Thrombophilia Prediction Using Machine Learning Algorithms

Aldina R. Avdić[1], Natasa Z. Djordjević[2], Ulfeta A. Marovac[1*], Lejlija M. Memić[1], Zana Ć. Dolićanin[3], Goran M. Babić[4]

[1] State University of Novi Pazar, Department of Technical and Technological Sciences, Vuka Karadžića 9, 36300 Novi Pazar, Serbia, e-mail: apljaskovic@np.ac.rs, umarovac@np.ac.rs, lmemic@np.ac.rs
[2] State University of Novi Pazar, Department of Natural and Mathematical Sciences, Vuka Karadžića 9, 36300 Novi Pazar, Serbia, e-mail: natasadj@np.ac.rs
[3] State University of Novi Pazar, Department of Biomedical Sciences, Vuka Karadžića 9, 36300 Novi Pazar, Serbia, e-mail: zdolicanin@np.ac.rs
[4] University of Kragujevac, Faculty of Medical Sciences, Svetozara Markovića 69, 34000 Kragujevac, Serbia, e-mail: ginbabic@medf.kg.ac.rs

* *Corresponding author*

DOI: 10.46793/ICCBI23.140A

Abstract: Thrombophilia in pregnancy is the result of a complex interaction of inherited and acquired factors, which increase blood coagulation and consequently placental ischemic conditions. Early identification of risk of developing thrombophilia in pregnancy is crucial for implementing preventive measures and personalized therapy. In this study, we propose a novel approach for prediction of thrombophilia in pregnancy utilizing machine learning (ML) algorithms with a particular focus on neural networks. The research is done using a dataset consisting of demographic, lifestyle, and clinical information from a 35 pregnant woman (22 healthy and 13 with thrombophilia). These features are used to train and evaluate different ML models with neural networks and decision trees. The evaluation of the proposed approach involves cross-validation and performance metrics assessment. The results highlight the effectiveness of decision trees and neural networks in accurately predicting thrombophilia in pregnancy risk.

Keywords: neural networks, decision trees, machine learning, thrombophilia in pregnancy, prediction

1. Introduction

Thrombophilia is inherited or acquired disorder of hemostasis and represents condition of increased hypercoagulability that predisposes patients to thromboembolic events. In normal pregnancy, hormonal changes induce hypercoagulability state, which is adaptive mechanism to reduce the risk of hemorrhage during and after the delivery. The risk of thromboembolic events in pregnancy and postpartum period is 5 times greater compared to non-pregnant women. Thrombophilia in pregnancy is the result of a

complex interaction of inherited and acquired factors, which increase blood coagulation. It is characterized by the generation of microthrombi, which causes reduction in uteroplacental blood flow and consequently placental ischemic conditions. Mutations FV Leiden (G1691A), FII (G20210A), MTHFR (C677T) and PAI-1 are the most common genetic risk factors for thromboembolism. Several environmental factors and stress are associated with the occurrence of thrombophilia in pregnancy. However, many environmental factors are not considered in the assessment of the risk of developing thrombophilia in pregnancy due to their still unknown influence on the etiology of this disease. Identification of environmental risk factors associated with genetic predisposition to thrombophilia in pregnancy would enable individual assessment of the risk of developing this disease, and therefore more adequate prevention and therapy [1].

Machine learning (ML) algorithms have emerged as powerful tools for analyzing complex clinical data and making accurate predictions. By leveraging ML techniques, researchers have made significant progress in various medical domains, including disease diagnosis, prognosis, and risk prediction [2-3]. The aim of this research was to develop a new approach to identify environmental risk factors in pregnant women with thrombophilia using ML algorithms, with a special focus on decision trees and neural networks. Accurately predicting the risk of developing thrombophilia can enable healthcare professionals to identify high-risk pregnant women early, which leads to the improvement of preventive measures, such as lifestyle modifications and the use of therapeutic prophylaxis. In the following sections, we will detail the methodology, results, and discussion of our study, presenting the effectiveness of machine learning algorithms in thrombophilia prediction.

2. Materials and Methods

The studied population consisted of 35 pregnant women who are accommodated at the Gynecology and Obstetrics Clinic of the University Clinical Center Kragujevac for treatment or delivery, and 13 of whom had a diagnosis of thrombophilia. Thrombophilia in pregnant women was diagnosed based on confirmed mutations: FV Leiden (G1691A), FII (G20210A), MTHFR (C677T) and PAI-1. Clinical (spontaneous abortion, method of conception, week of pregnancy, systolic blood pressure, number of erythrocytes, myometrial thickness, fetal femur length) and demographic (smoking in pregnancy, number of cigarettes before pregnancy, place of residence, family relative (brothers and sisters), number of persons in the household, assessment of a healthy lifestyle) data were collected for each participant. The data on the way emotions regulated were collected from each pregnant woman using the Affective styles questionnaire. The Ethics Committee of the University Clinical Center Kragujevac approved the study protocol, and all patients gave informed consent before data collection.

The raw data is transformed into a suitable format for ML algorithms: decision Trees, recurrent (RNN) and convolutional (CNN) neural networks [2-3]. To evaluate the performance of the model, multiple metrics were utilized, including precision, recall, accuracy and F1 score. The entire study was implemented using Weka and Python.

