An Inference Mechanism Using Bayes-based Classifiers in Pregnancy Care

Mário W. L. Moreira[1,2], Joel J. P. C. Rodrigues[1,3,4], Antonio M. B. Oliveira[2], Kashif Saleem[4], Augusto V. Neto[5]

[1] Instituto de Telecomunicações, Universidade da Beira Interior (UBI), Covilhã, Portugal
[2] Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Aracati, Brazil
[3] National Institute of Telecommunications (INATEL), Santa Rita do Sapucaí, MG, Brazil
[4] University of Fortaleza (UNIFOR), Fortaleza, CE, Brazil
[4] Center of Excellence in Information Assurance (CoEIA), King Saud University (KSU), Riyadh, Kingdom of Saudi Arabia
[5] Federal University of Rio Grande do Norte (UFRN), Natal, RN, Brazil

mario.moreira@it.ubi.pt; joeljr@ieee.org; amauroboliveira@gmail.com; ksaleem@ksu.edu.sa; augusto@dimap.ufrn.br

Abstract— **Significant advances on smart decision support systems (DSSs) development have influenced important results on pregnancy care. Nevertheless, even considering the efforts to reduce the number of women deaths due to problems related to pregnancy, this decrease presented less impact than other areas of human development. Hypertensive disorders in pregnancy, particularly pre-eclampsia and eclampsia, account for significant proportion of perinatal morbidity and maternal mortality. In this context, this paper proposes an inference model that uses data mining (DM) techniques capable for operating in a data set to extract patterns and assist in knowledge discovery. Identifying hypertensive crises that complicate pregnancy, it can impact in a meaningful reduction the incidence of sequelae and death of pregnant women. Comparison between two Bayesian classifiers is performed in this work to better classify the hypertensive disorders severity. Results showed that Naïve Bayes classifier had an excellent performance, presenting better precision and F-measure, compared to the other experimented classifiers. Even finding a good performance to predict hypertensive disorders, other Bayesian methods need to be evaluated, as well as other DM techniques such as those based on artificial intelligence (AI) and tree-based methods.**

Keywords— eHealth; Hypertension; Pregnancy; Decision Support Systems; Data Mining; Bayes Methods

I. INTRODUCTION

According to the World Health Organization (WHO), hypertensive disorders of pregnancy afflict about 10% of all pregnancies around the world. These are the leading causes of morbidity, disability, and death among mothers and babies [1]. These complications during pregnancy were an important cause of mortality in Latin America and the Caribbean, contributing to 22.1% of all maternal deaths in this region [2]. With providing timely and efficient care, the majority of deaths related to these complications could be avoided. Thus, optimization of health care for pregnant women to prevent and treat hypertensive disorders is needed.

As reported by the National High Blood Pressure Education Program, hypertension in pregnancy is classified into one of the following five categories: (*i*) Chronic hypertension, (*ii*) Preeclampsia, (*iii*) Chronic hypertension with superimposed preeclampsia, (*iv*) Gestational hypertension, and (*v*) Transient hypertension [3]. These categories are critical to differentiating preeclampsia, a pregnancy-specific syndrome of exaggerated vasoconstriction and reduced organ perfusion, from pre-existing chronic hypertension. Nevertheless, this complex multifactorial syndrome, that occurs in about 5 to 7% of pregnancies worldwide, has not an etymology established yet, i.e., this disease has still no agreement about its classification as well as on the timing of its occurrence during pregnancy. To evaluate this disorder is necessary to define the blood pressure status. If the pregnant woman is hypertensive, the health expert assesses its severity, possible secondary causes, and damage presence in organs, to plan treatment strategies. The treatment of chronic cases in the first trimester of pregnancy is critical since fetal loss rate is about 50% and maternal mortality is significant in these cases. Preeclampsia is more common in pregnant women that already suffered from chronic hypertension, with an incidence of approximately 25%. To prevent more severe problems is recommendable the early recognition of high-risk pregnancies, a constant clinical, laboratory, and intensive monitoring when indicated. In this sense, information and communication technologies (ICT) play a key role to improve the quality of life of pregnant women. With the development of intelligent systems for risk pregnancy monitoring, health experts can identify serious problems caused by gestational hypertension at its early stages, saving lives of both mothers and babies. Several technology solutions are already being deployed to combat preeclampsia in their most critical condition [4], [5]. Many approaches have achieved proper evaluation but are still unable to reduce the critical situation of maternal and fetal deaths by themselves, mainly, in developing countries. Smart DSSs are considered a goof tool capable to contribute to this goal. Then, this paper proposes an inference model that uses data mining (DM) techniques capable for operating in a data set to extract patterns and assist in knowledge discovery. Identifying hypertensive crises that complicate pregnancy, it can impact in a meaningful

