Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230207

Towards an Explainable AI-Based Tool to Predict Preterm Birth

Ilias KYPARISSIDIS KOKKINIDIS[a], Evangelos LOGARAS[a], Emmanouil S. RIGAS[a,1], Ioannis TSAKIRIDIS[b], Themistoklis DAGKLIS[b], Antonis BILLIS[a,2] and Panagiotis D. BAMIDIS[a,2]

[a] *Lab of Medical Physics and Digital Innovation, School of Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki, Greece*

[b] *3rd Department of Obstetrics and Gynecology, Aristotle University of Thessaloniki, Greece*

ORCiD ID: Ilias Kyparissidis Kokkinidis https://orcid.org/0000-0002-3307-4351
Evangelos Logaras https://orcid.org/0000-0003-4921-3819
Emmanouil S. Rigas https://orcid.org/0000-0002-8042-9135
Ioannis Tsakiridis https://orcid.org/0000-0003-4337-7871
Themistoklis Dagklis https://orcid.org/0000-0002-2863-5839
Antonis Billis https://orcid.org/0000-0002-1854-7560
Panagiotis D. Bamidis https://orcid.org/0000-0002-9936-5805

Abstract. Preterm birth (PTB) is defined as delivery occurring before 37 weeks of gestation. In this paper, Artificial Intelligence (AI)-based predictive models are adapted to accurately estimate the probability of PTB. In doing so, pregnant women' objective results and variables extracted from the screening procedure in combination with demographics, medical history, social history, and other medical data are used. A dataset consisting of 375 pregnant women is used and a number of alternative Machine Learning (ML) algorithms are applied to predict PTB. The ensemble voting model produced the best results across all performance metrics with an area under the curve (ROC-AUC) of approximately 0.84 and a precision–recall curve (PR-AUC) of approximately 0.73. An attempt to provide clinicians with an explanation of the prediction is performed to increase trustworthiness.

Keywords. Artificial Intelligence (AI)**,** Machine Learning (ML), Explainable Artificial Intelligence (XAI), Decision Support Systems, Clinical, Premature Birth, Obstetrics and Gynecology (specialty)

1. Introduction

Preterm birth (PTB) is defined as delivery before 37 weeks of pregnancy. One third of PTBs are medically indicated, primarily preeclampsia and/or fetal growth restriction (FGR), and the other two thirds are spontaneous [1]. According to data from 107 countries in 2014, 14.84 million births—or 10.6% of all births—were PTB [2].

PTB increases the risk of both short- and long-term health effects to the neonate and its later life. Regarding short-term effects, neonatal and childhood mortality are both mostly attributed to PTB [3]. Long-term risks for hypertension and other cardiovascular

[1] Corresponding Author: Emmanouil S. Rigas, Medical Physics and Digital Innovation Lab, School of Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki, Greece. E-mail: erigas@auth.gr.

[2] A. Billis & PD Bamidis share senior authorship

diseases, type 2 diabetes, kidney disease, respiratory difficulties, and developmental disabilities are elevated [4]. PTB is also linked to higher healthcare expenses [5].

Based on the above, novel approaches to reduce the risk of PTB is one of the main goals in obstetric care. The development of new technologically advanced digital tools and interventions can contribute to achieving this goal. The QUiPP application is one such example [6]. Taking into account the pregnant woman's medical history, current pregnancy information, and predictive clinical tests, the QUiPP application is able to predict her likelihood of giving PTB within clinically significant timeframes. In another study [7], authors proposed the PredictPTB model, a deep learning model, similarly able to predict PTB using routinely collected data from electronic health records (EHRs). Moreover, an intelligent mechanism based on the SVM algorithm has been introduced [8], capable of predicting PTB among others.

In this paper, the aim is to predict PTB in an interpretable way using ML algorithms, find the best-performing ones for PTB, then combine them with ensemble methods and a Multilayer Perceptron (MLP) neural network to increase prediction performance. Furthermore, an emphasis is being placed on the interpretability of the models. To provide explainability, Shapley Additive Explanations (SHAP) [9] were used. This is done to increase the trust of the models from clinicians, who are unlikely to trust the diagnostic recommendations of a black box system [10].

2. Methods

A dataset of 375 pregnancies (of which 128 were identified as preterm) was used in this study. The dataset contained 32 features, including demographics, social and medical history, and obstetrics variables. The information was collected pseudonymously by Hippokration General Hospital in Thessaloniki, as part of an ongoing prospective cohort study that was approved by AUTH’s Research Ethics and Conduct Committee (94521/2022), in the context of the HosmartAI project.[2] The dataset was collected by four medical professionals over a four-month period (May-August 2022). The dataset underwent preprocessing, including categorizing input features as numerical, ordinal, or nominal, and using one-hot encoding and a label encoder to map certain features into binary vectors or integer values. Some features with high levels of missing values were dropped, and others were imputed with the most frequent or median value. The dataset was also balanced using random under-sampling and Synthetic Minority Oversampling Technique (SMOTE). The final training set used, after the preprocessing and sampling methods were applied contained 120 entries identified as pre-term and 150 entries that were not, whereas in the test set 30 cases were identified as preterm and 45 as not.

