International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(Research Article)

Predicting polycystic ovary syndrome using SVM

Tanvir Mahmud [1, 2, *] and S A Sabbirul Mohosin Naim [3]

[1] *Department of Electrical and Electronic Engineering, Daffodil International University, Dhaka, Bangladesh.*
[2] *Department of Electrical and Computer Engineering, Lamar University, Beaumont, Texas 77710, USA.*
[3] *Department of Electrical Engineering, School of Engineering, San Francisco Bay University, Fremont, CA, USA.*

International Journal of Science and Research Archive, 2024, 13(02), 4400-4408

Publication history: Received on 28 September 2024; revised on 18 December 2024; accepted on 21 December 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.2.2153

Abstract

Polycystic ovary syndrome (PCOS) has been classified as a severe health problem common among women globally. Early detection and treatment of PCOS reduce the possibility of long-term complications, such as increasing the chances of developing type 2 diabetes and gestational diabetes. Therefore, effective and early PCOS diagnosis will help the healthcare systems to reduce the disease's problems and complications. Machine learning (ML) and ensemble learning have recently shown promising results in medical diagnostics. The main goal of our research is to provide model explanations to ensure efficiency, effectiveness, and trust in the developed model through local and global explanations. Feature selection methods with different types of SVM models are used to get optimal feature selection and best model.

Keywords: Polycystic Ovarian Syndrome; Machine Learning; PCOS Predict; Support Vector Machine; linear SVM; RBF; polynomial SVM

1. Introduction

PCOS is the most common endocrine disorder in women of reproductive age [1]. PCOS is characterized by the ovaries production of an abnormal number of androgens which are male sex hormones that are normally present in women in small amounts. These androgens can cause more problems with the women's menstrual cycle and' are a reason of many PCOS features [2]. PCOS has many symptoms including irregular menstrual cycles, acne, heavy periods, excess hair growth, thickened and dark areas of skin, weight gain, pelvic pain, oily skin, and difficulty in getting pregnant. It is specified by hyperandrogenism, insulin resistance, anovulation where the ovary does not release an oocyte during the menstrual cycle, and neuroendocrine disruption [3–4].

PCOS diagnosis can be tricky, because not everyone with PCOS has polycystic ovaries (PCO), nor does everyone with ovarian cysts have PCOS, so the pelvic ultrasound as a standalone diagnosis is not sufficient [5]. The full diagnostic plan is mainly a combination of a pelvic ultrasound besides blood tests of specific parameters that indicate the presence of PCOS. PCOS can be diagnosed well in adults opposed to the diagnosis of adolescents where in this age group, the symptoms of the PCOS is overlapped with the characteristics of puberty. The diagnosis in adults is set by three distinct sets, one is determined by the National Institute of Health Consensus Statement which defines PCOS as menstrual irregularity and evidence of hyperandrogenism [6]. Another set is determined by the ESHR/ASRM where the PCOS is defined as two of three features including anovulation or oligo-ovulation, hyperandrogenism, and polycystic ovaries by ultrasound [7]. The third and final set is defined by the Androgen Excess and PCOS Society, that diagnose PCOS as hyperandrogenism with ovarian dysfunction of polycystic ovaries [8]. Although the PCOS diagnosis is determined by one of the three sets, it is still difficult to certainly diagnose it, one reason according to Dr. Darche is there is no universal definition of the condition, "There are multiple expert-derived criteria for the syndrome, which means there is no universal diagnostic test or algorithm that doctors used to assess patients" she said. The other reason is that symptoms

[*] Corresponding author: Tanvir Mahmud

vary between women and does not affect them the same, this makes the diagnosis even more ambiguous for doctors. Furthermore, symptoms might not necessarily point to PCOS, but could be related to other endocrine issues, obesity, and hypothyroidism [9].

Since PCOS is a hard-to-diagnose widespread hormonal disorder, blood tests, symptoms, and other parameters with the help of a computer can form a new and easy method to diagnose it. By collecting clinical data and building a model by writing algorithms, Machine Learning has shown its efficiency in the health sector when it comes to diagnosing diseases accurately [10-15].

2. Literature Review

The authors applied ML models to PCOS from Kaggle to predict PCOS. For example, in [16], the authors applied gradient boosting, RF, LR, and a hybrid RFLR model that integrated RF with LR with a univariate feature selection (UFS) algorithm from the PCOS dataset. They split the dataset using holdout and cross-validation methods to train and test models. The result showed that RFLR with UFS achieved the highest performance.

