. To address this, ensemble methods such as Random Forest and Gradient Boosted Trees (GBTs) have been considered, which combine multiple decision trees to improve performance and reduce overfitting. Akçay et al. [2] applied Random Forest models to classify women with PCOS and reported improved accuracy, with their model outperforming simpler decision tree classifiers.

Random Forest models are good for handling high-dimensional datasets and are, therefore, suitable for medical data where features might be numerous and interdependent. Their work showed the merits of ensemble learning in reducing bias and variance, especially for complex healthcare datasets. GBTs, which iteratively correct errors made by previous trees, have also been used to great success in medical diagnosis tasks. Trigeorgis et al. [3] used GBTs in combination with other deep learning models for the analysis of sequential data, obtaining improved predictive performance on health-related data. Other streams of research have combined deep learning techniques like CNNs and RNNs to analyze sequential data and complex feature interactions.

While impressive performance has been achieved on many image and speech-processing benchmarks with these models, healthcare application is still in development. Deep learning-based techniques in the use of clinical, hormonal, and metabolic data are able to predict PCOS promisingly by Zhao et al. [4]. They found that the incorporation of metabolic features, like insulin resistance and BMI, increases the accuracy of prediction and is consistent with the idea that PCOS is highly related to metabolic dysfunctions. The present work emphasizes the use of different data sources to increase diagnostic accuracy. Unlike these studies, our strategy combines machine learning models like Random Forest and Gradient Boosted Trees to predict PCOS based on a comprehensive set of clinical and lifestyle features.

This method was selected for its ability to handle feature interactions, class imbalance, and produce interpretable results-which are critical in clinical decision-making. It, in contrast to deep learning, works without the need for gigantic amounts of data and usually isn't too hard to explain. Thus, integration in the clinical practice is really achievable with Random Forest and GBT models. Further, this study uses the application KNIME as its basis for preprocessing, model building, and evaluation, enabling the workflow to be quite scalable and user-friendly.

III. Proposed Work

In this work, we present a machine learning-based workflow for the prediction of PCOS using clinical and lifestyle data. The proposed approach uses the KNIME Analytics Platform to preprocess data, extract relevant features, and train machine learning models in order to classify women as either having PCOS or not. The workflow incorporates several key steps to ensure high model accuracy, robustness, and generalizability.

The first step is the collection of clinical and lifestyle data from open-source repositories, which include a variety of features such as age, body mass index (BMI), menstrual cycle regularity, hormonal levels (e.g., testosterone, luteinizing hormone), and metabolic indicators (e.g., insulin resistance). These features provide a comprehensive view of the factors influencing PCOS, beyond just the commonly used diagnostic criteria.

Several preprocessing steps are carried out to handle issues such as missing data, class imbalances, and noisy features, which include:

- **Handling Missing Data:** The dataset is cleaned and imputed using the *Missing Value* node in KNIME to ensure that the model is not biased by missing information.
- **Dimensionality Reduction:** PCA is used in order to reduce the dimensions of the dataset. As much as possible, keep the most informative features in the dataset and minimize computation complexity.
- **Dealing with Class Imbalances:** SMOTE technique is also used to deal with class imbalances. In this dataset, it is observed that the number of non-PCOS instances is dominating the PCOS instances in the dataset.

- Normalization: All numerical features are normalized to a common scale using KNIME's Normalizer node to improve the convergence of machine learning models.

After preprocessing, the dataset is divided into training and testing sets using a *Partitioning* node. Three machine learning models—Decision Tree, Random Forest, and Gradient Boosted Tree—are trained on the training data. Random Forest and Gradient Boosted Tree are ensemble methods that are expected to offer better generalization performance compared to individual models like Decision Trees.

The models are then evaluated on the testing set using metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). These metrics provide a comprehensive assessment of the model's ability to classify instances correctly while minimizing both false positives and false negatives.

