The 6th International Conference on Digital Arts, Media and Technology (DAMT) and
4th ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (NCON)

2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering | 978-1-6654-1569-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ECTIDAMTNCON51128.2021.9425764

Risk Assessment of Pregnancy-induced Hypertension Using a Machine Learning Approach

Sirinat Wanriko
Image Information and Intelligence Laboratory, Department of Computer Engineering, Faculty of Engineering, Mahidol University
Nakhon Pathom, Thailand
sirinat.wan@student.mahidol.ac.th

Narit Hnoohom*
Image Information and Intelligence Laboratory, Department of Computer Engineering, Faculty of Engineering, Mahidol University
Nakhon Pathom, Thailand
*Corresponding author:
narit.hno@mahidol.ac.th

Konlakorn Wongpatikaseree
Department of Computer Engineering, Faculty of Engineering, Mahidol University
Nakhon Pathom, Thailand
konlakorn.won@mahidol.ac.th

Anuchit Jitpattanakul
Intelligent and Nonlinear Dynamic Innovations Research Center, Department of Mathematics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
anuchit.j@sci.kmutnb.ac.th

Olarik Musigavong
Department of Obstetrics and Gynecology (Reproductive Endocrinology and Infertility) Chaophraya Abhaibhubejhr Hospital
Prachin Buri, Thailand
dr.olarik@gmail.com

***Abstract*—This research aimed to develop a predictive model of the risk assessment of pregnancy-induced hypertension using a machine learning approach. Pregnancy-induced hypertension is a complication that has a serious impact on pregnant women and fetuses. It is the world's top three cause of death among pregnant women [1]. Nowadays, the exact cause of pregnancy-induced hypertension is unknown and therefore cannot be prevented. Early detection and received treatment can reduce the severity and danger. A public dataset of Logan (2020) was used in this research [2]. The dataset was collected from a case-control study on the determinants of 83 pre-eclampsia and five eclampsia cases among 352 pregnant women delivering in county hospitals in Nairobi, Kenya. According to the dataset, 75 percent of the pregnant women were healthy. Only 25 percent of the pregnant women were pre-eclampsia and eclampsia. Thus, this would result in a problem of an imbalanced classification when one of the two classes had more data than the other class. As such, this problem was resolved with the Synthetic Minority Over-sampling Technique (SMOTE). Risk assessment of pregnancy-induced hypertension was performed on seven machine learning algorithms, which were logistic regression (LR), K-nearest neighbor (KNN), decision tree (DT), random forest (RF), multilayer perceptron neural network (MLP), support vector machines (SVM), and naive Bayes (NB). In the experimental results, RF had the highest accuracy at 89.62 percent compared to other machine learning algorithms.**

Keywords - Pregnancy-induced hypertension, machine learning, imbalanced classification, synthetic minority over-sampling technique

I. Introduction

Pregnancy-induced hypertension is a complication that has a serious impact on pregnant women and fetuses [2]. It mostly occurs after the twentieth week of pregnancy. In addition to death, there is a risk of serious complications; such as, premature placenta, pulmonary flood, abnormal blood clotting, temporary or permanent blindness, and bleeding in the brain. Nowadays, the exact cause of pregnancy-induced hypertension is unknown and therefore cannot be prevented. Early detection and received treatment can reduce the severity and danger. Data from the World Health Organization (WHO) in 2010 found that 99 percent of the pregnant women who died lived in the countryside and were impoverished [3]. Furthermore, in rural areas of Thailand, there is a shortage of doctors and medical personnel. Consequently, several agencies have established projects to manage the problem of distributing medical services to the countryside, but this problem still occurs. According to the statistics of Thailand in 2017, the Medical Council found that one doctor must take care of an average of 1,143 patients per year [4-5]. Thus, the ultimate objective of this research was to develop a predictive model of the risk assessment of pregnancy-induced hypertension using a machine learning approach. The purpose was to support doctors for early patient assessment, treatment planning, and recovery care.

The rest of the paper is arranged as follows. First, a brief introduction is given. The related work is explained in Section II. Section III contains an overview of the proposed method. Experiment results are included in Section IV. Finally, in Section V, the conclusions are discussed.