In [17], the authors reduced the number of features using Principal Component Analysis (PCA). They applied NB, KNN, LR, RF, and SVM with selected features to predict PCOS. The result showed that RF achieved the highest accuracy. In [6], the authors used correlation feature selection methodology to select a subset of features from the database. They applied different ML models: SVM, LR, RF, DT, KNN, Quadratic Discriminant Analysis (QDA), Linear Discriminant Analysis (LDA), GB, AdaBoost (AB), XGBoost (XB), and CatBoost, and obtained the optimal model based on correlation thresholds. The result showed that RF was the optimal model.

In [18], the authors compared different models, i.e., CNN, ANN, SVM, DT, and KNN, and applied feature selection methods to diagnose PCOS. RF achieved the best-performing model. In [19], the authors utilized Pearson correlation to determine the best features. The applied SVM, RF, and XG boost multi-layer perceptron with selected features to detect the accuracy rate of their SVM have the highest rate. In [20], the authors proposed a hybrid feature selection approach using filters and wrappers to reduce the number of features. Furthermore, they applied different ML models with selected features to predict PCOS. SVM achieved the highest accuracy.

In [21], they applied SVM, LR, NB, and KNN to detect whether a woman was suffering from PCOS. They used chi-square feature selection methods to select the top 30 features. The accuracy of RF has achieved the highest rate. In [16], the authors used RF, DT, SVM, LR, KNN, XGBRF, and CatBoost Classifier to detect whether a woman was suffering from PCOS. The result showed that CatBoost recorded the highest accuracy.

In [22], the authors used Gini importance to select features. They applied different ML models: KNN, DT, SVM, LR, and NB, to detect PCOS. Based on the accuracy, DT recorded the highest rate. In [23], the authors applied CatBoost, RF, LR, NB, DT, SVM, and DT. Furthermore, they compared their outcomes in terms of the evaluation matrix. CatBoost has the highest accuracy in predicting whether a woman should seek medical help for PCOS. In [24], the authors applied Chi-Square, ANOVA, and Mutual Information to identify insignificant features from the data. They used selected features to detect PCOS by applying SVM, LR, DT, NB, XGBRF, RF, and CatBoost. The CatBoost classifier performed with the best accuracy.

In [25], the authors used ML models: LR, DT, RF, SVM, NB, KNN, AdaBoost, XGBoost, and Extratrees and DL and proposed multi-stacking ML to predict PCOS. They used Explainable AI (XAI) techniques to make model predictions understandable, interpretable, and trustworthy. The result showed that multi-stacking ML recorded the best performance.

Logistic regression, K-Nearest Neighbor (KNN), Gaussian Naive Bayes, Random Forest Classifier, and Support Vector Machine (SVM) were used to identify polycystic ovary syndrome (PCOS) in [26].

A random forest classifier, decision tree and random forest chi-square algorithms were used in [27] to find the highest accuracy for PCOS detection keeping 20% of the dataset for testing and the remaining for training for every iteration.

3. Methodology

We applied different ML models: SVM, NB, LR, KNN, RF, DT, XGboost, and AdaBoost, with FS methods to predict PCOS.

We used the PCOS dataset from Kaggle [26], which includes 541 instances and 41 attributes. There are 178 instances of the positive class (1) and 363 instances of the negative class (3). For the experiments, Python programing language is used as a machine learning tool. For this, Anaconda distribution package, Scikit-learn library, Jupiter notebook, Spyder, Orange, etc. are used for the deployment of Python.