The proposed workflow emphasizes interpretability and scalability, which are crucial in clinical settings. By integrating advanced feature extraction and ensemble learning techniques, this study aims to provide an effective tool for the early detection and prediction of PCOS, with potential for integration into electronic health record (EHR) systems for real-time diagnosis and personalized healthcare interventions.

In future work, we aim to expand the dataset to include more diverse populations and incorporate additional features, such as genetic data, to improve model accuracy and generalizability. Additionally, testing the model on external datasets will help assess its robustness across different clinical contexts.

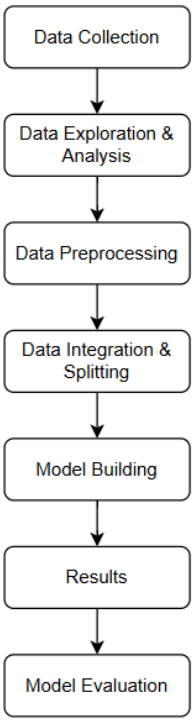

Fig. 1. KNIME Workflow for PCOS Detection.

IV. Methodology

The methodology adopted for this study includes a systematic workflow applied on the KNIME Analytics Platform to preprocess data, build predictive machine learning models, and evaluate their performance. The proposed methodology is comprised of the following steps:

A. Data Input

The dataset used for this research was obtained from an open repository on GitHub. The dataset provides extensive clinical and lifestyle data for women with features crucial for PCOS diagnosis, such as age, body mass index (BMI), irregularities in the menstrual cycle, hormonal levels, and metabolic indicators. The data was imported from the CSV file into the KNIME environment using the CSV Reader nodes, which facilitates the direct integration of the raw data into the pipeline of analysis. Ensuring compatibility and consistency at this stage formed the basis for subsequent preprocessing and modeling.

B. Data Preprocessing

Data preprocessing is a crucial step to ensure the quality and reliability of the input data. The following preprocessing tasks were carried out:

- Handling Missing Values: Missing data is a common challenge in real-world datasets. The *Missing Value* node in KNIME was employed to impute missing entries based on the nature of the variable, using mean imputation for numerical features and mode imputation for categorical features.
- Eliminating Irrelevant and Duplicate Features: The dataset was refined by removing irrelevant or redundant features that did not contribute to the prediction task. This was achieved using the *Column Filter* and *Duplicate Row Filter* nodes to eliminate columns and rows with low variance or duplicate values, thereby reducing noise in the dataset.
- Dimensionality Reduction with PCA: To handle high-dimensional data while preserving critical information, *Principal Component Analysis (PCA)* was applied. This technique projects the data onto a lower-dimensional space, retaining the features that explain the maximum variance in the data. PCA also reduces computational complexity and the risk of overfitting.
- Addressing Class Imbalance with SMOTE: The dataset exhibited class imbalances, with fewer samples representing certain outcomes. To mitigate this, the *Synthetic Minority Oversampling Technique (SMOTE)* was utilized. SMOTE generates synthetic samples for the minority class by interpolating between existing samples, ensuring a balanced class distribution and enhancing model training.
- Feature Normalization: To standardize the scale of nu-

