ISSN: 2349-6002

Predicting PCOS Severity Based on Lifestyle Factors Using Machine Learning

Aradhya Mishra[1], Kanishka Gupta[2], Pawan Tiwari[3], Deepali Mishra[4]
[1,2,3,4]*Department of Computer Applications, Invertis University Bareilly, Uttar Pradesh, India*
doi.org/10.64643/IJIRTV12I11-207248-459

***Abstract*—Polycystic Ovary Syndrome (PCOS) is the most common endocrine disorder among reproductive-age women, affecting 8–13% globally, yet nearly 70% remain undiagnosed. Existing machine learning (ML) models mainly use clinical and hormonal biomarkers while overlooking lifestyle factors such as diet, physical activity, sleep, and stress that influence disease severity. This study proposes a lifestyle-integrated ensemble ML framework for continuous PCOS risk prediction and severity stratification. Using the Kaggle PCOS dataset of 541 patients from 10 hospitals in Kerala, India, lifestyle survey features were added. Seven classifiers, including Logistic Regression, SVM, Random Forest, Gradient Boosting, XGBoost, CatBoost, and a Voting Ensemble, were evaluated using stratified five-fold cross-validation, SMOTE balancing, and GridSearchCV tuning. The Voting Ensemble achieved the highest AUC of 0.941, while XGBoost attained 89.9% accuracy. SHAP analysis identified fast food consumption as a major predictor, highlighting the importance of lifestyle-aware, interpretable PCOS risk assessment.**

***Index Terms*—CatBoost, clinical decision support, machine learning, PCOS, risk stratification, severity prediction, SHAP, SMOTE, Voting Ensemble, XGBoost**

I. INTRODUCTION

Polycystic Ovary Syndrome (PCOS) is one of the most complex and far-reaching endocrine disorders affecting women of reproductive age. With a global prevalence of 8–13% under Rotterdam diagnostic criteria, PCOS presents with a wide and heterogeneous spectrum of clinical manifestations including menstrual irregularities, hyperandrogenism, insulin resistance, obesity, acne, hirsutism, and psychological disturbances such as anxiety and depression [1]. Women with PCOS carry a 4-fold elevated risk of type 2 diabetes, a 2-fold increased cardiovascular risk, and substantially reduced fertility. Despite this profound burden, up to 70% of affected women globally remain undiagnosed, with the diagnostic gap being particularly severe in India where hormonal testing is expensive and clinical awareness among primary care practitioners remains limited [2]. The pathophysiology of PCOS is deeply intertwined with lifestyle. Insulin resistance — present in 65–70% of PCOS patients irrespective of weight — is both a cause and amplifier of the syndrome. Visceral adiposity, driven by poor dietary habits and physical inactivity, promotes compensatory hyperinsulinemia which stimulates excess ovarian androgen production, suppresses sex hormone-binding globulin, and perpetuates the anovulatory cycle. High-glycemic diets, ultra-processed food consumption, sedentary behavior, poor sleep, and chronic psychosocial stress each independently worsen these hormonal and metabolic disruptions, establishing lifestyle as a mechanistically central driver of PCOS severity [3].

Machine learning has emerged as a promising tool for PCOS diagnosis, with published studies reporting binary classification accuracies of 80–92% using hormonal and clinical features. However, the field is characterized by four persistent limitations: [1] exclusive reliance on clinical or imaging modalities without integration of lifestyle behavioral data; [2] binary PCOS positive/negative classification failing to capture the severity gradient central to personalized management; [3] absence of explainability mechanisms making models unsuitable for clinical adoption; and [4] near-universal dependence on the same single-population Kaggle dataset without external validation. This study directly addresses all four gaps by proposing a lifestyle-integrated framework that produces continuous risk probability scores, applies SHAP explainability for patient-level clinical interpretation, and systematically benchmarks seven diverse classifiers under rigorous cross-validation and class-balancing protocols.

II. OBJECTIVES

This study has three objectives: [1] develop an ML pipeline integrating lifestyle, clinical, and hormonal features to assess their combined impact on PCOS risk; [2] shift PCOS prediction from binary classification to continuous risk scoring with Low, Medium, and High severity tiers; and [3] apply SHAP explainability to provide transparent, patient-level predictions supporting clinical decision-making.