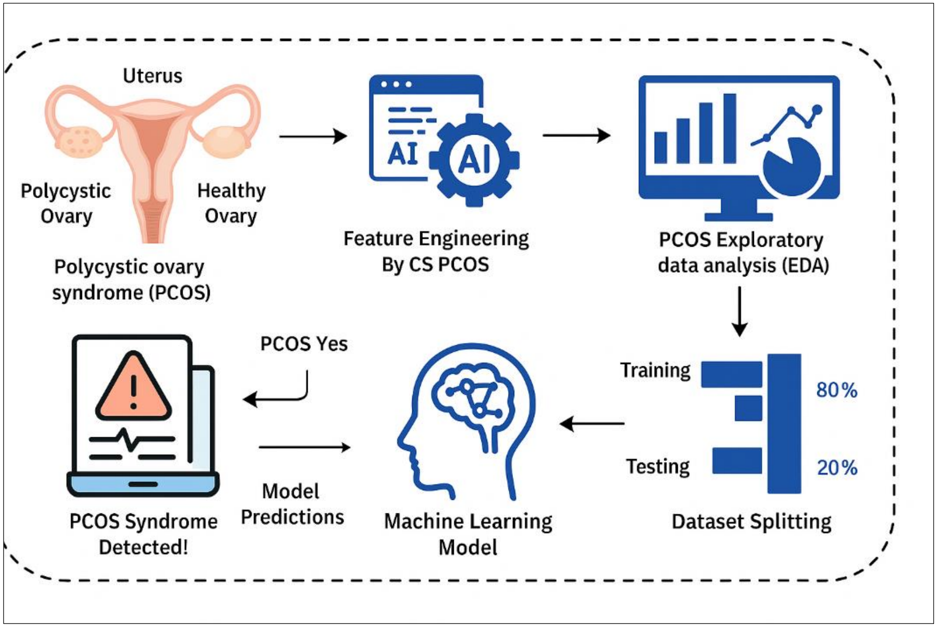


Figure 1 Work flow chart

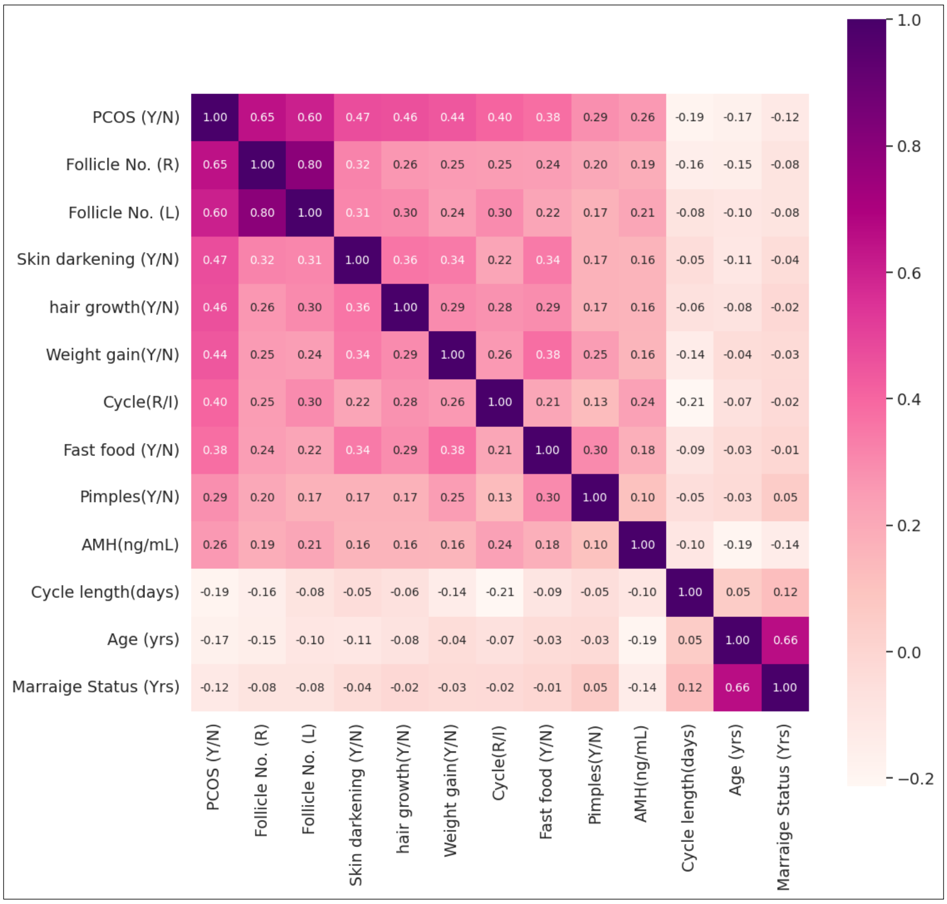


Figure 2 Features correlation

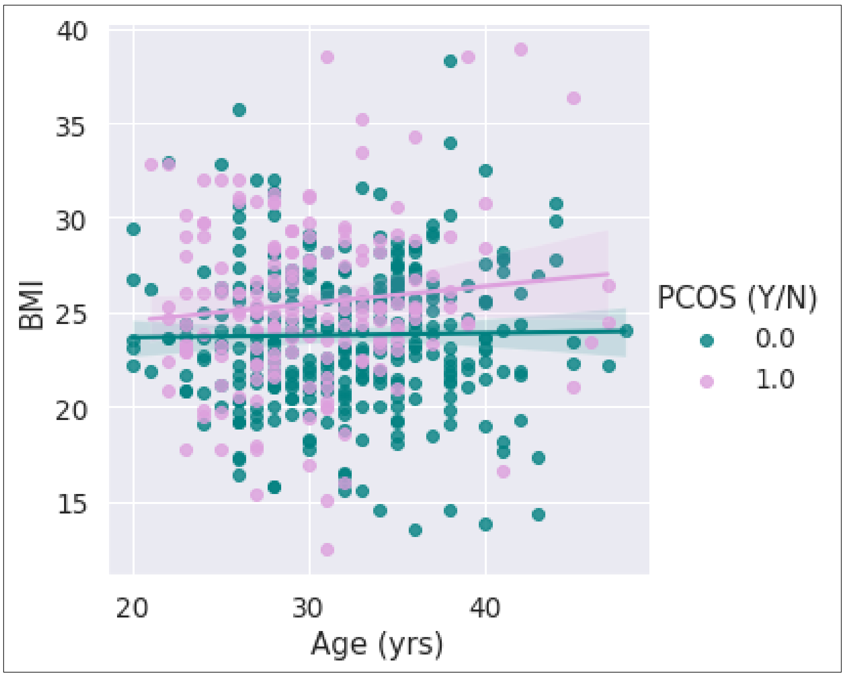


Figure 3 Pattern of weight gain (BMI) over years in PCOS and Normal.

Body mass index (BMI) in fig. 3 is showing consistency