To provide the prediction, the following models were tested: Logistic Regression, SVM, Random Forest, XGBoost, and an MLP perceptron from the scikit learn implementation. Various combinations of Voting and Stacking ensembles [11] were also tested to increase performance. Hyperparameters were determined via Bayesian optimization and error analysis. Respectively, regarding the interpretations, SHAP explanations were used. SHAP was chosen because model independence is a feature of SHAP values, which can be used to explain both generally for each model and locally for each prediction in a consistent way in all models.

[2] https://www.hosmartai.eu/

3. Results

3.1. Model evaluation

Six standard evaluation metrics are used to assess the predictions of all the classifiers. These include ROC-AUC, PR-AUC, Recall, Precision, Accuracy, and F1 score. ROC-AUC reflects the best balance between Sensitivity and Specificity whereas PR-AUC the balance between precision and recall. The developed models were able to predict PTB using the predefined variables from the dataset with accuracies of ~69% (SVM), ~57,3% (Logistic), ~73% (XGBoost), ~70% (Random Forest), 78% (Stacking) and ~81% (Voting) (Table 1). This indicates fair discriminative ability in Random Forest and XGBoost, whereas SVM and Logistic Regression were discarded because of low performance, especially at the Recall (SVM ~0.47, Logistic Regression ~0.5) measure. On all models, 5-fold cross-validation was performed to avoid overfitting.

Table 1. Performance metrics of the ML-based prediction models

ML Algorithm	AUC	Accuracy	Recall	Precision	F1 Score
Random Forest	0.83	0.7	0.47	0.78	0.59
XGBoost	0.83	0.73	0.82	0.65	0.65
Voting ensemble	**0.84**	**0.81**	**0.68**	**0.73**	**0.70**
Stacking	0.83	0.78	0.70	0.68	0.69

A graphic representation of the ROC-AUC and PR-AUC of all models can be seen in Figure 1.

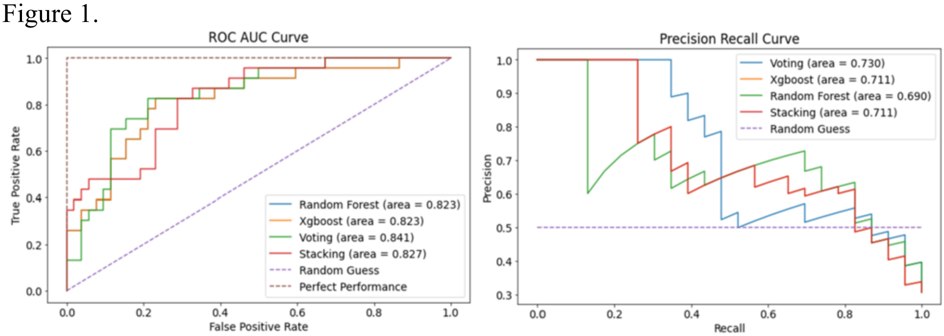

Figure 1. ROC-AUC and PR-AUC for all Classifiers Multi-part figure

In Figure 2 the Confusion Matrix regarding the Voting ensemble Classification algorithm for the preterm labor prediction is displayed. The percentage of the true negative predicted cases is ~57%, and the percentage of the true positive cases is ~22%.

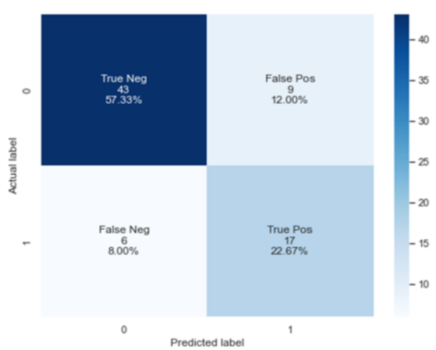

Figure 2. Confusion Matrix for the Voting Classifier

3.2. Feature ranking and Local explanations

In Figure 3(a) we can see the global feature ranking for the predictions of the model. After considering the global explanation for the predicted outcome of PTB in pregnancies, it is also essential to comprehend the output of the models for each specific case. Figure 3(b) illustrates an example of the SHAP explanation for one instance randomly selected from the preterm dataset. This output instance was predicted and confirmed as preterm. Here Placental cord insertion, Pappa, and Gravida play the most important role in the model's output for this instance. It is worth noting that local and global explanations can differ as presented in this instance.