II. Related Work

There is currently some research, which uses machine learning to create a predictive model for the risk of pregnancy-induced hypertension; such as, eclampsia and pre-eclampsia. Tahir et al. [6] proposed a prediction of the risk of the pre-eclampsia level in pregnant women during the pregnancy process using the neural network (NN) and deep learning (DL) algorithms. The number of attributes was reduced from 17 to nine using particle swarm optimization. DL provided the most accuracy with 95.68 percent. Nikolaides et al. [7] developed an early-stage indicator for the risk of pre-eclampsia. The dataset was composed using 6,838 pregnant women cases in the UK involving 24 variables. The dataset only contained 116 cases of pregnant women with pre-eclampsia. A multi-slab neural network provided the most accurate result at 93.8 percent. Moreover, Tahir et al. [8] proposed the detection of pre-eclampsia by using a neural network compared with other algorithms. The pre-eclampsia dataset was taken from the Haji General Hospital Surabaya, Indonesia. The neural network algorithm and LOO validation provided the most accuracy

with 96.66 percent. Likewise, Leemaqz et al. [9] presented a tiered pre-eclampsia predictive model focusing on the convergence of multiple models. They analyzed that Bayes' theorem could be used to integrate multiple models. The integrated model provided the most accuracy, where 81 percent was accurately identified at 20 weeks of gestation.

In addition to the above research, machine learning has been used to create a predictive model for the risk of other diseases. Khalilia et al. [10] proposed the Random Forest (RF) for predicting the risk of eight chronic diseases by using the database of the nationwide inpatient sample from samples of hospitals in the United States. The ensemble learning approach was used to solve the problem of the imbalanced data. The results from testing the accuracy was 88.79 percent. In addition, Pattekari and Parveen [11] proposed a prediction system for heart disease. They used mining techniques consisting of DT, naive Bayes (NB), and NN. This system could intelligently answer complex questions in diagnosing heart disease and help medical practitioners. This further assisted in enhancing and reducing the treatment costs. Moreover, Akhil jabbar et al. [12] proposed the K-nearest neighbor (KNN) and genetic algorithms to analyze heart disease using six datasets from the UCI Repository (UCI Machine Learning Repository) and one dataset from various hospitals in Andhra Pradesh, India. To enhance the accuracy of the research, KNN was used to collect all cases and classify new patients using a genetic algorithm to calculate the similarities. The results showed that using both methods together gave a 95.73 percent accuracy. Nayeem et al. [13] proposed a multilayer perceptron neural network (MLP) to predict heart disease, liver disease, and lung cancer using a feed-forward backpropagation neural network algorithm and MLP to distinguish between infected and uninfected individuals. They used the MIT-BIH Arrhythmia dataset, and the results of the research showed that the accuracy of the heart disease predictions was 82 percent, liver disease was 82 percent, and lung cancer was 91 percent, respectively.

III. The Proposed Method

In this section, the process was the development of the risk prediction model of pregnancy-induced hypertension as shown in Figure 1. The proposed method consisted of five processes: (1) Pregnancy data, (2) imbalanced data, (3) data preparation, (4) feature selection, and (5) evaluate machine learning algorithms.

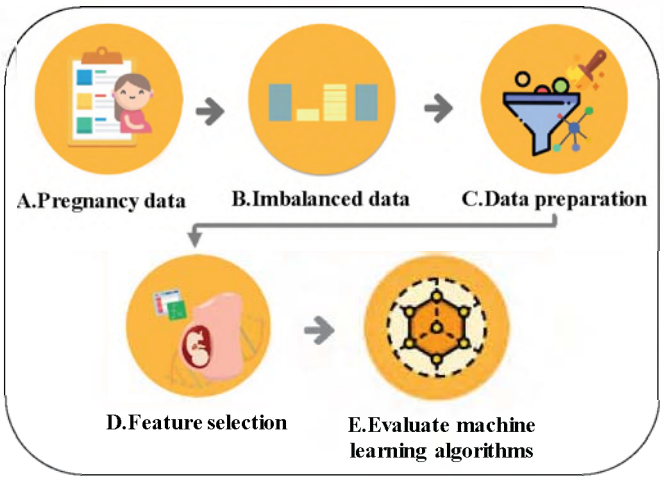
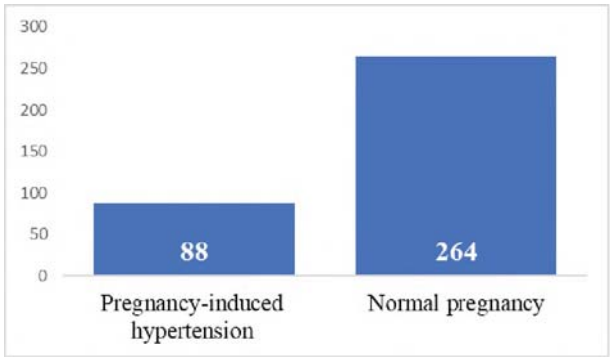


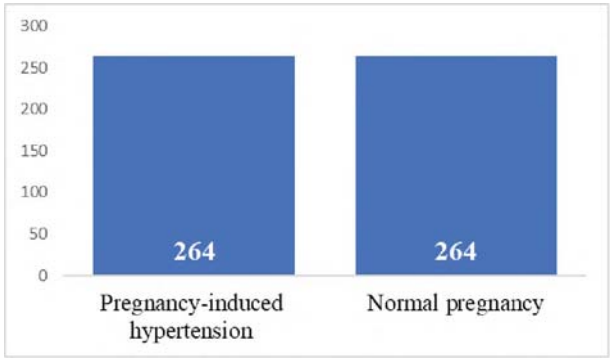
Fig. 1. Proposed method development processes.