978-1-5090-3370-6/16/$31.00

reduction the incidence of sequelae and death of pregnant women.

The paper is organized as follows. Section II presents related works discussing used methods on identifying high-risk pregnancies. Section III presents two modeling proposals that use Bayes-based classifiers capable of identifying hypertensive disorders as from its symptoms. Section IV performs the performance evaluation study of these methods and the results analysis considering the proposed plans. Finally, Section V provides the conclusion and suggestions for further works.

II. Related Work

This topic discusses recent works that are significant efforts to identify hypertensive disorders risks on pregnancy, and then, researches that use Bayes-based classifiers to collect valuable information from a data set.

LARIISA platform is an intelligence solution that aims to specify and implement, from an analytical and experimental study of DM methods, data warehouse, ontologies, and mashups [6]. It provides governance intelligence in the decision-making process in five traditional areas of management (systemic, legal, functional, clinical, and care). The knowledge management framework is the platform component responsible for turning it in a basis for building smart systems. This segment considers three blocks, namely, inference engine, semantic search, and DM. The intelligence management model comprises the adequacy of the mining block data to the knowledge management component (Intelligence). Comparative analysis of pattern recognition algorithms requires a deep study to define which is the best-suited DM model. Moreover, it is also important to analyze how the system should (or not) learn after the inclusion of new information using ML techniques. Taking into account this information classification performance evaluation can be executed.

Recent studies found hereditary factors in occurrence of preeclampsia. Dutta *et al.* conduct a study about critical genes that cause preeclampsia using microarray gene expression data [7]. In this research, a matrix with rows representing genes and columns representing expression levels of the genes is obtained. This method performs meta-heuristic algorithms assessment, to know the Variable Neighborhood Search (VNS) and the Differential Evolution (DE) in the critical genes identification. The k-Nearest Neighbor (kNN) algorithm determines the fitness of set key genes to classify samples into normal and diseased states. Results showed that the DE approach outperforms the other method regarding best fitness values and average when executed at the same runtime.

The leading causes of preeclampsia are not known. However, its risk factors have been already defined. Cheng *et al.* study the effects of these risk factors on the gestational age [8]. The authors examined whether the consequences of these factors are different on preeclampsia and gestational hypertension. This method applies varying coefficient models in this approach. Results showed that these effects behave differently between these two groups. As the criterion to assess the accuracy of the estimated varying coefficient model, the relative error was used.

The presence of the diastolic notch in the Doppler of uterine artery is the best predictive signal after twenty weeks of pregnancy. Das *et al.* propose a method based on Doppler signal analysis to identify automatically high-risk pregnancy conditions with higher sensitivity even when the clinical parameters are presenting healthy state [9]. The acquisition of uterine artery signals from both women with healthy and at-risk pregnancy is made using a color-based Doppler ultrasound system. An algorithm extracts characteristics from the Doppler flow waveform and compares them between the two types of pregnancies. This study presented the possibility of identifying changes in the quantity of the blood supply to the fetus, improving the performance of the Doppler uterine artery investigation in high-risk pregnancies.