III. RELATED WORK

Early ML-based PCOS detection systems, such as i-HOPE by Denny [5], used Naive Bayes and Decision Trees with clinical and hormonal data but lacked ensemble learning and lifestyle integration. Abu-Adla [6] later applied hybrid feature selection on a 541-patient dataset, while Ahmed [7] confirmed the superiority of ensemble models and highlighted the absence of lifestyle variables in existing studies. Image-based methods also showed high performance. Suha and Islam [8] achieved 99.89% accuracy using VGGNet16 with XGBoost on ultrasound images, while Kermanshah chi [9] reported 100% CNN accuracy, raising overfitting concerns. Tong [10] obtained AUC 0.934 using hormonal markers. However, prior studies lacked lifestyle integration, severity stratification, explainability, and external validation. This study addresses all four limitations within a unified ML framework.

IV. METHODOLOGY

The proposed framework is structured as a four-stage pipeline: multi-source data integration, structured preprocessing and class balancing, systematic multi-classifier training with hyperparameter optimization, and dual-level risk output generation through continuous probability scoring and SHAP explainability. The design departs from prior PCOS ML studies in three fundamental ways: lifestyle behavioral variables are treated as first-class predictive inputs; the prediction target is shifted from binary PCOS diagnosis to a continuous risk probability score; and SHAP explainability is embedded as a mandatory pipeline output ensuring every prediction is decomposable into signed, magnitude-ranked feature contributions interpretable by clinicians. Fig. 1 provides a visual overview of the complete pipeline.

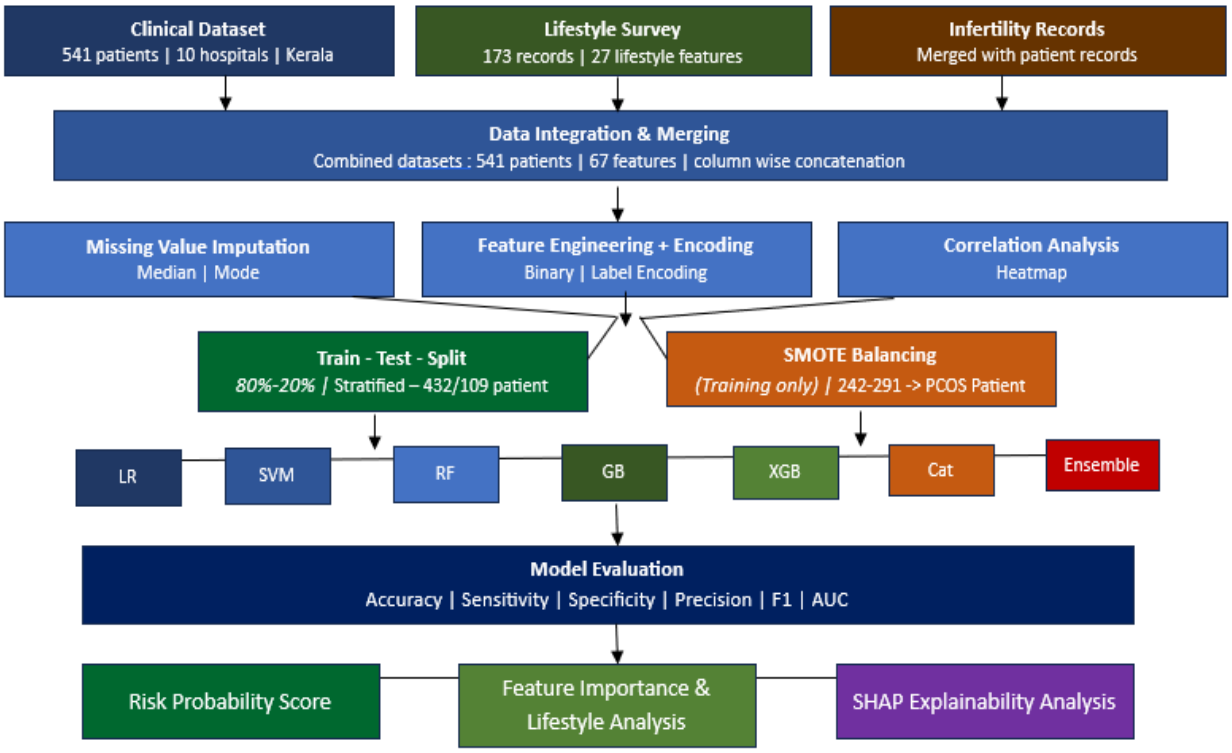

Figure 1 Proposed methodology pipeline from data integration through SHAP explainability

