Proposed model to predict preeclampsia using machine learning approach

Raden Topan Aditya Rahman1,2, Muhammad Modi Lakulu3, Ismail Yusuf Panessai1, Esti Yuandari4, Ika Mardiati Ulfa5, Fitriani Ningsih5, Lensi Natalia Tambunan5

1Doctoral Program Student, Faculty of Computing and Meta Technology, Sultan Idris Education University, Tanjung Malim, Malaysia
2Polytechnic of Health Borneo Citra Medika, Tanah Laut Regency, Indonesia
3Faculty of Computing and Meta Technology, Sultan Idris Education University, Tanjung Malim, Malaysia
4Faculty of Health, Sari Mulia University, Banjarmasin, Indonesia
5Betang ASI Raya Midwifery Academy, Pahandut, Indonesia

ABSTRACT

Pregnancy complications, which are the biggest cause of death in productive women, are more common in developing countries with low incomes. One of the contributors to death in pregnant women is preeclampsia which contributes 2-8% every day. Based on research results, more than 70% of the use of technology can be a solution for early prevention in detecting cases of pregnancy. The aim of this research is to build a model for early detection of preeclampsia using a machine learning approach. Sample using retrospective data with sample size 1.473. Based on the result, decision tree (DT) is the best model with accuracy 92.2% (area under curve (AUC): 0.91; Spec: 92.3; and Sens: 83.6), according to weigh correlation we can show 3 (three) highest features causes preeclampsia is history of hypertension, history of diabetes mellitus, and history of preeclampsia. The health of pregnant women is essential in the development of the fetus, so it needs optimal monitoring. Monitoring during pregnancy can now be done through technology-based examinations for assist health workers in making decisions during pregnancy.

1. INTRODUCTION

Preeclampsia is one of cause of maternal death after bleeding. World Health Organization (WHO) estimates that the cases of preeclampsia are seven times higher in developing countries than in developed countries. In the last two decades, there has been no significant decrease in the incidence of preeclampsia. The magnitude of this problem is not only because preeclampsia has an impact on the mother during pregnancy and childbirth, but also causes post-partum problems due to endothelial dysfunction in various organs, such as the risk of cardiometabolic diseases and other complications. Thus, preeclampsia becomes a serious medical problem and complex [1]. There is no active screening evaluation for preeclampsia so that efforts to prevent preeclampsia are not optimal which can lead to increased morbidity and mortality. Therefore, there is an urgency for recommendations based on scientific evidence to assist practitioners in diagnosing, evaluating, and managing preeclampsia.

To date, there have been various biomarker findings that can be used to predict the incidence of preeclampsia, but no single test has high sensitivity and specificity. This occurs because the features used are...
mostly related to characteristics. A series of checkups using more features is urgently needed through technological assistance to screen the risk of preeclampsia in pregnant women from the start of their pregnancy, so health practitioners can identify risk factors for preeclampsia and control them as a form of primary prevention [2]. Pregnancy screening at dr. H. Moch Ansari Saleh General Hospital is currently still being carried out conventionally, which is through antenatal checkups. However, this is still not optimal and has the potential to produce inaccurate results, considering that cases of preeclampsia still occur and are increasing every year [3].

According to a statement by a research expert who is also a specialist in Obstetrics and Gynecology, most health workers in dealing with pregnancy complications are still focused on treatment, which will definitely cost more than prevention in terms of funding [4]. dr. H. Moch Ansari Saleh General Hospital Banjarmasin as a hospital owned by the local government and also as a referral hospital, is a hospital that has a high incidence of preeclampsia. According to a preliminary study conducted during the last 3 (three) years, the number of cases has been increasing. The incidence of preeclampsia in 2020 was 168 (5.92%) of 3,007 mothers who gave birth; in 2021 there were 145 (5.17%) cases of 2,804 women who gave birth and increased again in 2022 as many as 178 (9.82%) of 1,813 cases of mothers giving birth [5], [6]. These data indicate the high incidence of preeclampsia in dr. H. Moch Ansari Saleh General Hospital Banjarmasin every year.