III. Modeling Proposals Using Bayes-based Classifiers

Hypertension in pregnancy occurs when there are high-pressure levels in pregnant women. Pregnancy-induced hypertension refers to the onset of hypertension as a result of pregnancy, occurring after 20 weeks of pregnancy and disappearing until six weeks after childbirth. Hypertension is the increase of blood pressure above 140/90mmHg. With the pregnant woman seated, the health expert should measure the blood pressure and confirm, after the rest period, for three times. Gestational hypertension diagnosis occurs when diastolic blood pressure is above 90mmHg or when it has a blood pressure increase above 15 mmHg of the value measured before first 20 weeks of gestation. Preeclampsia is the hypertension occurrence accompanied by loss of protein in the urine (proteinuria) after 20 pregnancy weeks with or no edema. About the protein loss, significant proteinuria is considered with the values equal or greater than 300mg of protein in urine collected for 24 hours. The diagnosis is clinical and laboratory: measurement of blood pressure, edema examination, and protein quantity in urine. Figure 1 shows a graphical contextualization of preeclampsia within hypertensive disorders.

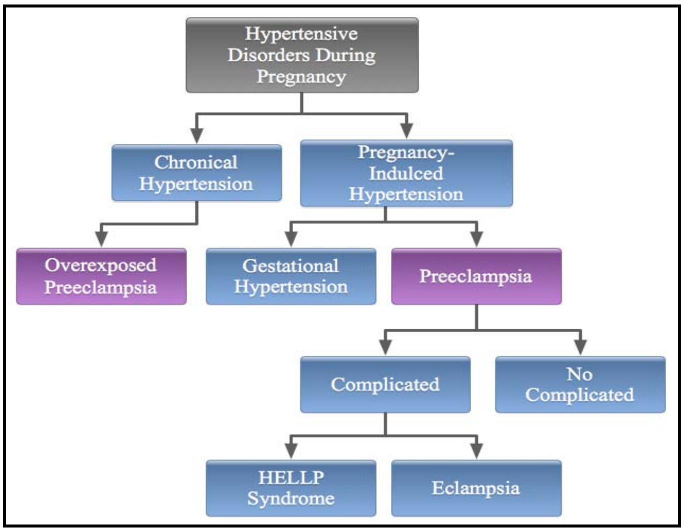

Fig. 1. Preeclampsia contextualization scheme.

A. The Naïve Bayes Classifier

The Naïve Bayes classifier is applicable in health care when there is a set of attributes that represents each risk factor. Each one of these attributes occurs in a particular disorder hypertensive. This classifier based on Bayes theorem is used to determine the probability of each hypertensive disease from symptoms never seen based on training examples.

Shaikh *et al.* develop an intelligent DSS using DM techniques to predict heart diseases [10]. This system uses an application based on the Naïve Bayes classifier in which patient answers predefined questions. It uses hidden data of a database and compares the user answers with a training data set. Results showed that this smart system can assist healthcare practitioners in the decision-making clinical process which other conventional decision support systems cannot.

According to a Bayes-based method, this work aims to classify new cases according to its most probable classification of hypertensive disorder in pregnancy given its set of symptoms $< a_1, a_2, \dots, a_n >$ using the Equation (1).

$$v_{MAP} = argmax_{v_{j \in V}} \frac{P\left(a_1, a_2, \dots, a_n \middle| v_j\right) P(v_j)}{P(a_1, a_2, \dots, a_n)} \tag{1}$$

In the Eq. (1), v_{MAP} is the case with maximum a posteriori probability, i.e., one that generates better results for data sets never seen before, considering the occurred training period. First of all, it is necessary to estimate from the training set, the probability $P(v_j)$ that is straightforward to determine. However, assuming that this set is limited in terms of size, it becomes difficult to predict $P(a_1, a_2, \dots, a_n|v_j)$ because there are probably few or no occurrence identical in the training set. If the training set is too large, this second probability could be well estimated. The Naïve Bayes classifier makes a simplification assuming that attributed values to the symptoms are independent, i.e., the probability of observing $a_1, a_2, \dots, a_n$ is precisely the product of the probabilities associated with each attribute, as shown in equation (2).

$$P\left(a_1, a_2, \dots, a_n \middle| v_j\right) = \prod_i P(a_i|v_j) \tag{2}$$

Thus, the Naïve Bayes classifier is a simplification, which is given by equation (3).

$$v_{NB} = argmax_{v_j \in V} P\left(v_j\right) = \prod_i P(a_i|v_j) \tag{3}$$

Figure 2 illustrates graphically the structure of Naïve Bayes classifier.