for normal cases. Whereas for PCOS the BMI increases with age.

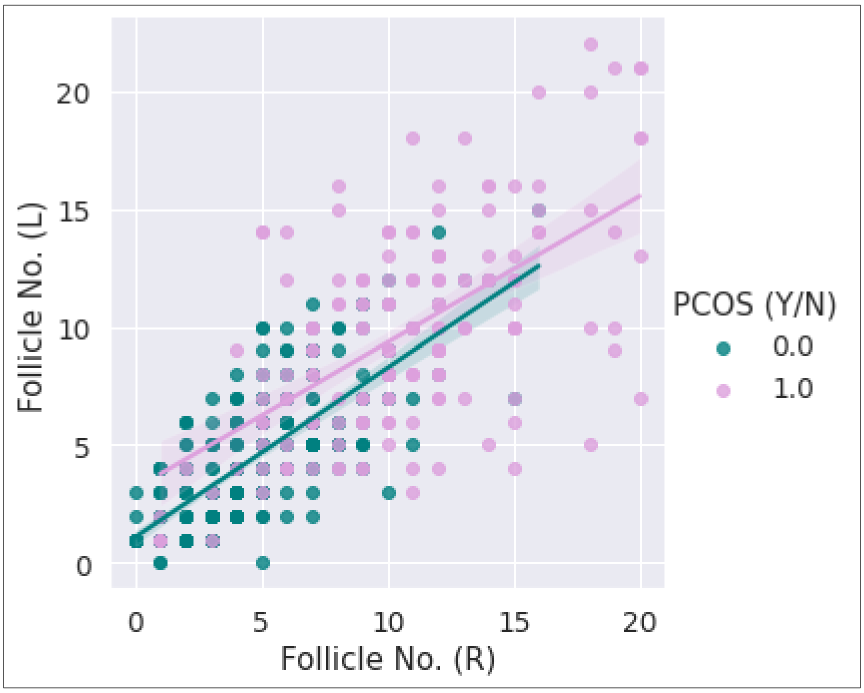


Figure 4 Distribution of follicles in both ovaries.

The distribution of follicles in fig. 4 in both ovaries Left and Right are not equal for women with PCOS in comparison with the "Normal" patient.

