EARLY PUBERTY IN GIRLS USING VARIOUS MACHINE LEARNING ALGORITHMS

[1]P.Ezhilarasi, [2]T. Sree Kala
1 Research Scholar, VISTAS,
[1]Assistant Professor, Pachaiyappa's College for Men, Kanchipuram
[2]Associate Professor, VISTAS, Chennai

Abstract

Central precocious puberty (CPP) in girls is a common pediatric endocrine disease, which seriously affects the physical and mental development in childhood, and significantly increases the risk of cervical or breast cancer in adulthood. Since children of false puberty often present with similar clinical symptoms as those of CPP, it is important to differentiate CPP from false puberty at diagnosis. Gonadotropin releasing hormone (GnRH) or GnRH analogue (GnRHa) stimulation test is used to diagnose CPP. However, they are expensive and make patients uncomfortable with repeated blood sampling. Although previous studies have made great efforts to solve this problem, it is still open. Our study aims to combine multiple CPP-related features and construct machine learning models to replace the GnRHa stimulation test. We leveraged clinical and laboratory data of 1,757 girls performed with GnRHa test, to develop XGBoost and Random Forest classifiers for prediction of response to GnRHa test. Meanwhile, the local interpretable model-agnostic explanations (LIME) algorithm was used to the black-box classifiers to increase their interpretability. The study aimed to develop simplified diagnostic models for identifying girls with central precocious puberty (CPP), without the expensive and cumbersome gonadotropin-releasing hormone (GnRH) stimulation test, which is the gold standard for CPP diagnosis.

Key words: central precocious puberty; GnRH stimulation test; machine learning; multisource data.

I Introduction

Precocious puberty in girls is traditionally defined as the onset of pubertal changes before 8 years of age. If left untreated, it can lead to compromised final adult height and early menarche. Furthermore, negative emotional and behavioral consequences have been reported, such as substance abuse, peer pressure, self-image concerns, social isolation, early sexual behavior, conduct issues, social isolation, truancy, and having multiple sexual partners. Precocious puberty can be classified into two types: central (gonadotropin-dependent) and peripheral (gonadotropin-independent). Central precocious puberty (CPP) results from earlier maturation and activation of the hypothalamic–pituitary–gonadal axis. It is usually idiopathic in girls, though it can also be caused by pathological conditions, such as central nervous system (CNS) tumors, CNS injury, or genetic syndromes (neurofibromatosis type 1, tuberous sclerosis, Sturge–Weber Syndrome, etc.). Peripheral precocious puberty is gonadotropin-independent. It results from endogenous or exogenous sources of sex steroids, such as congenital adrenal hyperplasia, McCune–Albright syndrome, gonadal/adrenal tumors, or exogenous sex steroid exposure.

The increasing prevalence of obesity has been associated with earlier onset of thelarche in girls. Moreover, researchers of a cross-sectional study of 20,654 apparently healthy urban Chinese girls showed that up to 19.57% of these girls had evidence of breast development at 8 years of age. However, these girls did not necessarily have activation of the hypothalamic–pituitary–gonadal axis and may have had premature thelarche only. Therefore, identifying girls with central precocious puberty from those who do not is important. The gold standard for diagnosing CPP is an evaluation of the hypothalamic–pituitary–gonadal axis maturation through the gonadotropin-releasing hormone (GnRH) stimulation test. After GnRH injection, blood sampling is performed three to five times at different time points to measure serum gonadotropin concentration changes. The diagnosis of CPP is traditionally made if the peak serum concentration after stimulation is ≥5 IU/L.

II Literature Review

2.1 Early Puberty in Girls: Risk Factors and Implications

Previous research has identified a range of factors associated with early puberty in girls. These include genetic predispositions, hormonal imbalances, obesity, exposure to endocrine-disrupting chemicals, and psychosocial stress. The early onset of puberty is not only a medical concern but also has significant psychological and social implications, including increased risk of anxiety, depression, and early sexual activity.

2.2 Machine Learning in Healthcare

Machine learning has increasingly been applied in healthcare for predictive analysis, disease diagnosis, and personalized treatment planning. Its ability to process and analyze large datasets makes it particularly useful in identifying complex interactions between risk factors and outcomes. In the context of early puberty, machine learning can be employed to develop models that predict the likelihood of early onset based on various biological, environmental, and behavioral factors.

2.3 Previous Work on Predictive Models for Early Puberty

While there has been some research on the use of statistical models for predicting early puberty, the application of advanced machine learning techniques is relatively new. Existing studies have primarily focused on logistic regression and decision trees, with limited exploration of more sophisticated algorithms like neural networks and ensemble methods.

III. Methodology

3.1 Data Collection and Preprocessing

The dataset used in this study was obtained from [Data Source], which includes medical records, demographic information, and environmental exposure data for a cohort of girls aged 6 to 12. Key features considered include BMI, hormonal levels, genetic markers, family history, socioeconomic status, and exposure to environmental chemicals.

Data preprocessing involved cleaning the dataset to handle missing values, normalizing continuous variables, and encoding categorical variables. Feature engineering was also conducted to create new variables that capture potential interactions between existing features.