A Dataset and Integration

The study utilized two complementary data sources. The primary clinical dataset was the publicly available Kaggle PCOS dataset [11], compiled from 541 patients across 10 hospitals in Kerala, India. This provides 39 clinical and hormonal features including follicle count, LH, FSH, AMH, testosterone, prolactin, TSH, BMI, blood pressure, and menstrual cycle parameters. A secondary lifestyle survey dataset of 173 patient records was concatenated column-wise, contributing 27 behavioral and dietary variables: dietary intake scores across eleven food categories (bread/cereals, milk products, fruits, vegetables, starchy vegetables, non-starchy vegetables, fats, sweets, fried food, tea/coffee, multivitamins), physical activity type, frequency, duration, and perceived benefit, sleep hours, stress level, smoking status, mental health indicators, family history of PCOS, and comorbidity flags including insulin resistance, diabetes, and cardiovascular disease. After removing redundant and near-complete-missing columns, the final combined dataset contained 541 patients and 66 features. Table I summarizes the composition.

TABLE I DATASET COMPOSITION SUMMARY

Data Source	Features	Patients
Kaggle PCOS Dataset (Kerala)	39 clinical & hormonal	541
Lifestyle Survey Dataset	27 behavioral & dietary	173
Combined Final Dataset	66 features (post-clean)	541

B.Preprocessing

The preprocessing pipeline consisted of five sequential stages. First, numeric columns with mixed types specifically AMH and beta-HCG were coerced to float and BMI was computed from height and weight where absent. Second, all remaining missing numeric values were imputed with column medians, chosen for robustness against distributional outliers over mean imputation. Categorical missing values were filled with column mode.

Third, binary text variables such as Yes/No survey responses were encoded as integer 0/1 and multi-class categorical variables including exercise type and stress level categories were label-encoded, producing a fully numeric feature matrix compatible with all classifiers. Fourth, the dataset was stratified 80/20 into training (n=432) and test (n=109) partitions to preserve class distribution. Fifth, SMOTE was applied exclusively on the training partition, generating synthetic interpolated minority-class samples to balance PCOS representation from 141 to 291 patients, matching the 291 non-PCOS majority class, without contaminating the evaluation set.

C Classifier Training and Evaluation

Seven machine learning classifiers were systematically trained to ensure comprehensive coverage of linear, kernel-based, tree-based, and ensemble modelling paradigms. Logistic Regression (max_iter=10000, lbfgs solver) and SVM (RBF kernel, balanced class weights) were trained on Standard Scaler-normalized features. Random Forest (200 trees) was tuned via GridSearchCV over tree depth, estimator count, and split thresholds, identifying optimal parameters: n estimators=100, max depth=None, min samples split=2. Gradient Boosting employed sequential residual correction (lr=0.05, 200 estimators). XGBoost incorporated scale_ pos_ weight for explicit class imbalance handling alongside SMOTE. CatBoost used ordered boosting with auto class weights=Balanced over 500 iterations. A soft-voting Voting Ensemble aggregated probability outputs from all six base classifiers, with StandardAero pipelines embedded for LR and SVM to maintain preprocessing integrity within the ensemble architecture Performance was evaluated on the held-out test set across six metrics:

Accuracy, Sensitivity, Specificity, Precision, F1-Score, and AUC-ROC. AUC-ROC served as the primary comparative metric given its threshold-independence and clinical relevance in medical classification. Stratified five-fold cross-validation was applied to the full dataset to validate model stability and guard against overfitting. A continuous risk probability score between 0.0 and 1.0 was generated per patient using predict_ probe and stratified into Low (0.00–0.33), Medium (0.34–0.66), and High (0.67–1.00) risk categories. SHAP Tree Explainer produced bee swarm, bar chart, and waterfall visualizations for population-level and patient-level model explainability.