Artificial intelligence can be applied in health care for modeling, diagnosis, early detection, and monitoring. Another potential difficulty is the noisy class of outcomes due to the disease’s varied gene expressions. Such issues can be solved using machine learning technologies [7]–[9]. It has the potential to assist in decision-making and improve medical care. Machine learning is well-suited for predictive modeling of pregnancy outcomes [10], [11]. Machine learning, also known as supervised, semi-supervised, unsupervised, or reinforcement learning, is a subset of artificial intelligence that involves the use of algorithms and computer models to achieve a certain goal. In decision-making settings, machine-learning algorithms are commonly utilized to achieve higher predictive accuracy than traditional generalized linear models [12]. Furthermore, deep learning is a machine learning technique that employs neural networks, similar to neurons in the human brain, to extract many levels of data representation from a given input in order to solve a problem [13].

Preeclampsia is still a serious problem that needs to be handled optimally. Until now, the etiology of preeclampsia has not been clearly explained [14]. So that, early diagnosis of preeclampsia remains a clinical challenge [15]. Researcher found several issues that have not been addressed in the studies reviewed, which indicate gaps in the literature:

First, based on the statement of one of the research experts who is also a specialist in Obstetrics and Gynecology the pathogenesis of preeclampsia still not fully explained [4].

Second, so far, early detection of preeclampsia is done conventionally through antenatal care examinations, so there is a need for artificial technology support that can diagnose the incidence of preeclampsia [3].

Third, artificial intelligence-based technology is expected to be able to improve health services, especially in pregnancy. Therefore, it is hoped that further research is highly recommended to find the most effective model in diagnosing pregnancy specifically preeclampsia [16].

According to the background above, it can be concluded that the case of preeclampsia still becomes the crucial case which requires the optimal solution. Regarding the preeclampsia checking that is still carried out conventionally, this creates an impact on the management of preeclampsia which is still limited to only the medication to date. Even if there are already several studies related to preeclampsia using technological approach through machine learning, there are no results that can be preventive actions in handling preeclampsia. Furthermore, the results of previous studies only focused on the accuracy number of a model, so there are still needs for the model development that is applied by providing features to show the primary causes of preeclampsia. Therefore, by the results of this study that compare performance metric and feature of importance, the researcher believes that it can result a new finding which gives contribution to the health field in providing the early prediction tool for preeclampsia as the preventive steps, as well as the shift of conventional method to the technological one. Based on the introduction above, researcher is interested to propose model to predict preeclampsia using machine learning approach.

2. PROPOSED MODEL

Conducted a preliminary study using the literature review on predicting the incidence of preeclampsia using machine learning. From the results of the literature review, articles that are relevant to the research are selected, which is then analyzed on theories related to preeclampsia, gaps, problem statements, methods used, and recommendations for further research. Based on the results of the analysis of the article, it is then used as the basis for preparing the background according to the topic to be studied.
Figure 1 shows the working process auto model of the machine learning approach. It has stages such as preliminary study, development of machine learning model, and evaluation model. The preliminary study begins by asking for permission to collect pregnancy data at the hospital. The data taken were preeclampsia data as a case sample, and data on pregnant women who were not preeclampsia as a control sample.

The next phase is model development. This stage begins with identify related to the algorithm model used and identify the factors that cause preeclampsia. We developed a machine learning model to predict the incidence of preeclampsia. The data that has been obtained is then analyzed through three stages, namely the design process, the qualification process and the validation process. Model validation using machine learning that begins with the stage of building a model using auto model. Auto model provides with a selection of models that are relevant for the problem [17]. If there is no time constraint, the best option is probably to build all of them, and compare their performance once they are finished. Typically, we have to decide on priority like the accuracy of the finished model and time it takes to build it. Auto model helps to arrive at a reasonable compromise.

The next step, enter testing data with parameter threshold value of all features and build model algorithm. After this stage is complete the algorithm will get the output of the best model, and the last to the preprocessing stage of the model to evaluation. The next phase after validation is model evaluation. Evaluate the model by looking at the level of accuracy, sensitivity, and specifications. From these results it can be seen which model is the best can be used in predicting the incidence of preeclampsia.