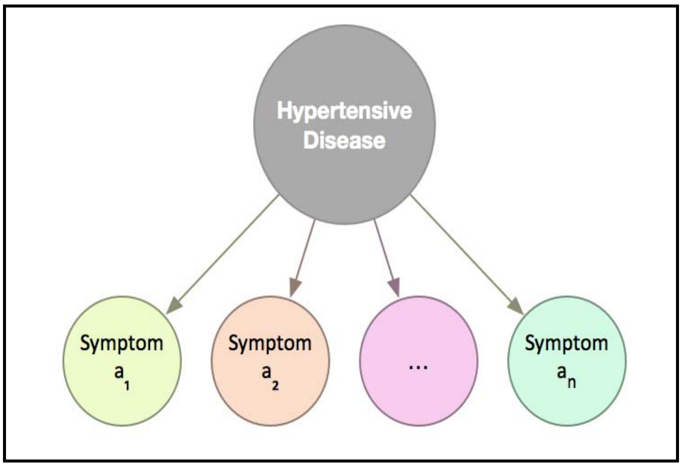

Fig. 2. Illustration of the Naïve Bayes classifier structure where each node representing a symptom is independent.

B. The Averaged One-dependence Estimators classifier (AODE)

The Averaged One-dependence Estimators classifier (AODE) adopts another approach to minimize the dependence of attributes. In fact, it extends the structure of the Naive Bayes classifier also including the relationship of each attribute to with other. Figure 3 illustrates this model. This method shows the influence of characteristics to each other, unlike the previous classifier. In [11], the authors discuss the AODE classifier, considering it as one of the most attractive semi-Naïve Bayesian classifiers and, hence, a great alternative to the Naïve Bayes classifier. This promising classifier obtains low error rates, maintaining under control the computational complexity. Results show the importance of selecting a proper number of nodes on the learning of both classifiers. The number of nodes is decisive to provide accurate results.

In healthcare, Kovács and Hajdu use a novel strategy for the segmentation of the vascular system in retina images [12]. Based on the Hidden Markov Random Fields approach, this work considers the tangent vector field of the image to improve the connectivity of the vascular system. Thus, it extends the optimization problem of this model. To improve the probability estimation during the solution of this optimization problem, the AODE classifier is used instead of the well-known Naïve Bayes classifier, because it uses a weaker assumption than the total independence of nodes. Figure 3 shows the AODE classifier structure in which each attribute has relations with others.

This method seeks to estimate the probability of each class given a specified set of features $a_1, \dots, a_n$, $P(v_j|a_1, \dots, a_n)$ using the equation (4).

$$\hat{P}\left(v_j \middle| a_1, \dots a_n\right) = \frac{\sum_{i:1 \le i \le n \wedge F(a_i) \ge m} \hat{P}(v_j, a_n) \prod_{j=1}^{n} \hat{P}(a_j|v_j, a_i)}{\sum_{v_j' \in V} \sum_{i:1 \le i \le n \wedge F(a_i) \ge m} \hat{P}(v_j', a_i) \prod_{j=1}^{n} \hat{P}(a_j|v_j', a_i)} \tag{4}$$

In the eq. (3), $\hat{P}$ denotes an estimate of P. F represents the rate in which the argument appears in the sample data. A term must have minimum frequency m to be used in the summation. Usually, m is set at 1.