A. Pregnancy data

The researcher explored the acquired public dataset relating to the risk of pregnancy-induced hypertension. The public dataset of Logan (2020) was used in this research [2]. The dataset was collected from a case-control study on the determinants of 83 pre-eclampsia and five eclampsia cases among 352 pregnant women delivering in county hospitals in Nairobi, Kenya. The researcher selected 17 attributes, which were matched to the attributes of the maternal and child health handbook in Thailand, and this was used to record the mother and child health data [14]. The 17 attributes comprised maternal age, age of the first pregnancy, diabetes family history, diabetes personal history, hypertension family history, hypertension personal history, first antenatal care visit, number of antenatal care visits, distance of pregnancy, cesarean section delivery, multifetal pregnancy, gravidity, parity, province, residence, alcohol use, and tobacco use. The researcher then converted the data for appropriateness in the imbalanced data process.



(a) *Imbalanced data.*



(b) *Balanced data.*

Fig. 2. A comparison of the imbalanced data and balanced data.

B. Imbalanced data

In the pregnancy data [13], 75 percent of pregnant women were healthy and only 25 percent were pre-eclampsia and eclampsia. The problem of the imbalanced classification arose when one of the two classes had more data than the other class, and this data affected the classification of the minority class. Typical classification was effective when each data class had a similar number, and the prediction was inclined to the majority class.

To compare the performance of the imbalanced data and balanced data approaches, the researcher performed the Synthetic Minority Over-sampling Technique (SMOTE) algorithm on the dataset [15]. The SMOTE works by utilizing

the KNN algorithm to construct synthetic data. The SMOTE began by selecting random data from the minority class after which the KNNs were created from the data. The random data and the randomly chosen KNN were then merged to construct synthetic data. The process was replicated until the minority class's proportion equaled that of the majority class.

The researcher plotted the class distribution to show the imbalanced data in the selected dataset as shown in Figure 2(a). Figure 2(b) shows the balanced data was handled with the SMOTE.

C. Data preparation

In the data preparation process, the researcher used data transformation in order to convert the data to better reveal the structure of the classification problem for the risk of pregnancy-induced hypertension. The Python program and scikit-learn library were utilized to develop the predictive model in this research. The 17 attributes were a redistributed or rescaled dataset using the three preprocessing methods, which consisted of the MinMaxScaler, StandardScaler, and Normalizer.

D. Feature selection

Feature selection for machine learning is the method of choosing features that are relevant to each other to reduce the number of features to only the feature necessary in order to develop a predictive model. Principal component analysis (PCA) is a data reduction technique that is commonly used in combination with linear algebra to minimize the dimensionality of a dataset by compressing a dataset of attributes while preserving most of the information in the dataset. PCA was used to extract the three principal components on the 17 attributes, imbalanced data, and balanced data.

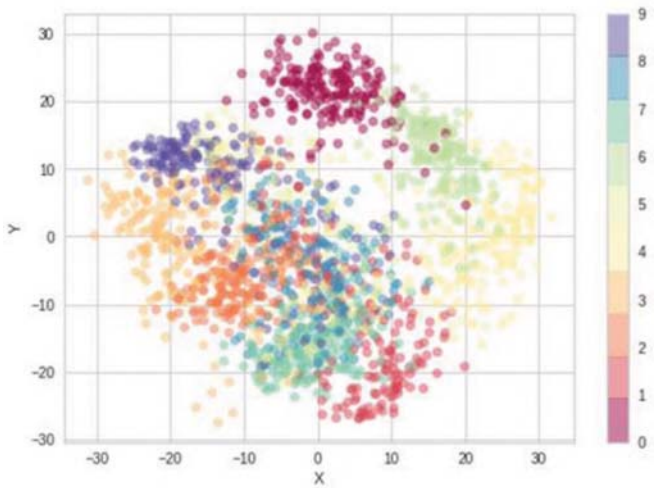


Fig. 3. Post-processing data with PCA.