The Figure 1 below is the working process of the machine learning method system using auto model algorithm in predicting the incidence of preeclampsia.

3. METHOD

Current machine learning techniques have been extensively developed to solve differential equations, particularly in the field of healthcare. Proposed model to predict preeclampsia using machine learning approach was developed in three phases, which are preliminary study, select input and model types, as well as performance metrics.
3.1. Phase one: preliminary study

The preliminary study begins to prepare the data by changing the data according to the specified coding before being entered into the machine learning model. The first technical stage that would be carried out were checking and handling missing or incomplete data, then mapping categorical and nominal data into ordinal data. The next step was the pre-processing stage with feature scaling using normalization and standardization techniques. For permission to collect pregnancy data at the hospital, the entire population in this study was pregnant women. Meanwhile, the sample was divided into 2 (two), namely pregnant women with preeclampsia and pregnant women without preeclampsia. The sample was divided into 2 (two), namely pregnant women with preeclampsia and pregnant women without preeclampsia. Based in result Aung et al [18] using a comparison of case data and control data of 1:2 the best ratio. Based on data in medical record from ansari saleh general hospital. This study used a sample of 491 cases (pregnant women with preeclampsia) and a control sample (pregnant women without preeclampsia) of 982, so the total sample is 1.473. Sampling technique using systematic random sampling. Features using in this research 14 (fourteen) such as: mother age, profession, education, gravidity, parity, antenatal care (ANC), pregnancy interval, body mass index, hemoglobin, history of abortion, history of cesarean section (CS), history of diabetes mellitus, history of hypertension, and history of preeclampsia. and for label feature is preeclampsia.

3.2. Phase two: select input and model types

The next stage is select input and model types using auto model. Split data is a validation technique which divides the data into two parts randomly, some as training data and some others as testing data. By using split data will be carried out by training trials based on a predetermined split ratio before, for then the remainder of the split ratio training data will be considered as data testing. Training data is data that will be used in training doing learning while data testing is data that has never been used as learning and will serve as data testing the truth or accuracy of the results learning [19]. The selected input explains about features selection from attributes which use in this research. Ineligible features on this research will be exclude from the model. After that build model using auto model such as Naïve Bayes (NB), generalized linier model (GLM), logistic regression (LR), fast large margin (FLM), deep learning, decision tree (DT), random forest (FR), gradient boosted trees (GBT), and support vector machine (SVM).

3.3. Phase three: performance metrics

The next phase after validation was model evaluation. Table 1 performance metrics as a evaluated the model by looking at the level of area under curve (AUC), accuracy, sensitivity, and specificity. Accuracy is a metric describes how accurately the model can classify the correct ratio (positive and negative) with the overall data. AUC is a prediction function on each training set, evaluates the true performance (or the true) probability of the correct ranking of two randomly selected observations, where one is a positive sample and the other a negative sample in the appropriate validation set, and finally, takes an above average validation set. The true value of this target parameter is random, as it depends on the separation of the sample data into training sets and the fit of the corresponding prediction function. Sensitivity (true positive rate) is a metric measures the proportion of negatives that are correctly identified. And recall or specificity (true negative rate) is a metric measures the proportion of negatives that are correctly identified as negative or had no condition. From these results, it can be seen which model is the best can be used as the best model in predicting the incidence of preeclampsia [20].

<table>
<thead>
<tr>
<th>Actual label</th>
<th>Predicted labelling</th>
<th>Predicted labelling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive class</td>
<td>True positive (tp)</td>
<td>False negative (fn)</td>
</tr>
<tr>
<td>Negative class</td>
<td>False positive (fp)</td>
<td>True negative (tn)</td>
</tr>
</tbody>
</table>

4. RESULTS AND DISCUSSION

The model developed in this research uses an auto model that is simulated using a rapid miner. The model performance results are then compared for each algorithm by looking at the accuracy value. Then look at the correlation weights in the selected algorithm to determine the important features which are important features for the occurrence of preeclampsia.