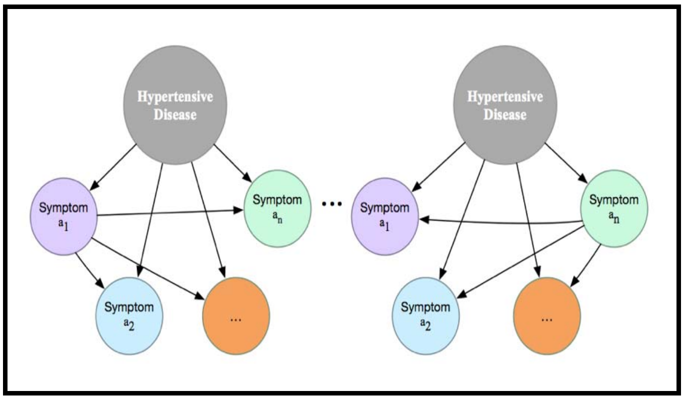

Fig. 3. Illustration of the AODE classifier structure where each attribute has relations with others.

IV. Performance Evaluation and Result analysis

The healthcare dataset used for classification includes twenty-five hypertension disorders cases, each one with eleven attributes provided by health experts.

To evaluate the performance of the proposed approaches, the 5-fold cross-validation method was used. This technique assesses the ability of generalization of a model from a data set. In problems where the goal of the modeling is the prediction, this method is widely used. The primary aim is to estimate how much a model is accurate. This research used a confusion matrix, the cross-validation approach, and the F-measure to evaluate the accuracy of the models constructed by both classifiers. The confusion matrix is a table where TP (true positives), TN (true negatives), FP (false positives), and FN (false negatives) are represented. The cross-validation method allows the use of all the database data for training and testing. This work adopted five folds in this method. F-measure is the mean between precision and recall, measuring the ability to recognize negative and positive cases.

Based on [13], this work has chosen the best nodes for modeling. The main symptoms presented in pregnant women suffering from hypertensive disorders are a headache, epigastric pain, nausea or vomiting, blurred vision, giddiness, hyperreflexia, edema, oliguria, hypertension, and proteinuria. These are the nodes that make up the proposed model. Tables I and II show the evaluation results of both methods.

TABLE I. Performance Evaluation of the Naïve Bayes Classifier.

TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.333	0.211	0.333	0.333	0.333	0.684	CH
0.714	0.111	0.714	0.714	0.714	0.817	P
0.167	0.158	0.250	0.167	0.200	0.509	OP
0.333	0.316	0.250	0.333	0.286	0.632	GH

TABLE II. Performance Evaluation of the AODE Classifier.

TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area	Class
0.333	0.316	0.250	0.333	0.286	0.675	CH
0.714	0.167	0.625	0.714	0.677	0.794	P
0.000	0.158	0.000	0.000	0.000	0.412	OP
0.167	0.263	0.167	0.167	0.167	0.623	GH

CH (Chronic hypertension), P (Preeclampsia), OP (Chronic hypertension with superimposed preeclampsia), and GH (Gestational hypertension)

Table III presents the performance evaluation of both approaches using a weighted average.

TABLE III. Global Performance Evaluation of both Bayes-based Classifiers

	TP Rate	FP Rate	Precision	Recall	F-measure	ROC Area
NB	0.400	0.195	0.400	0.400	0.397	0.667
AODE	0.320	0.224	0.275	0.320	0.295	0.633

Figure 4 shows the ROC (Receiver Operator Characteristic) curve. The ROC curves describe the discriminative ability of a diagnostic experiment. It allows highlighting the values in which there is further optimization of the sensitivity according to the specificity. The point in a receiver operator characteristic curve where this happens is the one that is nearest to top left of the diagram. Optimal models show points closest to the point (0,1).