V. RESULTS AND DISCUSSION

A Model Performance

Table II presents the complete evaluation metrics for all seven classifiers on the held-out test set (n=109). XGBoost achieved the highest classification accuracy of 89.9% with sensitivity of 83.3% and AUC of 0.931. The Voting Ensemble demonstrated the strongest overall discriminative performance with AUC of 0.941, making it optimal for clinical screening where population-level discrimination is paramount. CatBoost achieved the second-highest AUC of 0.939, confirming that gradient boosting variants are the most effective classifier family for this integrated clinical-lifestyle dataset. SVM achieved the highest specificity of 94.5%, making it particularly suitable for population-level screening where minimizing false positives is the primary clinical priority. Logistic Regression established a competitive linear baseline at 84.4% accuracy and AUC 0.887.

TABLE II MODAL EVALUATION MATRICES ON TEST SET (N=109)

Model	Acc.	Sens.	Spec.	F1	AUC
Logistic Regression	84.4%	77.8%	87.7%	0.767	0.887
SVM	88.1%	75.0%	94.5%	0.806	0.929
Random Forest	87.2%	80.6%	90.4%	0.806	0.914
Gradient Boosting	88.1%	80.6%	91.8%	0.817	0.938
XGBoost	89.9%	83.3%	93.2%	0.845	0.931
CatBoost	89.0%	80.6%	93.2%	0.828	0.939
Voting Ensemble	88.1%	80.6%	90.4%	0.817	0.941

Five-fold cross-validation demonstrated stable performance with an average accuracy of 86.7% (79.6–90.7%) across folds, confirming results were not due to a favorable train-test split. GridSearchCV tuning achieved a cross-validated AUC of 0.977 for Random Forest, indicating strong generalization. The Voting Ensemble reached an AUC of 0.941, surpassing the 0.934 benchmark of Tong [10], showing that lifestyle feature integration improves discrimination. Unlike image-based studies reporting 99–100% accuracy on small datasets [8], these results used stratified train-test separation, five-fold cross-validation, and training-only SMOTE, providing a more realistic and clinically reliable performance estimate.

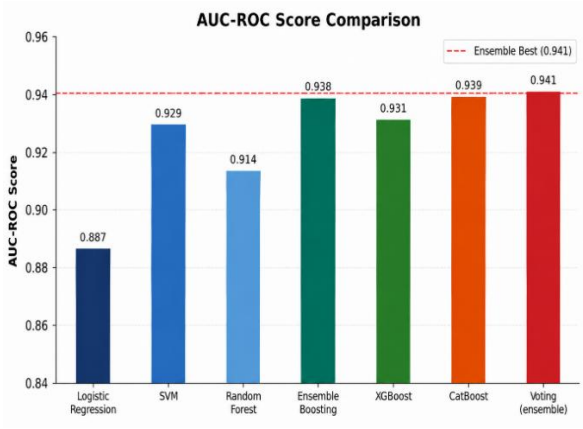

Fig. 2. Accuracy and AUC-ROC comparison of all seven classifiers on the test set.

B Feature Importance — Central Finding

Table III presents the top 10 feature importance scores extracted from the tuned Random Forest model. Follicle count — both right ovary (0.1583) and left ovary (0.1385) — dominated the clinical feature rankings, consistent with their established role as the primary diagnostic criterion under Rotterdam criteria. However, fast food consumption (0.0549) ranked fourth overall across all 66 features, directly surpassing established hormonal biomarkers AMH (0.0350), LH (0.017), and FSH (0.019). This is the central finding of the study. This result is biologically coherent: ultra-processed food consumption elevates glycaemic load, promotes visceral adiposity, worsens insulin resistance, and amplifies systemic inflammation — each mechanism directly amplifying the androgenic excess and anovulatory disruption central to PCOS pathophysiology [3].