4.1. Performance metrics

Preprocessing the data before analysis is carried out to ensure that there are no errors in the data input so that missing does not occur during analysis. After handling the missing value, the next step is to
divide the dataset into 80% for testing data and 20% for training data [21]. After that, carry out analysis with an auto model which will be the final result by comparing the accuracy values for each algorithm. The figures below are the modeling results of each algorithm and runtime. Figure 2 presents the total time needed for modelling on each algorithm. The production of RF algorithm needs the longest time among other models, which is 1 minute 16 seconds, while the quickest production of algorithm is LR, which is 6 seconds.

Figure 3 shows the comparison of metrics performance results on each algorithm. DT becomes the model with the highest accuracy (92.2%) and NB becomes the model with lowest accuracy (89.5%). Table 2 show the result performance metric on the selected models. Based on the modelling, we can see the DT have the best model with accuracy 92.2%, AUC 0.91, precision 92.3, sensitivity 83.6, specificity 96.5 with the total time 7 second.

Receiver operating characteristics (ROC) comparison showed in Figure 4 below is the result of classification problem in determining thresholds in separate models. RUV curve shows the trade-off between sensitivity (TPR) and specificity (1-FPR). All in all, the created model gives closer curve to the top left corner, which means the algorithm model shows the good performance randomly classifies the points located along the diagonal (FPR = TPR). However, DT model shows better performance compared to other models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc</th>
<th>AUC</th>
<th>Prec</th>
<th>Sens</th>
<th>Spec</th>
<th>Classification error</th>
<th>Runtime (second)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>90.7</td>
<td>0.95</td>
<td>86.9</td>
<td>83.0</td>
<td>93.6</td>
<td>9.3</td>
<td>16 s</td>
</tr>
<tr>
<td>GBT</td>
<td>91.2</td>
<td>0.92</td>
<td>90.6</td>
<td>82.1</td>
<td>95.7</td>
<td>8.8</td>
<td>34 s</td>
</tr>
<tr>
<td>RF</td>
<td>91.9</td>
<td>0.93</td>
<td>91.4</td>
<td>83.6</td>
<td>96.1</td>
<td>8.1</td>
<td>1 min 16 s</td>
</tr>
<tr>
<td>Deep learning</td>
<td>90.0</td>
<td>0.92</td>
<td>88.4</td>
<td>80.7</td>
<td>94.6</td>
<td>10.0</td>
<td>11 s</td>
</tr>
<tr>
<td>FLM</td>
<td>90.7</td>
<td>0.92</td>
<td>92.5</td>
<td>78.6</td>
<td>96.8</td>
<td>9.3</td>
<td>9 s</td>
</tr>
<tr>
<td>LR</td>
<td>90.5</td>
<td>0.92</td>
<td>92.4</td>
<td>77.9</td>
<td>96.8</td>
<td>9.5</td>
<td>6 s</td>
</tr>
<tr>
<td>GLM</td>
<td>90.5</td>
<td>0.92</td>
<td>92.4</td>
<td>77.9</td>
<td>96.8</td>
<td>9.5</td>
<td>9 s</td>
</tr>
<tr>
<td>DT</td>
<td>92.2</td>
<td>0.91</td>
<td>92.3</td>
<td>83.6</td>
<td>96.5</td>
<td>9.5</td>
<td>7 s</td>
</tr>
<tr>
<td>NB</td>
<td>89.5</td>
<td>0.92</td>
<td>85.4</td>
<td>82.9</td>
<td>92.9</td>
<td>7.8</td>
<td>11 s</td>
</tr>
</tbody>
</table>
4.2. Features of importance