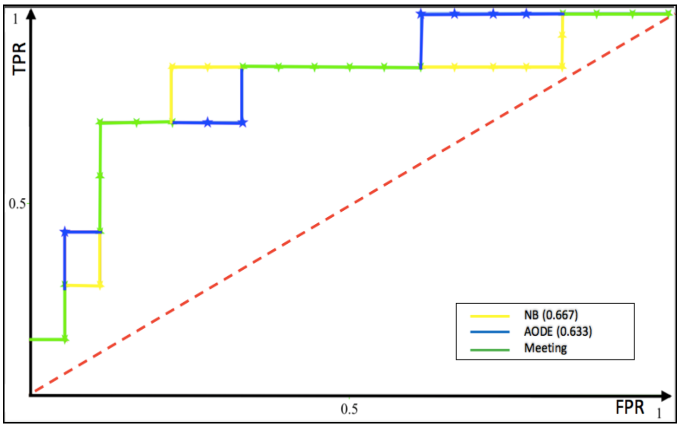

Fig. 4. Receiver Operator Characteristic (ROC) curve presenting the Naïve Bayes classifier (yellow line), the AODE classifier (blue), and the meeting between these curves (green).

The Naïve Bayes classifier showed the same value for the accuracy and F-measure equal to 0.714, while the AODE classifier presented 0.625 and 0.677, respectively. Both classifiers achieved the same performance for the recall of 0.714. These classifiers also presented a ROC area of 0.817 (Naive Bayes) and 0.794 (AODE). Great models present results for this indicator close to 1.000. In comparison with other works, in general, with [8] and [14], that show an accuracy about 0.860 and 0.836 respectively, this proposal presents a good performance given the small amount of used data. The Kappa statistic evaluates the prediction performance of classifiers, deducting the expected number of correct results. The Naïve Bayes classifier had a performance of 0.1987, while the AODE classifier had a performance of 0.0899. A possible performance interpretation of these proposed models, from Kappa statistics, comes from the fact that the presented results had a reasonable performance.

V. CONCLUSION

This paper discussed a model for hypertensive diseases in risk pregnancies using DM techniques. For identification of the mortality leading cause during pregnancy, namely, the preeclampsia, two classifiers were considered and evaluated. Both of them presented a good performance. The main result of this study shown that minimizing the dependence of all the attributes they reduced the accuracy of the model. Further works will discuss the relationship between nodes, and they will evaluate other classifiers. Decision trees and AI based classifiers can be the key to find a better relationship between nodes, making possible a more precise model. Experimenting these models on a larger database will also show if the amount of data positively affects the accuracy.

This work was an initial effort to develop a better intelligent mechanism, which aims to provide smart governance in the decision-making process by various actors in the public health area. For health professionals that work with pregnancy is important to be conscious of the medical history and changes in the pregnant woman's clinical state because these changes may not come always with a high blood pressure, complicating the decision-making process. A more specialized care is essential for preeclampsia monitoring with the purpose to personalize assistance for the prevention, promotion, and health recovery. To minimize complications and fatal consequences specific strategies with agility and efficiency are needed. In this context, information and communication technologies are essential and play a key role to reduce these consequences.

ACNOWLEDGEMENTS

This work has been partially supported by the *Instituto de Telecomunicações*, Next Generation Networks and Applications Group (NetGNA), Portugal, by National Funding from the FCT - *Fundação para a Ciência e a Tecnologia* through the UID/EEA/50008/2013 Project, by Finep, with resources from Funttel, Grant No. 01.14.0231.00, under the Radiocommunication Reference Center (*Centro de Referência em Radiocomunicações* - CRR) project of the National Institute of Telecommunications (*Instituto Nacional de Telecomunicações* - Inatel), Brazil, by the National Plan of Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, Kingdom of Saudi Arabia, Award Number (12-INF2817-02), and by *Ciências sem Fronteiras* of *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), Brazil, through the process number 207706/2014-0.

REFERENCES

[1] WHO, “Recommendations for Prevention and treatment of pre-eclampsia and eclampsia,” WHO Library Cataloguing-in-Publication Data, Geneva, Switzerland, 2011.

[2] L. Say, D. Chou, A. Gemmill, Ö. Tunçalp, A. Moller, J. Daniels, A. M. Gülmezoglu, M. Temmerman, and L. Alkema, “Global causes of maternal death: a WHO systematic analysis,” The Lancet Global Health, Vol. 2, N. 6, pp. 323-333, 2014.