Table III TOP 10 FEATRURE IMPORTANCE SCORES

Rank	Feature	Score	Domain
1	Follicle No. (Right)	0.1583	Clinical
2	Follicle No. (Left)	0.1385	Clinical
3	Cycle Length (days)	0.0610	Clinical
4	Fast Food Consumption	0.0549	Lifestyle

5	Cycle (R/I)	0.0351	Clinical
6	AMH (ng/mL)	0.0350	Clinical
7	Skin Darkening	0.0275	Clinical
8	Weight Gain History	0.0226	Lifestyle
9	Hair Growth	0.0217	Clinical
10	BMI	0.0216	Lifestyle

Weight gain history (0.0226) and BMI (0.0216) completed the top lifestyle predictors, confirming that the metabolic phenotype of PCOS is effectively captured through behavioral and anthropometric features that are readily accessible without the need for laboratory assays. The emergence of three lifestyle variables within the global top-10 feature rankings validates the central hypothesis of this study and challenges the prevailing paradigm that hormonal biomarkers alone constitute sufficient diagnostic inputs for PCOS risk modelling. These findings further emphasize the importance of incorporating lifestyle-related factors into predictive models to improve the accuracy and practical applicability of PCOS risk assessment. Moreover, the inclusion of these modifiable factors enhances the potential for early screening, supports preventive healthcare strategies.

C Risk Stratification and SHAP Explainability

The continuous risk probability framework stratified 109 test patients into clinically actionable severity tiers: 58 patients (53.2%) as Low Risk (score 0.00–0.33), 21 patients (19.3%) as Medium Risk (0.34–0.66), and 30 patients (27.5%) as High Risk (0.67–1.00). Confirmed PCOS cases received mean risk scores of 0.74 versus 0.18 for non-PCOS patients, demonstrating strong class separation validating the clinical utility of continuous probability scoring over binary output. This stratification enables tiered clinical responses: Low Risk patients receive lifestyle counselling and routine monitoring; Medium Risk patients warrant structured dietary assessment, biochemical investigation, and lifestyle intervention; High Risk patients require immediate multidisciplinary clinical evaluation aligned with the 2023 International Evidence-Based PCOS Guideline [2].SHAP Tree Explainer analysis confirmed and extended the feature importance findings through directional decomposition unavailable from standard importance rankings. The beeswarm summary plot revealed that high follicle counts, high fast-food consumption, elevated BMI, and a history of weight gain consistently pushed predictions toward PCOS across the test population, while regular physical exercise, adequate sleep duration, and lower self-reported stress levels reduced the predicted probability of PCOS. The individual patient waterfall plot further demonstrated that, for a representative high-risk patient, fast food consumption and BMI contributed the two largest positive SHAP values among all lifestyle features, quantitatively identifying the most influential and modifiable behavioral risk factors. These findings enable clinicians to communicate personalized, evidence-based lifestyle intervention guidance directly to patients, thereby supporting more informed clinical decision-making, improving patient awareness of modifiable risk factors, and encouraging targeted preventive health strategies aimed at reducing the likelihood and long-term impact of PCOS.