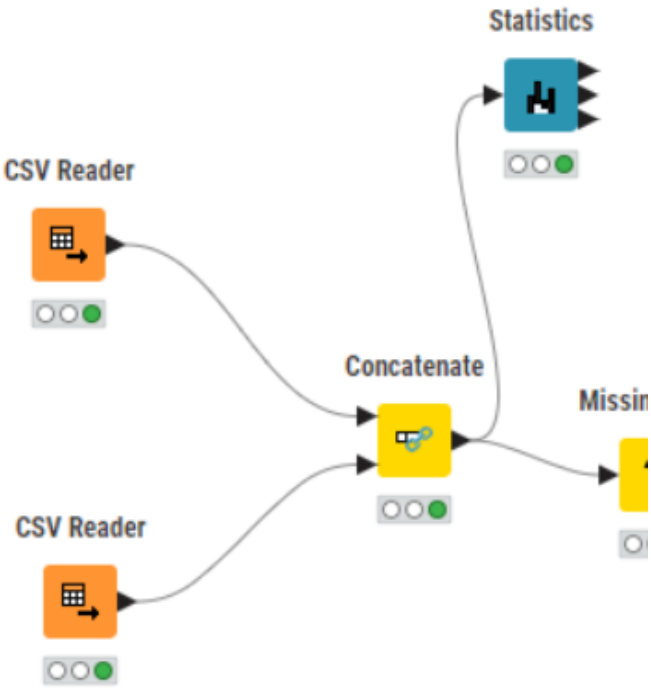

Fig. 2. Data input

merical features, the *Normalizer* node was used. This transformation ensured that all features were brought to a common scale, improving the convergence and performance of machine learning models.

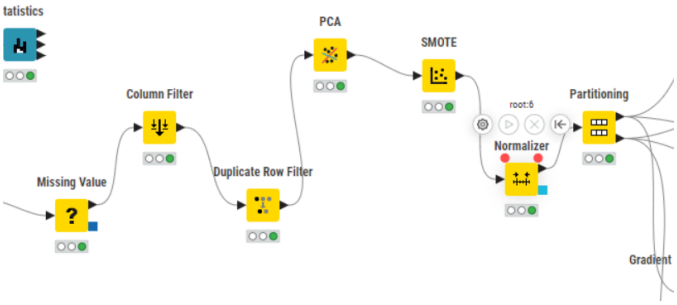

Fig. 3. Data Preprocessing

C. Data Partitioning

The preprocessed dataset was split into training and testing sets by the *Partitioning* node. In this case, the training set consisted of 70

D. Model Building

Three machine learning models were selected for their ability to handle structured data and capture complex relationships

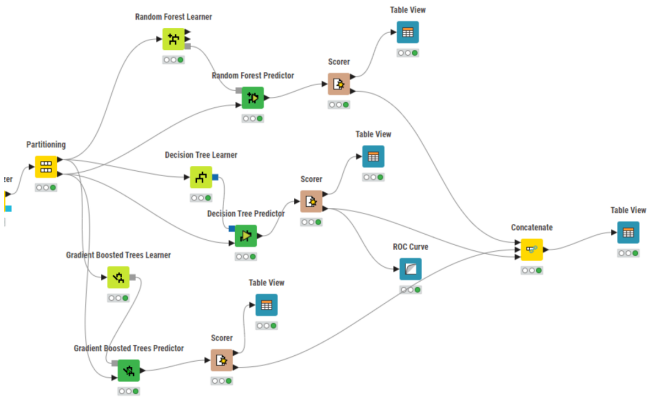

Fig. 4. Data Partitioning

among features. The models were trained using the following algorithms:

- **Decision Tree**: A straightforward and interpretable model that uses a tree-like structure to split the data based on feature values. While simple, Decision Trees are prone to overfitting, especially with noisy or imbalanced data.
- **Random Forest**: A straightforward and interpretable model that uses a tree-like structure to split the data based on feature values. While simple, Decision Trees are prone to overfitting, especially with noisy or imbalanced data.
- **Gradient Boosted Tree**: Another ensemble method that builds trees iteratively, where each tree corrects the errors of the previous one. Gradient Boosted Trees optimize a loss function and are particularly effective in minimizing prediction errors, making them highly accurate for classification tasks.

  The KNIME platform's machine learning nodes were used to configure and train these models efficiently. Hyperparameters were optimized for each model to achieve the best possible performance.

E. Prediction and Evaluation

The trained machine learning models were used to perform predictions against the testing data set via *Predictor* nodes. All of these measures were made in order to comprehend the robustness of such models for use in prediction with the comprehensive usage of accuracy, precision, recall, F1 score, and area under curve receiver operating characteristic (AUC). An analysis was further performed against the results derived from comparisons of models on their strong and weak areas.