[3] National High Blood Pressure Education Program, “The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure,” National Institutes of Health, Bethesda (MD), USA, 2004.

[4] J. F. Torrado, I. Farro, F. Farro, D. Bia, Y. Zocalo, C. Sosa, S. Scasso, S. Zunini, R. L. Armentano, and J. Alonso, “Carotid-radial pulse wave velocity as an alternative tool for the evaluation of endothelial function during pregnancy: potential role in identifying hypertensive disorders of pregnancy,” 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2012), San Diego (CA), USA, Aug. 28 – Sept. 1, 2012, pp. 5473–5474.

[5] H. D. Tagare, K. Rood, and I. A. Buhimschi, “An algorithm to screen for preeclampsia using a smart phone,” IEEE Healthcare Innovation Conference (HIC 2014), Seattle (WA), USA, Oct. 8-10, 2014, pp. 52-55.

[6] M. Oliveira, L. O. M. Andrade, M. Santos, G. Teles, and N. Agoulmine, “Towards a cost-effective homecare for a public health management system in Brazil,” IEEE 16th International Conference on e-Health Networking, Applications and Services (HEALTHCOM 2014), Natal, Brazil, Oct. 15-18, 2014, pp. 206–211.

[7] J. Dutta, S. Biswas, S. Saha, and S. Acharyya, “Identification of Disease-Critical Genes Causing Preeclampsia: Meta-heuristic Approaches,” IEEE UP Section Conference on Electrical Computer and Electronics (UPCON 2015), Allahabad, Dec. 4-6, 2015, pp. 1–6.

[8] W. Cheng, L. Fang, L. Yang, H. Zhao, P. Wang, and J. Yan, “Varying Coefficient Models for Analyzing the Effects of Risk Factors on Pregnant Women’s Blood Pressure,” 13th International Conference on Machine Learning and Applications (ICMLA 2015), Detroit (MI), Dec. 3-5, 2014, pp. 55–60.

[9] R. Das, P. Vajinepalli, R. Sisodia, and L. Gupta, “Automatic doppler signal analysis to assess utero-placental circulation for identifying high risk pregnancies,” 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS 2013), Osaka, Japan, July 3-7, 2013, pp. 3905–3908.

[10] S. Shaikh, A. Sawant, S. Paradkar, K. Patil, and A. E. System, “Electronic Recording System-Heart Disease Prediction System,” International Conference on Technologies for Sustainable Development (ICTSD2015), Mumbai, India, Feb. 4-6, 2015, pp. 2–6.

[11] M. J. Flores, J. A. Gámez, A. M. Marténez, and A. Salmerón, “Mixture of truncated exponentials in supervised classification: Case study for the naive Bayes and averaged one-dependence estimators classifiers,” International Conference on Intelligent Systems Design and Applications (ISDA 2011), Córdoba, Spain, Nov. 22-24, 2011, pp. 593–598.

[12] G. Kovács and A. Hajdu, “Extraction of vascular system in retina images using Averaged One-Dependence Estimators and orientation estimation in Hidden Markov Random Fields,” IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2011), Chicago (IL), USA, March 30 – April 2, 2011, pp. 693–696.

[13] M. W. Moreira, J. J. P. C. Rodrigues, A. M. Oliveira, R. F. Ramos, and K. Saleem, “A Preeclampsia Diagnosis Approach using Bayesian Networks,” IEEE International Conference on Communications (ICC 2016), Kuala Lumpur, Malaysia, May 23-27, 2016, pp. 1432-1436.

[14] C. K. Neocleous, P. Anastasopoulos, K. H. Nikolaides, C. N. Schizas, and K. C. Neokleous, “Neural networks to estimate the risk for preeclampsia occurrence,” International Joint Conference on Neural Networks (IJCNN 2009), Atlanta (GA), USA, June 14-19, 2009, pp. 2221-2225.