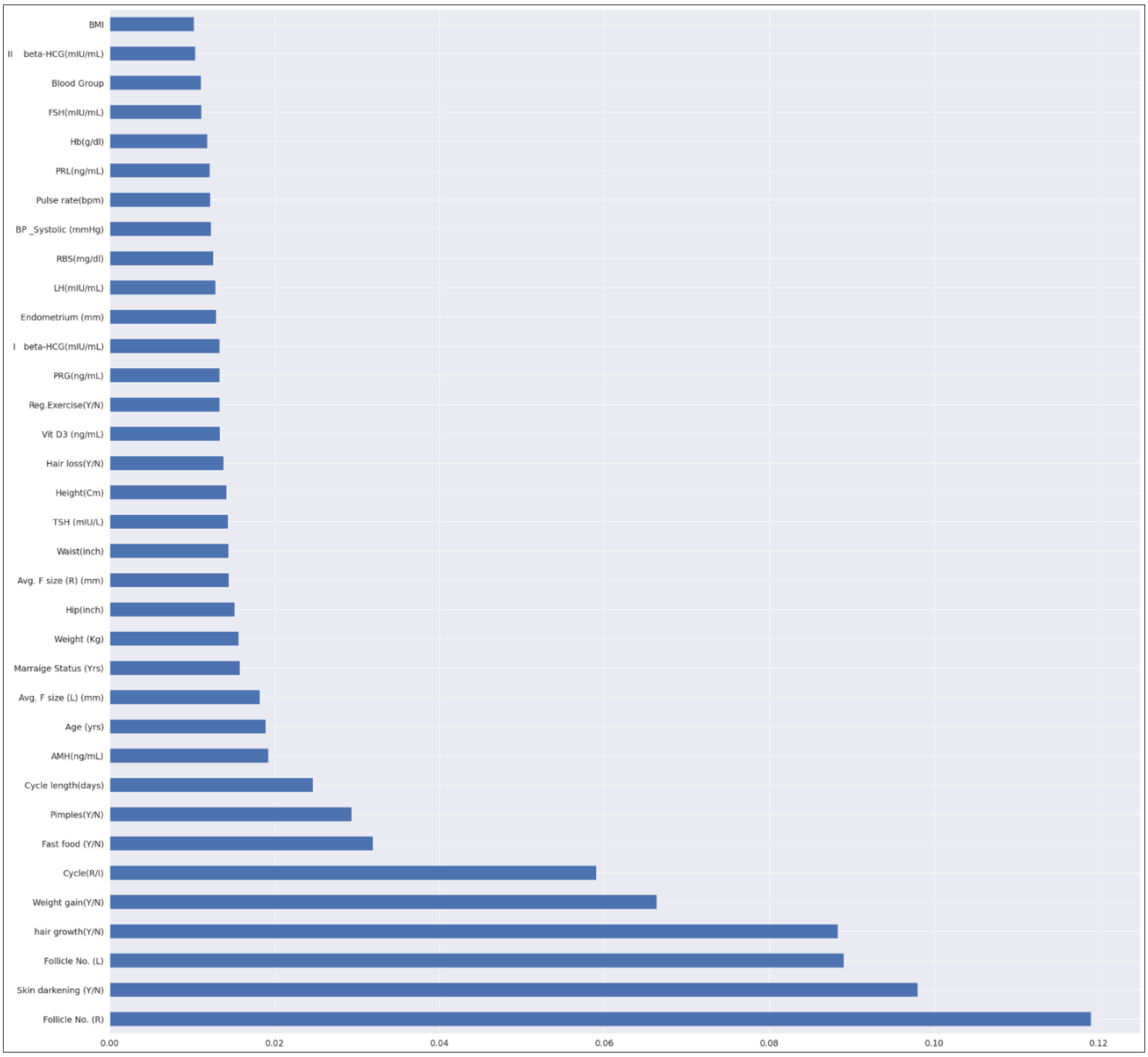
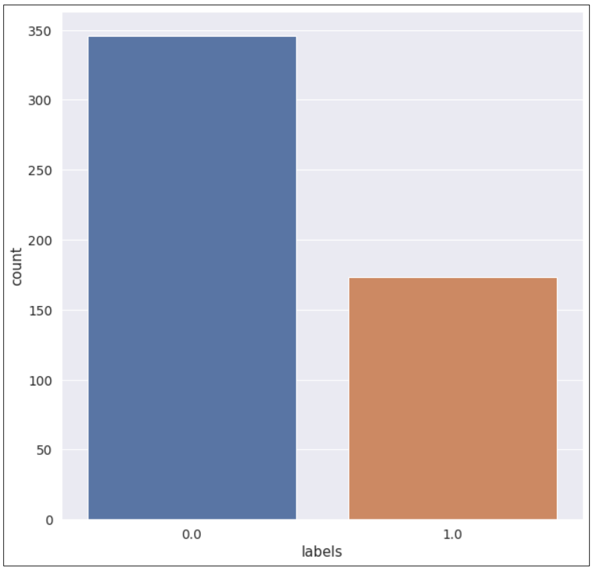
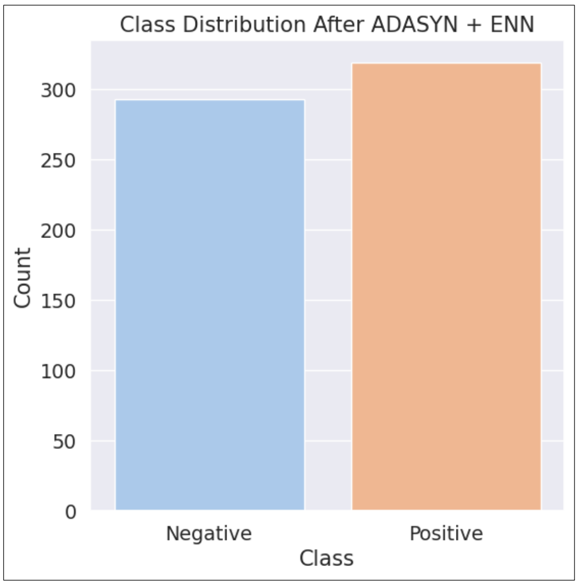