E. Evaluate machine learning algorithms

To evaluate machine learning algorithms on the dataset, the predictive model was performed on seven machine learning algorithms, which were LR, KNN, DT, RF, MLP, SVM, and NB.

To estimate the performance of the machine learning algorithm on the unseen data, the researcher first split the dataset, both the imbalanced data and balanced data, into two parts: training and testing. For this case study, the researcher took a 77: 33 ratio keeping 77 percent of the dataset for training and the remaining 33 percent for testing. The researcher trained a machine learning algorithm on the first part then evaluated the predictions on the test set against the expected results.

IV. Experimental Results

To obtain the predictive model of the risk assessment of pregnancy-induced hypertension, there were three main steps. Firstly, the three preprocessing methods were performed on the imbalanced data and balanced data using scikit-learn with the MinMaxScaler, StandardScaler and Normalizer. Next, PCA was used to extract the three principal components on the imbalanced data and balanced data. Finally, the three principal components were performed on seven machine learning algorithms, which were logistic regression (LR), K-nearest neighbor (KNN), decision tree (DT), random forest (RF), multilayer perceptron neural network (MLP), support vector machines (SVM), and naive Bayes (NB). The evaluation results are shown in Tables I-VI.

TABLE I. Results Of The Imbalanced Data With StandardScaler.

Machine Learning	Feature Selection
	PCA
Logistic Regression	69.01
K-Nearest Neighbor	64.79
Decision Tree	61.97
Random Forests	70.42
Multilayer Perceptron	71.83
Naive Bayes	28.17
Support Vector Machines	70.42

The results can be seen in Table I. The highest accuracy rate was 71.83 percent, shown by the underlined letter, which belonged to the MLP algorithm.

TABLE II. Results Of The Imbalanced Data With MinMaxScaler.

Machine Learning	Feature Selection
	PCA
Logistic Regression	71.83
K-Nearest Neighbor	67.61
Decision Tree	60.56
Random Forests	71.83
Multilayer Perceptron	71.83
Naive Bayes	28.17
Support Vector Machines	71.83

As shown in Table II, the highest accuracy rate was 71.83 percent, shown by the underlined letter, which belonged to the LR, RF, MLP and SVM algorithms.

The results can be seen in Table III. The highest accuracy rate was 73.24 percent, shown by the underlined letter, which belonged to the KNN algorithm.

As shown in Table IV, the highest accuracy rate was 89.62 percent, shown by the underlined letter, which belonged to the RF algorithm.

TABLE III. RESULTS OF THE IMBALANCED DATA WITH NORMALIZER.

Machine Learning	Feature Selection
	PCA
Logistic Regression	71.83
K-Nearest Neighbor	73.24
Decision Tree	59.15
Random Forests	67.61
Multilayer Perceptron	71.83
Naive Bayes	28.17
Support Vector Machines	71.83

TABLE IV. RESULTS OF THE BALANCED DATA WITH STANDARDSCALER.

Machine Learning	Feature Selection
	PCA
Logistic Regression	69.81
K-Nearest Neighbor	72.64
Decision Tree	82.08
Random Forests	89.62
Multilayer Perceptron	45.28
Naive Bayes	48.11
Support Vector Machines	78.30

TABLE V. RESULTS OF THE BALANCED DATA WITH MINMAXSCALER.

Machine Learning	Feature Selection
	PCA
Logistic Regression	66.04
K-Nearest Neighbor	68.87
Decision Tree	79.25
Random Forests	85.85
Multilayer Perceptron	45.28
Naive Bayes	50.00
Support Vector Machines	59.43

The results can be seen in Table V. The highest accuracy rate was 85.85 percent, shown by the underlined letter, which belonged to the RF algorithm.

TABLE VI. RESULTS OF THE BALANCED DATA WITH NORMALIZER.

Machine Learning	Feature Selection
	PCA
Logistic Regression	47.17
K-Nearest Neighbor	69.81
Decision Tree	73.58
Random Forests	85.85
Multilayer Perceptron	45.28
Naive Bayes	48.11
Support Vector Machines	45.28

As shown in Table VI, the highest accuracy rate was 85.85 percent, shown by the underlined letter, which belonged to the RF algorithm.

According to the experiments conducted with the dataset, the RF algorithm yielded the best result based on the balanced data. The highest accuracy rate was 89.62 percent.