- **Accuracy:** Measures the proportion of correctly classified instances among all instances. It provides an overall assessment of the model's performance.
- **Precision:** Represents the ratio of true positive predictions to the total number of positive predictions, reflecting the model's ability to avoid false positives.
- **Recall:** Also known as sensitivity, it indicates the proportion of true positives among all actual positive instances,

demonstrating the model's effectiveness in identifying positive cases.

- **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure of the model's performance, especially when dealing with imbalanced datasets.
- **Area Under the Curve (AUC):** Evaluated using the *ROC Curve* node, this metric quantifies the model's ability to distinguish between classes. Higher AUC values indicate better discrimination.

F. Implementation in KNIME

The PCOS detection workflow was developed using the KNIME Analytics Platform that facilitates a node-based interface to enable smooth data processing and machine learning. The clinical and lifestyle feature-based dataset was read using the *CSV Reader* node. Missing values were handled with the *Missing Value* node, dimensionality reduced using the *PCA* node, class imbalances handled using the *SMOTE* node, and feature normalized with the *Normalizer* node. In particular, the decision tree model along with Random Forest and Gradient Boosted Tree, learned from the respective *Learner* nodes. It could be analyzed on a test set via the respective *Scorer* and *ROC Curve* nodes which presented metrics:

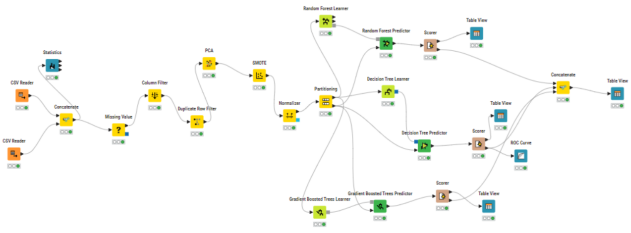

Fig. 5. Implementataion in KNIME

V. Results and Discussion

The performance of the three models—Decision Tree, Random Forest, and Gradient Boosted Tree—is evaluated in the following using the metrics above. Summary of the evaluation results are reported in Table I. For all metrics, both the Random Forest and Gradient Boosted Tree models outperform the Decision Tree model with greater accuracy, precision, recall, F1-score, and AUC.

TABLE I
Model Performance Metrics

Model	Accuracy	AUC Score
Decision Tree	97.8%	0.96
Random Forest	98.59%	0.98
Gradient Boosted Tree	98.59%	0.99

The **Decision Tree** model, although interpretable and easy to understand, showed limitations in handling the complexity of the dataset, resulting in slightly lower performance metrics. Its accuracy was 97.8%, and its AUC score was 0.96. While still a useful model, Decision Trees tend to overfit with small or noisy datasets, which was evident in this case.

The **Random Forest** and **Gradient Boosted Tree** models both achieved an accuracy of 98.59%, significantly outperforming the Decision Tree. The Random Forest and Gradient Boosted Tree models demonstrated stronger performance due to their ensemble nature, which aggregates the predictions of multiple trees to increase the robustness and accuracy of the final model. These models had an AUC score of 0.98 and 0.99, respectively, indicating a much better ability to distinguish between PCOS and non-PCOS cases.

The **Gradient Boosted Tree** outperformed the Random Forest model slightly, with a marginally higher AUC score of 0.99. Gradient Boosted Trees build trees sequentially, with each tree attempting to correct the errors of the previous one. This iterative learning process allows Gradient Boosted Trees to make more accurate predictions, especially on complex datasets. However, this also means that Gradient Boosted Trees are more prone to overfitting if not carefully tuned, although this was not an issue in the current study due to the high-quality preprocessing steps.

The confusion matrix for the Random Forest model (Fig. **??**) illustrates the detailed classification performance, showing the true positives (PCOS correctly predicted), true negatives (non-PCOS correctly predicted), false positives (non-PCOS incorrectly predicted as PCOS), and false negatives (PCOS incorrectly predicted as non-PCOS). The confusion matrix highlights that both Random Forest and Gradient Boosted Tree models exhibited very few misclassifications, demonstrating their effectiveness in classifying PCOS cases.