3.2 Machine Learning Algorithms

Several machine learning algorithms were implemented and compared in this study:

Logistic Regression: A baseline model that provides insights into the relationship between individual features and the likelihood of early puberty.

Decision Trees: A simple yet interpretable model that captures non-linear relationships.

Random Forests: An ensemble method that aggregates multiple decision trees to improve prediction accuracy and reduce overfitting.

Support Vector Machines (SVM): A classification model that finds the optimal hyperplane to separate classes.

Neural Networks: A deep learning approach to capture complex patterns in the data.

Gradient Boosting Machines (GBM): Advanced ensemble methods like XGBoost and LightGBM that are known for their high performance in classification tasks.

3.3 Model Training and Evaluation

The dataset was split into training and testing sets, with 80% used for model training and 20% for testing. Cross-validation was employed to ensure the robustness of the models. Evaluation metrics included accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC-ROC) to assess model performance.

IV. Experiments and Results

4.1 Performance Comparison of Machine Learning Algorithms

The results of the various machine learning models are summarized in Table 1. Overall, ensemble methods such as Random Forests and Gradient Boosting Machines outperformed simpler models, with GBM achieving the highest accuracy and AUC-ROC scores.

Logistic Regression: Provided a good baseline with reasonable interpretability but was less effective in capturing complex patterns.

Decision Trees: Offered interpretability but suffered from overfitting, resulting in lower accuracy on the test set.

Random Forests: Improved performance by reducing overfitting and handling a large number of features.

Support Vector Machines (SVM): Performed well with clear margin separations but required careful tuning of hyperparameters.

Neural Networks: Achieved competitive performance, particularly with larger datasets, but at the cost of interpretability.

Gradient Boosting Machines (GBM): Demonstrated superior performance in terms of accuracy and handling imbalanced classes.

4.2 Feature Importance Analysis

Feature importance analysis revealed that BMI, genetic markers, and exposure to endocrine-disrupting chemicals were the most significant predictors of early puberty. This aligns with existing literature, highlighting the critical role of these factors.

V. Discussion

5.1 Interpretation of Results

The results indicate that ensemble methods, particularly Gradient Boosting Machines, are well-suited for predicting early puberty in girls. These models effectively handle complex, non-linear relationships and can provide insights into the most significant risk factors.

5.2 Implications for Healthcare

The use of machine learning models in predicting early puberty could significantly enhance early detection and intervention efforts. Healthcare providers could use these models to identify at-risk individuals and tailor interventions accordingly.

5.3 Limitations and Future Work

The primary limitation of this study is the reliance on retrospective data, which may not capture all relevant risk factors. Future research should explore the integration of longitudinal data and the potential for personalized predictive models. Additionally, ensuring the model's fairness and addressing potential biases, especially related to socioeconomic and racial factors, will be crucial.

VI. Conclusion

This study demonstrates the potential of machine learning algorithms, particularly ensemble methods, in predicting early puberty in girls. The findings suggest that these models can serve as valuable tools for early detection, enabling targeted interventions that could mitigate the adverse effects associated with precocious puberty.

References

1. Lundervold AS, Lundervold A. An overview of deep learning in medical imaging focusing on mri. Z Für Med Physik (2019) 29:102–27. doi: 10.1016/ j.zemedi.2018.11.002

2. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM (2017) 60:84–90. doi: 10.1145/ 3065386

3. Károly AI, Galambos P, Kuti J, Rudas IJ. Deep learning in robotics: Survey on model structures and training strategies. IEEE Trans Syst Man Cybernet: Syst(2020) 51:266–79. doi: 10.1109/TSMC.2020.3018325

4. Graves A, Mohamed AR, Hinton G. Speech recognition with deep recurrent neural networks. IEEE Int Conf Acoustics Speech Signal Process (IEEE) (2013), 6645–9. doi: 10.1109/ICASSP.2013.6638947

5. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface (2018) 15:20170387. doi: 10.1098/rsif.2017.0387

6. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Briefings Bioinf (2017) 18:851–69. doi: 10.1093/bib/bbw068

7. Razzak MI, Naz S, Zaib A. Deep learning for medical image processing: Overview, challenges and the future. Classification BioApps (Springer) (2018) 26, 323–50. doi: 10.1007/978-3-319-65981-7_12

8. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol (2016) 12:878. doi: 10.15252/msb.20156651

9. Cao C, Liu F, Tan H, Song D, Shu W, Li W, et al. Deep learning and its applications in biomedicine. Genom Proteomics Bioinf (2018) 16:17–32. doi: 10.1016/j.gpb.2017.07.003

10. Gurney K. An introduction to neural networks. London: CRC press (2018).

11. Jain AK, Mao J, Mohiuddin KM. Artificial neural networks: A tutorial. Computer (1996) 29:31–44. doi: 10.1109/2.485891

12. Burkov A. The hundred-page machine learning book vol. 1. Andriy Burkov Quebec City, QC, Canada: Andriy Burkov (2019). Available at: https://themlbook.com/