Figure 5 Feature Selection using Extra Trees Classifier



Class imbalance {0: 'Negative', 1: 'Positive'}



Class balancing {0: 'Negative': 298, 1: 'Positive': 328}

Figure 6 Class balancing

Table 1 SVM Performance

Model	**Class**	**precision**	**recall**	**F1 score**	**Accuracy**
Linear	0	0.94	0.88	0.91	0.90
	1	0.87	0.93	0.90	
RBF	0	0.84	0.94	0.89	0.87
	1	0.92	0.79	0.85	
Polynomial	0	0.82	0.96	0.88	0.86
	1	0.93	0.75	0.83	

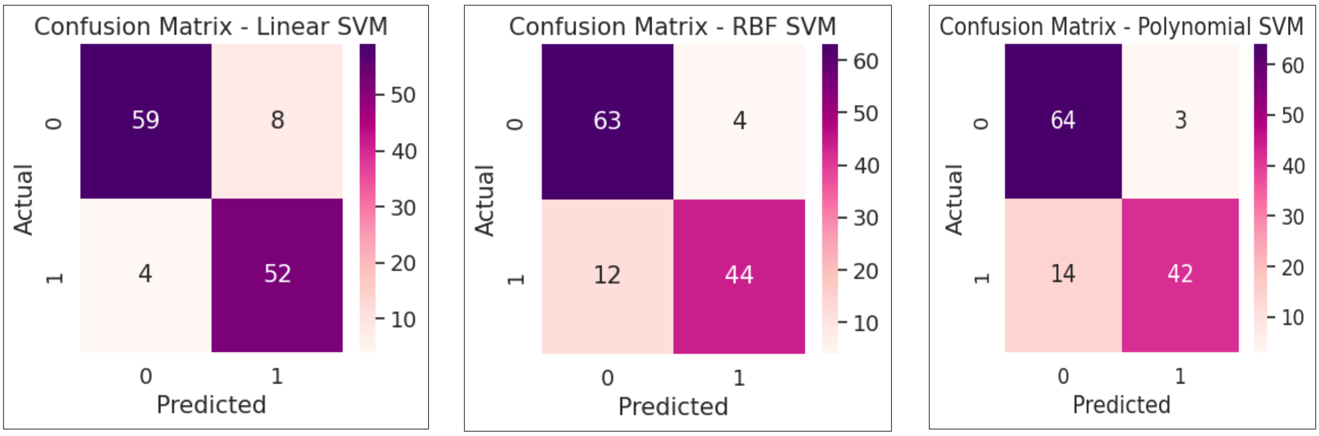


Figure 7 Confusion matrix

Figure 6 displays confusion matrices. True Positives and True Negatives are prominent in the Linear SVM matrix, confirming its strong predictive power for both PCOS-positive and negative classes. False Negatives are slightly higher in the RBF and Polynomial kernels, which might explain their lower F1 scores for the positive class.

4. Conclusion

The Linear SVM model outperforms both the RBF and Polynomial kernels, achieving the highest overall accuracy of 90%. For the Linear SVM, the F1 scores are 0.91 for class 0 (Negative) and 0.90 for class 1 (Positive), indicating a strong balance between precision and recall for both classes. Although the RBF and Polynomial SVMs shows good precision (particularly for the positive class), their recall values are lower, especially for class 1 (79% and 75% respectively), which affects the F1 scores.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Goodman, Neil F., et al. "American Association of Clinical Endocrinologists, American College of Endocrinology, and androgen excess and PCOS society disease state clinical review: guide to the best practices in the evaluation and treatment of polycystic ovary syndrome-part 1." Endocrine Practice 21.11 (2015): 1291-1300.

[2] Polycystic ovary syndrome (PCOS). Johns Hopkins Medicine. (n.d.). Retrieved December 25, 2021, from https://www.hopkinsmedicine.org/health/conditions-anddiseases/polycystic-ovary-syndrome-pcos.

[3] U.S. Department of Health and Human Services. (n.d.). What are the symptoms of PCOS? Eunice Kennedy Shriver National Institute of Child Health and Human Development. Retrieved December 25,2021, from https://www.nichd.nih.gov/health/topics/pcos/conditioninfo/symptoms

[4] Crespo, Raiane P., et al. "An update of genetic basis of PCOS pathogenesis." Archives of endocrinology and metabolism 62 (2018): 352-361.