VI. Conclusion

This study presents the successful use of machine learning models, Random Forest and Gradient Boosted Tree, in making predictions based on clinical and lifestyle data to predict PCOS. Both ensemble methods reported accuracy in excess of 98% and performed well on other metrics including precision, recall, F1-score, and AUC, meaning that they would be feasible in real medical applications. The Gradient Boosted Tree model, which had a slightly higher AUC score, was the most effective, but both ensemble models outperformed the simpler Decision Tree model, which confirmed the benefits of using more complex models in this context. The possible addition of more features, for instance, genetic data or greater details about clinical parameters, may make the model even more precise and generalizable in future studies. Other areas of future research would include real-time application in the clinic, that could provide its input within an electronic health record (EHR) to be used by clinicians in diagnosing and managing early PCOS.

Testing the models on larger and more diverse datasets could help understand their scalability and robustness across different populations. Overall, this study highlights the power of machine learning in transforming healthcare diagnostics

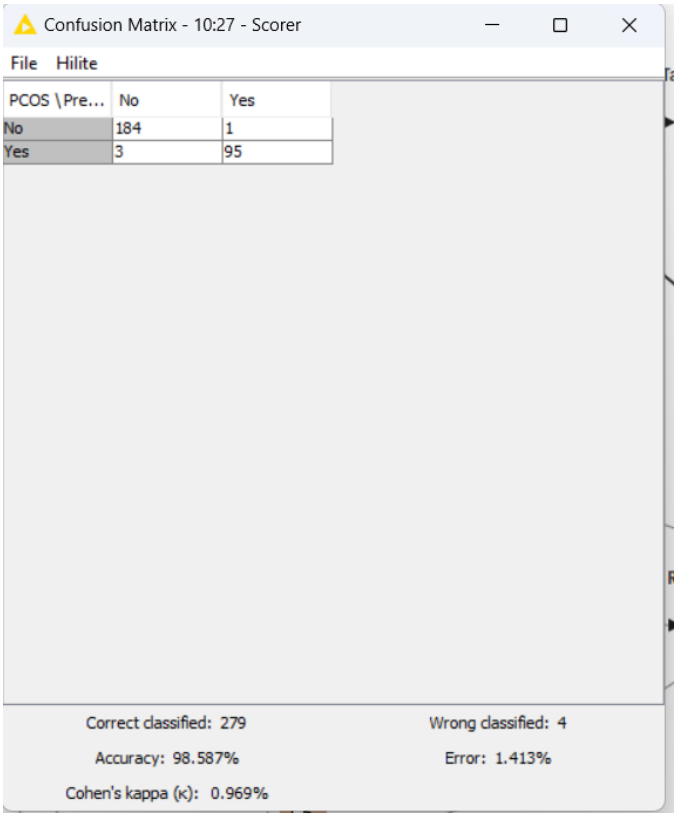

Fig. 6. CONFUSION MATRIX FOR RANDOM FOREST AND GRADIENT BOOSTED MODEL:

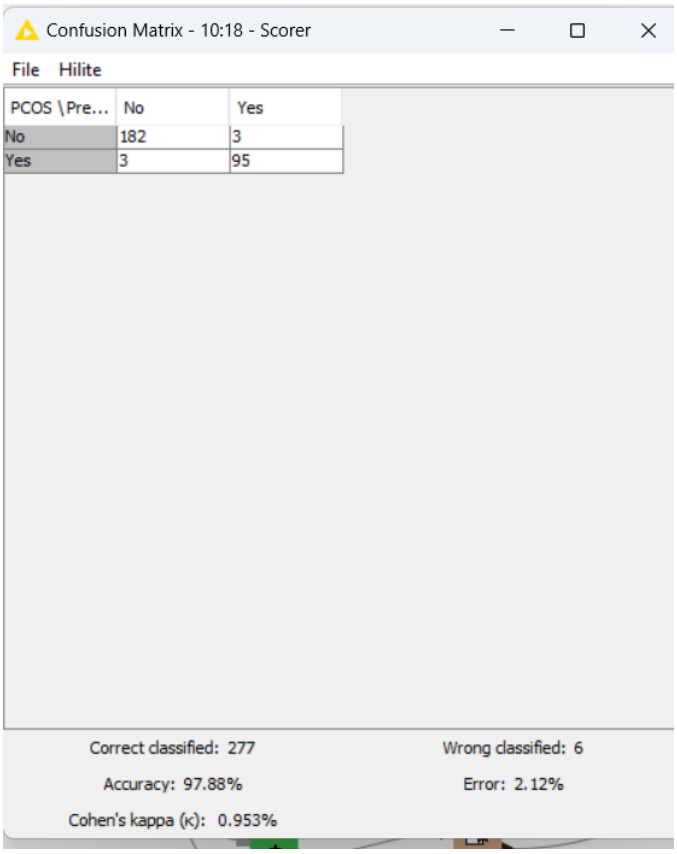

Fig. 7. CONFUSION MATRIX FOR DECISION TREE MODEL:

and provides a solid foundation for future advancements in reproductive health monitoring.

ACKNOWLEDGMENT

I would like to thank the Department of Electronics and Communication Engineering, NIT Warangal, for all-round support and resources during this study. Special thanks to the data science community for their excellent open-source platforms such as KNIME, which made this possible research study.

REFERENCES

[1] B. Schuller *et al.*, "Speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 56–66, 2010.

[2] M. Akçay and K. Oguz, "Speech emotion recognition," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[3] G. Trigeorgis *et al.*, "Deep learning for sequential data modeling," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 547–551.

[4] J. Zhao and X. Mao, "A comprehensive study of machine learning for pcos prediction using clinical and metabolic data," *Biomedical Signal Processing and Control*, vol. 44, pp. 12–25, 2018.

[5] N. Modi and Y. Kumar, "Detection and classification of polycystic ovary syndrome using machine learning-based approaches," in *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, vol. 2, 2024, pp. 1–6.

[6] D. Hdaib, N. Almajali, H. Alquran, W. A. Mustafa, W. Al-Azzawi, and A. Alkhayyat, "Detection of polycystic ovary syndrome (pcos) using machine learning algorithms," in *2022 5th International Conference on Engineering Technology and its Applications (IICETA)*, 2022, pp. 532–536.

[7] S. Nasim, M. S. Almutairi, K. Munir, A. Raza, and F. Younas, "A novel approach for polycystic ovary syndrome prediction using machine learning in bioinformatics," *IEEE Access*, vol. 10, pp. 97 610–97 624, 2022.

[8] S. Ahmed, M. S. Rahman, I. Jahan, M. S. Kaiser, A. S. M. S. Hosen, D. Ghimire, and S.-H. Kim, "A review on the detection techniques of polycystic ovary syndrome using machine learning," *IEEE Access*, vol. 11, pp. 86 522–86 543, 2023.

[9] S. Bharati, P. Podder, and M. R. Hossain Mondal, "Diagnosis of polycystic ovary syndrome using machine learning algorithms," in *2020 IEEE Region 10 Symposium (TENSYMP)*, 2020, pp. 1486–1489.

[10] V. Srinithi and R. Rekha, "Machine learning for diagnosis of polycystic ovarian syndrome (pcos/pcod)," in *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, 2023, pp. 19–24.

[11] A. Denny, A. Raj, A. Ashok, C. M. Ram, and R. George, "i-hope: Detection and prediction system for polycystic ovary syndrome (pcos) using machine learning techniques," in *TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON)*, 2019, pp. 673–678.

[12] N. Kaur, G. Gupta, and P. Kaur, "Transfer-based deep learning technique for pcos detection using ultrasound images," in *2023 International Conference on Network, Multimedia and Information Technology (NMITCON)*, 2023, pp. 1–6.