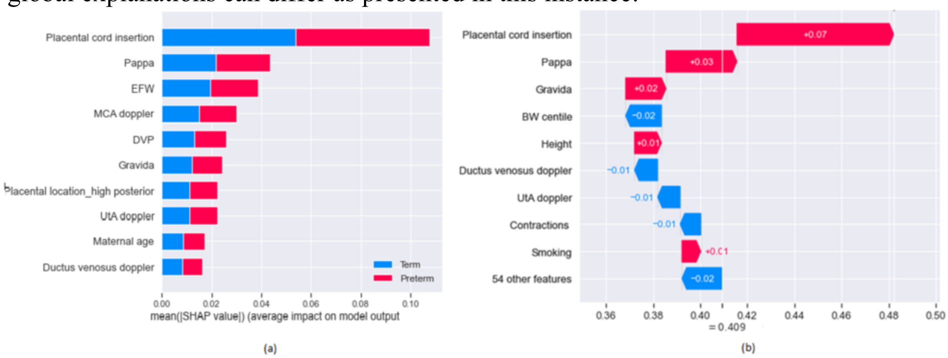

Figure 3. (a) [Global] depicts the influence of features on the constructed classifiers. It illustrates both negative and positive impacts on predictions (red and blue bars). (b)[Local] depicts feature influence on one randomly selected prediction.

4. Discussion and Future work

This study aims to improve machine learning algorithm predictions by providing local explanations that are easy for obstetricians to understand. This can help them make better decisions and provide feedback to improve the model's accuracy in future iterations. The research can also aid in efficient pregnancy screening by giving doctors options and allowing them to understand how each model reaches its decisions.

Given the relatively small dataset that was available for this study, we argue that the initial results that were presented are promising toward the goal of predicting PTB with high accuracy. We observe that currently the Stacking and the Voting ensemble seem to perform better. This is not surprising since ensemble models have been shown to outperform others in diagnostic tasks [12,13]. Moreover, ensemble techniques have been shown to deal better with imbalanced datasets like ours [14]. Compared to previously presented methods for the prediction of PTB, the results in terms of ROC-AUC have been similar. However, we notice an increase of the PR-AUC to ~0.72 compared to other applications [7,8] with the ensemble methods in our models.

Future work will optimize models on a larger dataset and apply 10-fold cross-validation instead of the 5-fold that was employed here to all models.

Acknowledgement

This project has been partially supported by the HosmartAI project – "Hospital Smart development based on AI", which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016834

References

[1] Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. Lancet. 2008;371(9606):75–84.

[2] Chawanpaiboon S, et. al., regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. Lancet Glob Health. 2019 Jan;7(1):e37-e46. doi: 10.1016/S2214-109X(18)30451-0. Epub 2018 Oct 30. PMID: 30389451; PMCID: PMC6293055.

[3] Liu L, Johnson HL, Cousens S, Perin J, Scott S, Lawn JE, Rudan I, Campbell H, Cibulskis R, Li M, Mathers C. Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. The lancet. 2012 Jun 9;379(9832):2151-61. doi: 10.1016/S0140-6736(12)60560-1.

[4] Raju TN, Pemberton VL, Saigal S, Blaisdell CJ, Moxey-Mims M, Buist S. Long-term healthcare outcomes of preterm birth: an executive summary of a conference sponsored by the National Institutes of Health. The Journal of pediatrics. 2017 Feb 1;181:309-18. doi: 10.1016/j.jpeds.2016.10.015.

[5] Jacob J, Lehne M, Mischker A, Klinger N, Zickermann C, Walker J. Cost effects of preterm birth: a comparison of health care costs associated with early preterm, late preterm, and full-term birth in the first 3 years after birth. The European Journal of Health Economics. 2017 Nov;18(8):1041-6. doi: 10.1007/s10198-016-0850-x.

[6] Dehaene I, Steen J, Vandewiele G, Roelens K, Decruyenaere J. The web-based application "QUiPP v. 2" for the prediction of preterm birth in symptomatic women is not yet ready for worldwide clinical use: ten reflections on development, validation and use. Archives of Gynecology and Obstetrics. 2022 Aug;306(2):571-5.

[7] AlSaad R, Malluhi Q, Boughorbel S. PredictPTB: an interpretable preterm birth prediction model using attention-based recurrent neural networks. BioData Mining. 2022 Feb 14;15(1):6.

[8] Moreira MW, Rodrigues JJ, Marcondes GA, Neto AJ, Kumar N, Diez ID. A preterm birth risk prediction system for mobile health applications based on the support vector machine algorithm. In2018 IEEE International Conference on Communications (ICC) 2018 May 20 (pp. 1-5). IEEE. doi: 10.1109/ICC.2018.8422616.

[9] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30. Retrieved from https://arxiv.org/abs/1705.07874

[10] Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Computer Methods and Programs in Biomedicine. 2022 Sep 27:107161.

[11] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011 Nov 1;12:2825-30.

[12] Sahin H, Subasi A. Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques. Applied Soft Computing. 2015 Aug 1;33:231-8.

[13] Bashir S, Qamar U, Khan FH. BagMOOV: A novel ensemble for heart disease prediction bootstrap aggregation with multi-objective optimized voting. Australasian physical & engineering sciences in medicine. 2015 Jun;38:305-23.

[14] Firdous N, Bhardwaj S. Handling of derived imbalanced dataset using XGBoost for identification of pulmonary embolism—a non-cardiac cause of cardiac arrest. Medical & Biological Engineering & Computing. 2022 Feb 1:1-8. https://doi.org/10.1007/s11517-021-02455-2