[5] Frossing, Signe, et al. "Quantification of visceral adipose tissue in polycystic ovary syndrome: dual-energy X-ray absorptiometry versus magnetic resonance imaging." Acta Radiologica 59.1 (2018): 13-17.

[6] Bharati, S., Podder, P., Mondal, M.R.H., Surya Prasath, V.B., Gandhi, N. (2022). Ensemble Learning for Data-Driven Diagnosis of Polycystic Ovary Syndrome. In: Abraham, A., Gandhi, N., Hanne, T., Hong, TP., Nogueira Rios, T., Ding, W. (eds) Intelligent Systems Design and Applications. ISDA 2021. Lecture Notes in Networks and Systems, vol 418. Springer, Cham. https://doi.org/10.1007/978-3-030-96308-8_116

[7] Rotterdam ESHRE/ASRM-Sponsored PCOS Consensus Workshop Group. "Revised 2003 consensus on diagnostic criteria and long-term health risks related to polycystic ovary syndrome (PCOS)." Human reproduction 19.1 (2004): 41-47.

[8] Azziz, Ricardo, et al. "Criteria for defining polycystic ovary syndrome as a predominantly hyperandrogenic syndrome: an androgen excess society guideline." The Journal of Clinical Endocrinology & Metabolism 91.11 (2006): 4237-4245.

[9] Tian, Lifeng, et al. "Androgen receptor gene mutations in 258 Han Chinese patients with polycystic ovary syndrome." Experimental and Therapeutic Medicine 21.1 (2021): 1-1.

[10] Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective." Artificial Intelligence in medicine 23.1 (2001): 89-109.

[11] Bharati, Subrato, Prajoy Podder, and M. Rubaiyat Hossain Mondal. "Diagnosis of polycystic ovary syndrome using machine learning algorithms." 2020 IEEE Region 10 Symposium (TENSYMP). IEEE,2020.

[12] Amsy Denny et al., "I-HOPE: detection and prediction system for polycystic ovary syndrome (PCOS) using machine learning techniques", TENCON 2019-2019 IEEE Region 10 Conference (TENCON), 2019.

[13] Md Boktiar Hossain and Khandoker Hoque, “Machine Learning approaches in IDS”, International Journal of Science and Research Archive, 2022, 07(02), 706-715.

[14] Halbouni, A., Gunawan, T. S., Habaebi, M. H., Halbouni, M., Kartiwi, M., & Ahmad, R. (2022). Machine learning and deep learning approaches for cybersecurity: A review. IEEE Access, 10, 19572-19585.

[15] Yasmin Akter Bipasha, “Blockchain technology in supply chain management: transparency, security, and efficiency challenges”, International Journal of Science and Research Archive, 2023, 10(01), 1186-1196.

[16] Bharati S., Podder P., Mondal M.R.H. Diagnosis of polycystic ovary syndrome using machine learning algorithms; Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP); Dhaka, Bangladesh. 5–7 June 2020; pp. 1486–1489.

[17] Denny A., Raj A., Ashok A., Ram C.M., George R. i-hope: Detection and prediction system for polycystic ovary syndrome (pcos) using machine learning techniques; Proceedings of the TENCON 2019—2019 IEEE Region 10 Conference (TENCON); Kochi, India. 17–20 October 2019; pp. 673–678. [Google Scholar]

[18] Anda D., Iyamah E. Comparative Analysis of Artificial Intelligence in the Diagnosis of Polycystic Ovary Syndrome. [(accessed on 17 March 2023)]. Available online: https://www.researchgate.net/publication/366320486_Comparative_Analysis_of_Artificial_Intelligence_in_the_Diagnosis_of_Polycystic_Ovary_Syndrome.

[19] Bhardwaj P., Tiwari P. Proceedings of Academia-Industry Consortium for Data Science: AICDS 2020. Springer; New York, NY, USA: 2022. Manoeuvre of Machine Learning Algorithms in Healthcare Sector with Application to Polycystic Ovarian Syndrome Diagnosis; pp. 71–84.

[20] Adla Y.A.A., Raydan D.G., Charaf M.Z.J., Saad R.A., Nasreddine J., Diab M.O. Automated detection of polycystic ovary syndrome using machine learning techniques; Proceedings of the 2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME); Werdanyeh, Lebanon. 7–9 October 2021; pp. 208–212.

[21] Thakre V., Vedpathak S., Thakre K., Sonawani S. PCOcare: PCOS detection and prediction using machine learning algorithms. Biosci. Biotechnol. Res. Commun. 2020;13:240–244. doi: 10.21786/bbrc/13.14/56.