Python libraries such as TensorFlow, Keras, and scikit-learn were employed for implementing neural networks, deep learning techniques, and performance metric calculations. To ensure the reliability and robustness of the model, a cross-validation technique, such as 10-fold cross-validation, was employed. By following these materials and methods, we aimed to develop a predictive model for thrombophilia in pregnancy using neural networks and decision trees.

3. Experiments

The following ML algorithms based on decision trees were applied to the demographic, lifestyle and clinical data: a DecisionTree classification in Python, and a RandomTree classification in Weka. Since the demographic and lifestyle data contained textual attributes, it was necessary to encode them using the OneHotEncoder package from the sklearn library, for DecisionTree classification. In Weka, it was necessary to use a filter for clinical data, ie. for the last class, NumericToNominal, to apply the RandomTree classification. The examples of following decision tree rules for demographic and lifestyle data were obtained and shown in Table 1.

Table 1. The decision tree rules

Demographic and lifestyle data	Clinical data
|--- SpontaneousAbortion <= 1.50 | |--- CigarettesNumberBeforePregnancy <= 7.50 | | |--- class: 2 | |--- CigarettesNumberBeforePregnancy > 7.50 | | |--- class: 0 |--- SpontaneousAbortion > 1.50 | |--- class: 2	WeekOfPregnancy < 39.52 | SystolicBloodPressure < 122.5 | | MyometrialThickness < 0.77 | | | FetalFemurLength < 7.78 : 2 (10/0) | | | FetalFemurLength >= 7.78 : 0 (1/0) | | MyometrialThickness >= 0.77 | | | FetalFemurLength < 7.71 | | | | ErythrocytesNumber < 4.69 : 0 (6/0) | | | | ErythrocytesNumber >= 4.69 : 2 (1/0) | | | FetalFemurLength >= 7.71 : 2 (2/0) | SystolicBloodPressure >= 122.5 : 0 (3/0) WeekOfPregnancy >= 39.52 : 0 (12/0)

Table 2. Classification results on demographic and lifestyle data.

	Classifier	Accuracy	Precision	Recall	F1-score
Demographic and lifestyle data	DecisionTree	71.67%	62.92%	70.83%	62.67%
	RandomTree	62.86%	62.30%	62.90%	62.50%
Clinical data	DecisionTree	85.83%	82.08%	85.00%	81.45%
	RandomTree	82.86%%	82.80%	82.90%	82.50%

Both trees take smoking during pregnancy and previous spontaneous abortions as key factors for predicting thrombophilia in pregnancy, while RandomTree considers the living environment as important factors. The methods were applied to clinical data in the same way, and again a simpler tree was obtained for the DecisionTree classifier,

where Fe level is crucial, while RandomTree showed that week of pregnancy, systolic blood pressure, myometrial thickness, fetal femur length and number of erythrocytes are important. There are very good classification results, over 80%, which means that clinical factors give us a more accurate prediction than demographic ones (Table 1 and 2).
Algorithms based on neural networks were applied to demographic and clinical data, namely MultilayerPerception in Weka, then using CNN and RNN in Python. Classification results for demographic and clinical data using the MultilayerPerception algorithm and classification accuracy using CNN and RNN are in Table 3. The results show that good prediction (above 85%) is achieved using MultilayerPerception (MP) on clinical data, and that RNN is more suitable for this data, which is expected considering its application. CNN should rather be used in future if the dataset will be enlarged using some images (some health recording, for example ultrasound etc.).

Table 3. Classification results on clinical and demographic data using MP, CNN and RNN

Data	Accuracy MP	Precision MP	Recall MP	F1-score MP	Accuracy CNN	Accuracy RNN
Clinical	85.71%	87.4%	85.7%	85.9%	66.67%	79.17%
Demographic	74.29%%	74.8%	74.3%	74.5%	32.5%	24.85%

4. Conclusions

In this study, we developed a prediction model of thrombophilia in pregnancy using decision trees and neural network techniques based on a comprehensive data set consisting of clinical, lifestyle and demographic information. Through our experiments, we demonstrated the potential of decision trees and neural networks in predicting thrombophilia in pregnancy. The model achieved promising results, outperforming the baseline models, and showing high precision, recall, accuracy and F1 score. These findings highlight the effectiveness of ML algorithms, with over 85% accuracy achieved on clinical data in identifying women at risk of developing thrombophilia in pregnancy. Despite the promising results, there are several areas for future research. We can expect better results using FE techniques, other ML algorithms or enlarging the dataset. The developed model is promising in the precise identification of women with risk of developing thrombophilia during pregnancy, which leads to the development of personalized preventive and therapeutic measures.

References

[1] M. J. Kupferminc, *Thrombophilia and pregnancy*, Curr Pharm Des., 11(2005) 735-748.
[2] F. Shamout, T. Zhu, D. A. Clifton, *Machine learning for clinical outcome prediction*, IEEE Rev. Biomed. Eng., 14 (2021) 116–126.
[3] M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang, *Disease prediction by machine learning over big data from healthcare communities*, IEEE Access, 5 (2017) 8869–8879.