V. CONCLUSION

This paper presented the risk assessment of pregnancy-induced hypertension using a machine learning approach. A public dataset of Logan (2020) was used in this research. In dealing with imbalanced data, the researcher performed the SMOTE algorithm on the dataset. In the data preparation process, the researcher used data transformation, which consisted of MinMaxScaler, StandardScaler and Normalizer, in order to convert data to better reveal the structure of the prediction problem. PCA was used to extract the three principal components on the imbalanced data and balanced data. In the experimental results, the RF algorithm achieved the best performance when compared to other machine learning algorithms. The prediction model yielded an accuracy rate of up to 89.62 percent, which was based on the balanced data. In future work, the researcher should apply the predictive model on the pregnancy-induced hypertension dataset from the Chaophraya Abhaibhubejhr Hospital, Prachin Buri.

ACKNOWLEDGMENT

This research was supported by the Department of Computer Engineering, Faculty of Engineering, Mahidol University. This research was also supported the National Research Council of Thailand for the project "Mom's buddy: AI chatbot for pregnancy health information". Furthermore, this research was supported by Chaophraya Abhaibhubejhr Hospital on medical knowledge.

REFERENCES

[1] Thanomrat Prasith-thimet, and Kasem Wetsutthanon, "Causes of maternal deaths in Regional Health 4 during Fiscal Year 2014-2016," Journal of Health Science, vol.26, pp.5, 2017.

[2] Logan Gorbee, "Replication Data for: Determinants of preeclampsia and eclampsia among women delivering in county hospitals in Nairobi, Kenya", Harvard Dataverse, version 1.0, 2020. [online] Available: http://www.doi.org/10.7910/DVN/BYFL3J. [Accessed: Nov.18, 2020].

[3] WHO, UNICEF, UNFPA, "The World Bank Trends in maternal mortality: 1990 to 2010", 2010. [online] Available: https://www.who.int/reproductivehealth/publications/monitoring/9789241503631/en/. [Accessed: Jan.1, 2019].

[4] The Medical Council of Thailand, "Medical statistics", 2017. [online] Available: http://www.tmc.or.th/pdf/01_stat _med2560.pdf. [Accessed: Jan.1, 2019].

[5] Official Statistics Registration Systems, "The population in each age, Nationwide", 2017. [online] Available: http://stat.dopa.go.th/stat/statnew/upstat_age_disp.php. [Accessed: Jan.1, 2019].

[6] Muhlis Tahir, Tessy Badriyah, Iwan Syarif, "Classification Algorithms of Maternal Risk Detection For Preeclampsia With Hypertension During Pregnancy Using Particle Swarm Optimization," EMITTER International Journal of Engineering Technology, vol.6, pp.236-250, 2018.

[7] Costas K. Neocleous, Panagiotis Anastasopoulos, Kypros H. Nikolaides, Christos N. Schizas, Kleanthis C. Neokleous, "Neural networks to estimate the risk for preeclampsia occurrence," in Proc. IEEE International Joint Conference on Neural Networks, 2009, pp.2221-2224.

[8] Muhlis Tahir, Tessy Badriyah, and Iwan Syarif, "Neural Networks Algorithm to Inquire Previous Preeclampsia Factors in Women with Chronic Hypertension During Pregnancy in Childbirth Process," in Proc. IEEE International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), 2018, pp.51-55.

[9] S.Y. Leemaqz, G.A. Dekker and C.T. Roberts "Tiered Prediction System for Preeclampsia: an integrative application of multiple models," in Proc. 20th International Congress on Modelling and Simulation (MODSIM), 2013, pp.2041-2045.

[10] Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu, "Predicting disease risks from highly imbalanced data using random forest," BMC Medical Informatics and Decision Making, vol.11, pp.1-13, 2011.

[11] Shadab Adam Pattekari and Asma Parveen, "Prediction system for heart disease using naive Bayes," Biomedical Research, vol.29, pp.2646-2649, 2018.

[12] M.Akhil jabbar, B.L Deekshatulu, and Priti Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," International Conference on Computational Intelligence: Modeling Techniques and Applications, CIMTA, Kalyani, Kolkata, India, September 27, 2013, pp.85-94.

[13] Md. Osman Goni Nayeem, Maung Ning Wan, and Md. Kamrul Hasan, "Prediction of Disease Level Using Multilayer Perceptron of Artificial Neural Network for Patient Monitoring," International Journal of Soft Computing and Engineering (IJSCE), vol.5, pp.17-23, September 2015.

[14] Department of Health, and National health security office (nhso) thailand, "mother and child health records or pink notebooks of general hospitals in Thailand," 2018. [online] Available: http://www.oic.go.th/FILEWEB/CABINFOCENTER17/DRAWER002/GENERAL/DATA0001/00001375.PDF. [Accessed: Jan.1, 2019].

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," journal Of Artificial Intelligence Research, vol.16, pp. 321-357, 2002.