Features of Important is an index in a data set that shows the relative importance of features in predicting the target variable, namely Preeclampsia. To determine the accuracy of important features using weight of correlation modelling. The Figure 5 shows the results of the correlation weights starting from the largest.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>History of Hypertension</td>
<td>0.459</td>
</tr>
<tr>
<td>History of Diabetes</td>
<td>0.322</td>
</tr>
<tr>
<td>History of Preeclampsia</td>
<td>0.294</td>
</tr>
<tr>
<td>History of Caesarean</td>
<td>0.201</td>
</tr>
<tr>
<td>Profession</td>
<td>0.123</td>
</tr>
<tr>
<td>Parity</td>
<td>0.082</td>
</tr>
<tr>
<td>Gravidity</td>
<td>0.038</td>
</tr>
<tr>
<td>Pregnancy interval</td>
<td>0.024</td>
</tr>
<tr>
<td>History of Abortion</td>
<td>0.022</td>
</tr>
<tr>
<td>BMI</td>
<td>0.016</td>
</tr>
<tr>
<td>Hemoglobin</td>
<td>0.013</td>
</tr>
<tr>
<td>Education</td>
<td>0.012</td>
</tr>
<tr>
<td>ANC</td>
<td>0.002</td>
</tr>
<tr>
<td>Age</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Based on the table above, we can see the 3 (three) highest features causes preeclampsia is history of hypertension, history of diabetes mellitus, and history of preeclampsia. Hypertension can be caused by a variety of individual and environmental factors, including genetics, and continuous exposure to stress. Hypertension is a sign of preeclampsia [22]. Hypertension is closely related to the occurrence of preeclampsia. According to the results of the study, hypertension in the 1st trimester has a high risk of preeclampsia, especially in primigravida [23]. Hypertension suffered since before pregnancy has resulted in disruption or damage to important organs. With pregnancy, the body will gain weight, resulting in more severe disorders again with the onset of edema and proteinuria [24].
Pregnancy diabetes, chronic renal illness, and systemic lupus erythematosus were all found to be highly linked to severe preeclampsia [25]. Women with a history of pre-gestational diabetes have more likely to experience preeclampsia compared to women without pre-gestational diabetes [26]. Women who experience resistance insulin before pregnancy can cause marked vascular damage with chronic inflammation, facilitating atherogenic and prothrombotic processes will affect normal blood vessels and normal placentation [27]. Having a history of preeclampsia is a very high risk factor for the recurrence of preeclampsia [28]. Pregnant women with a history of previous preeclampsia will have an increased risk of future pregnancies due to preeclampsia is a disease that are at risk of recurrence [29]. Pregnant women with a history of preeclampsia have an increased risk of cardiovascular disease (CVD) and dementia in subsequent pregnancies. One of the diseases in the CVD system is hypertension which is a symptom of preeclampsia [30].

5. CONCLUSION

The proposed model has been proven to be the simulation and experiment which can be used to test and validate its implementation in pregnancy screening through correct feature selection which definitely yields contribution to the field of technology. The findings of this study possess strengths compared to previous studies. They are completed with the features of importance that has never been carried out in the previous studies. Furthermore, this study also has its gap which the results that cannot be generalized. This explains the complexity of technology-based health service that has been served in the health field. Moreover, this success evaluation of model development provides implication in the field of health to help health workers in carrying out pregnancy screening, the trust of expectant on the service quality given surely.

REFERENCES


BIOGRAPHIES OF AUTHORS

Raden Topan Aditya Rahman is a lecturer at the Polytechnic of Health Borneo Citra Medika, South Kalimantan, Indonesia. Currently he is completing his doctoral at Sultan Idris Education University, Malaysia. His research areas are Pregnancy, Preeclampsia, Artificial Intelligence, and Machine Learning. He teaches research methodology and statistics courses. He has published books with ISBNs that focus on research methodology and statistics, with the title Statistical Analysis of Health Research in 2015 and Research Methodology and Statistics in 2017. He has also received several research grants from the Ministry of Education of the Republic of Indonesia from 2018 to 2020. He has produced several studies that have been published in reputable international and nationally accredited journals. He can be contacted at email: topanaditya85@gmail.com.