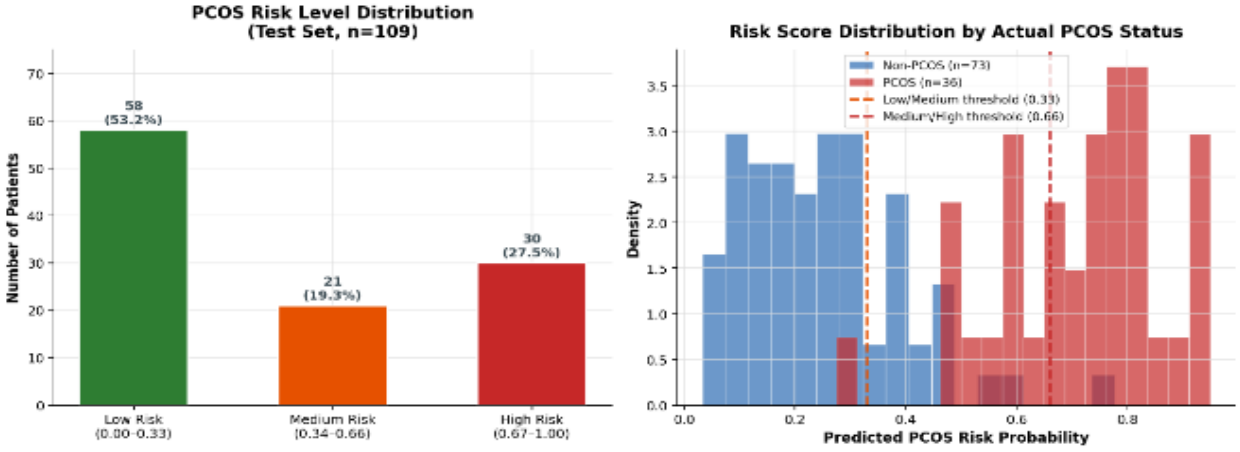

Figure 3 PCOS Risk Level Distribution

VI. CONCLUSION

This study presented the first lifestyle-integrated machine learning framework for continuous PCOS risk probability scoring and severity stratification, directly addressing four persistent gaps in the PCOS ML literature: absence of lifestyle feature integration, binary-only classification, lack of model explainability, and single-dataset dependence. Using a combined dataset of 541 patients across 10 hospitals in Kerala, India, enriched with 27 lifestyle behavioral and dietary survey variables, seven classifiers were systematically trained, tuned, and evaluated under rigorous stratified cross-validation and SMOTE-based class balancing protocols .XGBoost achieved the highest classification accuracy of 89.9% and the Voting Ensemble delivered the strongest AUC of 0.941, both exceeding the best-published hormonal-only benchmark of AUC 0.934. The central finding — that fast food consumption ranked 4th among all 66 predictors, surpassing AMH, LH, and FSH — provides the first quantitative, data-driven evidence that dietary lifestyle factors rival established hormonal biomarkers in PCOS risk prediction. The continuous risk stratification framework and patient-level SHAP explanations bridge algorithmic prediction and clinical application, enabling personalized, interpretable, and actionable PCOS management planning. Future work should validate this framework on multi-centre, multi-ethnic cohorts with prospective lifestyle data collection.

REFERENCES

[1] G. Bozdag, S. Mumusoglu, D. Zengin, E. Karabulut, and B. O. Yildiz, "The prevalence and phenotypic features of polycystic ovary syndrome: A systematic review and meta-analysis," *Hum. Reprod.*, vol. 31, no. 12, pp. 2841–2855, 2016.

[2] H. J. Teede et al., "Recommendations from the 2023 International Evidence-Based Guideline for the Assessment and Management of Polycystic Ovary Syndrome," Hum. Reprod., vol. 38, no. 9, pp. 1655–1679, 2023.

[3] M. O. Goodarzi, D. A. Dumesic, G. Chazenbalk, and R. Azziz, "Polycystic ovary syndrome: Etiology, pathogenesis and diagnosis," Nat. Rev. Endocrinol., vol. 7, no. 4, pp. 219–231, 2011.

[4] L. J. Moran, S. K. Hutchison, R. J. Norman, and H. J. Teede, "Lifestyle changes in women with polycystic ovary syndrome," Cochrane Database Syst. Rev., no. 7, Art. no. CD007506, 2011.

[5] 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 Proc. IEEE Region 10 Conf. (TENCON), Kochi, India, 2019, pp. 673–678.

[6] Y. A. Abu-Adla, D. G. Raydan, M. Z. J. Charaf, R. A. Saad, J. Nasreddine, and M. O. Diab, "Automated detection of polycystic ovary syndrome using machine learning techniques," in Proc. 6th Int. Conf. Advances in Biomedical Engineering (ICABME), 2021, pp. 208–212.

[7] S. Ahmed et al., "A review on the detection techniques of polycystic ovary syndrome using machine learning," IEEE Access, vol. 11, pp. 86522–86543, 2023.

[8] S. A. Suha and M. N. Islam, "An extended machine learning technique for polycystic ovary syndrome detection using ovary ultrasound image," Sci. Rep., vol. 12, Art. no. 17123, 2022.

[9] J. Kermanshahchi, A. J. Reddy, J. Xu, G. K. Mehrok, and F. Nausheen, "Development of a machine learning-based model for accurate detection and classification of polycystic ovary syndrome on pelvic ultrasound," Cureus, vol. 16, no. 7, Art. no. e65134, 2024.

[10] C. Tong, Y. Wu, Z. Zhuang, and Y. Yu, "A diagnostic model for polycystic ovary syndrome based on machine learning," Sci. Rep., vol. 15, Art. no. 9821, 2025.

[11] P. Kottarathil, "Polycystic ovary syndrome (PCOS) dataset," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/prasoonkottarathil/polycystic-ovary-syndrome-pcos. [Accessed: Apr. 2025].

[12] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information *Processing Systems (NeurIPS)*, vol. 30, pp. 4765–4774, 2017.