[22] Chauhan P., Patil P., Rane N., Raundale P., Kanakia H. Comparative analysis of machine learning algorithms for prediction of pcos; Proceedings of the 2021 International Conference on Communication information and Computing Technology (ICCICT); Mumbai, India. 25–27 June 2021; pp. 1–7.

[23] Rathod Y., Komare A., Ajgaonkar R., Chindarkar S., Nagare G., Punjabi N., Karpate Y. Predictive Analysis of Polycystic Ovarian Syndrome using CatBoost Algorithm; Proceedings of the 2022 IEEE Region 10 Symposium (TENSYMP); Mumbai, India. 1–3 July 2022; pp. 1–6.

[24] Aggarwal N., Shukla U., Saxena G.J., Kumar M., Bafila A.S., Singh S., Pundir A. Computational Intelligence: Select Proceedings of InCITe 2022. Springer; New York, NY, USA: 2023. An Improved Technique for Risk Prediction of Polycystic Ovary Syndrome (PCOS) Using Feature Selection and Machine Learning; pp. 597–606.

[25] Khanna V.V., Chadaga K., Sampathila N., Prabhu S., Bhandage V., Hegde G.K. A Distinctive Explainable Machine Learning Framework for Detection of Polycystic Ovary Syndrome. Appl. Syst. Innov. 2023;6:32. doi: 10.3390/asi6020032.

[26] F. J. Ananna, A. Khan, M. S. Ashraf, F. T. Zohora, M. T. Reza and M. M. Rahman, "Evaluating Machine Learning Model Performance in Predicting Polycystic Ovarian Syndrome," 2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), Thiruvananthapuram, India, 2023, pp. 339-344, doi: 10.1109/WIECON-ECE60392.2023.10456391.

[27] P. K B, B. V. Iyer, B. C, K. M. Thambanda and H. R. Kanasu, "Implementation of Various Machine Learning Algorithms to Predict Polycystic Ovary Syndrome," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-6, doi: 10.1109/INCET57972.2023.10170497.

[28] Amit Deb Nath, Rahmanul Hoque, Md. Masum Billah, Numair Bin Sharif, Mahmudul Hoque . Distributed Parallel and Cloud Computing: A Review. International Journal of Computer Applications. 186, 16 (Apr 2024), 25-32. DOI=10.5120/ijca2024923547

[29] Mobasher Hasan, Jubair Bin Sharif, Md. Kwosar, Md. Faysal Ahmed, Daniel Lucky Michael . Maximizing Business Performance through Artificial Intelligence. International Journal of Computer Applications. 186, 54 (Dec 2024), 9-15. DOI=10.5120/ijca2024924252

[30] Md Bahar Uddin, Md. Hossain and Suman Das, “Advancing manufacturing sustainability with industry 4.0 technologies”, International Journal of Science and Research Archive, 2022, 06(01), 358-366.

[31] Md Boktiar Hossain, Khandoker Hoque, Mohammad Atikur Rahman, Priya Podder, Deepak Gupta, “Hepatitis C Prediction Applying Different ML Classification Algorithm”, International Conference on Computing and Communication Networks 2024 (ICCCNet 2024) (Aceepted)

[32] Md Maniruzzaman, Md Shihab Uddin, Md Boktiar Hossain, Khandoker Hoque, "Understanding COVID-19 Through Tweets using Machine Learning: A Visualization of Trends and Conversations", European Journal of Advances in Engineering and Technology, 10(5), 108-114.

[33] Khamparia, A., Mondal, R. H., Podder, P., Bhushan, B., de Albuquerque, V. H. C., & Kumar, S. (Eds.). (2021). Computational intelligence for managing pandemics (Vol. 5). Walter de Gruyter GmbH & Co KG.

[34] Rahmanul Hoque, Masum Billah, Amit Debnath, S. M. Saokat Hossain and Numair Bin Sharif, "Heart Disease Prediction using SVM", International Journal of Science and Research Archive, 2024, 11(02), 412–420.

[35] Hoque, M., Hasan, M. R., Emon, M. I. S., Khalifa, F., & Rahman, M. M. (2024, September). Medical image interpretation with large multimodal models. In CEUR workshop proceedings.

[36] Emon, M., Hoque, M., Hasan, M., Khalifa, F., & Rahman, M. (2024, September). Fingerprint identification of generative models using a multiformer ensemble approach. In CEUR workshop proceedings. https://ceur-ws.org/Vol-3497/.

[37] Emon, M. I. S., Hoque, M., Hasan, M. R., Khalifa, F., & Rahman, M. (2025, April). A novel vision transformer-based approach to detect generative model fingerprint. In Medical Imaging 2025: Imaging Informatics (Vol. 13411, pp. 336-342). SPIE.