Muhammad Modi Lakulu is an associate professor in the Faculty of Computing and Meta-Technology at Universiti Pendidikan Sultan Idris, Malaysia. From 2013-2019, he was the head of the computing department and from 2019-2021, he was Deputy Dean (Research and Innovation) and currently he is the Director of the Quality Management Centre at Sultan Idris Education University, Moreover, he received his Ph.D. degree in Computer Science (Knowledge Management) from the Universiti Putra Malaysia (UPM), in 2012, his M.Sc. in Software Engineering from UK in 2002 and his B.Sc. in Computer Science from Universiti Teknologi Malaysia, in 1998. His research focuses on educational technology, information system and artificial intelligence. His research works have been published in journal, books and conferences. He can be contacted at email: modi@meta.upsi.edu.my.

Ismail Yusuf Panessai successfully studied for his Diploma of Telecommunication Engineering at Universitas Hasanuddin, Indonesia. Bachelor of Industrial Engineering at UJ Jakarta, Indonesia (completed 2005). Master of Science in Information and Communication Technology at Department of Artificial Intelligence, Technical University of Malaysia Malacca (UTeM), Malaysia (completed in August 2010), and Ph.D. at Department of Artificial Intelligence in Universiti Malaya, Malaysia (completed in February 2013). He followed the professional engineer program (Insinyur, Ir.) at Universitas Andalas Indonesia and completed in 2021. The main field of research studied by the author includes the application of artificial intelligence (expert systems, genetic algorithms, artificial neural networks, and fuzzy logic) which are used to solve problems in operations research, transportation, control systems, and multi-objective optimization. He can contact at email: ismailyusuf.panessai@yahoo.com.

Proposed model to predict preeclampsia using machine .... (Raden Topan Aditya Rahman)
Esti Yuandari is a lecturer at the Faculty of Health, Sari Mulia University, South Kalimantan, Indonesia. Currently she is completing his master at Diponegoro University, Semarang, Indonesia. Her research areas are Public Health, Adolescence, and Health of Promotion. Her teaches health policy administration, research of methodology, and several other courses related to public health. He has published books with ISBNs with title Research Methodology and Statistic in 2017. She has also received several research grants from the Ministry of Education of the Republic of Indonesia with title Problems of early marriage in teenagers in 2019, and factors causing early marriage in 2020. She has produced several studies that have been published in reputable international and nationally accredited journals. She can be contacted at email: yuandareti@gmail.com.

Ika Mardiatul Ulfa is a lecturer at the Betang Asi Raya Midwifery Academy, Palangka Raya, Central Kalimantan. The most recent education he completed was a Masters in Public Health in the Field of Reproductive Health at Respati Indonesia University, Jakarta Indonesia. The field of science studied is midwifery and pregnancy. She has received a research grant from ministry of higher education regarding the evaluation of the use of KIA books for pregnant women in 2019 and she has published an ISBN book about midwifery care in pregnancy in 2016, 2020, and 2022. She can be contacted at email: ikamardiatululf@gmail.com.

Fitriani Ningsih is a lecturer at the Betang Asi Raya Midwifery Academy, Palangka Raya, Central Kalimantan. The most recent education he completed was a Masters in Public Health in the Field of Reproductive Health at Respati Indonesia University, Jakarta Indonesia. The field of science studied is midwifery in childbirth and newborns. She has received a research grant from ministry of higher education on the effectiveness of a combination of oxytocin massage and hypno breastfeeding on optimizing breast milk production in postpartum women in 2019 and the lactation management control model at Pahandut Health Center, Palangka Raya City in 2020 and has published an ISBN book on the lactation management control model in 2020. She can be contacted at email: feghanz@gmail.com.

Lensi Natalia Tambunan is a lecturer at the Betang Asi Raya Midwifery Academy, Palangka Raya, Central Kalimantan. The most recent education he completed was a Masters in Public Health in the Field of Reproductive Health at Respati Indonesia University, Jakarta Indonesia. The field of science studied is midwifery and pregnancy. She has received a research grant from ministry of higher education on the effectiveness of a combination of endorphin massage and rose aroma therapy on reducing pain in the first stage of labor in 2019 and has published an ISBN book on the effectiveness of a combination of endorphin massage and rose aroma therapy on reducing pain in the first stage of labor in 2020. She can be contacted at email: len,enci@gmail.com.