Impact of Feature Selection

by Setio B

Submission date: 06-Aug-2023 07:04AM (UTC+0700)
Submission ID: 2141800177
File name: Data_Augmentation_for_Pregnancy_Risk_Detection_in_Indonesia.pdf (1.54M)
Word count: 5143
Character count: 27999

International Journal on
Advanced Science
Engineering
Information Technology

Vol.12 (2022) No. 6
ISSN: 2088-5334

Impact of Feature Selection and Data Augmentation for Pregnancy Risk Detection in Indonesia

Muhammad Irfan [a], Setio Basuki [a,*], Yufis Azhar [a]
[a] Informatics Engineering, Universitas Muhammadiyah Malang, Indonesia
Corresponding author: [*]setio_basuki@umm.ac.id

***Abstract*— This paper aims to develop an automatic system for pregnancy risk detection in Indonesia. The system requires a sophisticated approach to achieve the required performance as a sensitive field. Existing works are developed using small-sized datasets and limited classification features. Moreover, all features treated equally make the detection results hard to interpret which features contribute more. To address these issues, we propose to combine more complex features, data augmentation methods, and feature selection techniques. We prefer to use all 118 pregnancy indicators and 400 instances from Puskesmas as an original dataset. Next, the new datasets are used to build two data augmentation methods, i.e., GMM and CTGAN. Each data augmentation method generates 2,000 new synthetic instances. Following this, five machine learning methods combined with three feature selection approaches, i.e., RFE, Random Forest, and Chi-Square, are implemented in all datasets. Through experiments, we observed that feature selection techniques play an essential role in improving classification accuracies. While the GMM-based augmentation demonstrated performance improvement, the CTGAN-based synthetic dataset depicted low performances. The best accuracy on all experiment settings reached 95%. By using Random Forest combined with RFE on a GMM-based dataset, the highest accuracy was achieved using only five features. Another notable result is that both XGBoost and Decision Tree reached the same 95% accuracy on the GMM-based dataset on only nine features. The overall results show that appropriate data augmentation and feature selection are a matter for achieving better performance in this research.**

***Keywords*— Ctgan; data augmentation; feature selection; pregnancy risk detection.**

Manuscript received 26 Aug. 2021; revised 25 Mar. 2022; accepted 20 May 2022. Date of publication 31 Dec. 2022.

I. Introduction

The Maternal Mortality Rate (MMR) in Indonesia is considered high. The 2016 Indonesian Health survey stated that the high mortality rate is closely related to pregnancy and childbirth [1]. Inter-Census Population Survey defines the MMR as the percentage for every 100,000 live births. The survey reveals that the MMR decreases to 305 maternal deaths per 100,000 live births, but it is still considered above the pre-stated limit. The low awareness of pregnant women of potential pregnancy disorders contributes to the increase of MMR. This situation leads to the Indonesian Government promoting massive actions to provide better health services.

The Indonesian Ministry of Health, through Public Health Centre (*Puskesmas*), delivers the first level and individual health support through promotive and preventive actions. The purpose of the support is to reduce the MMR as public health service indicator in case of pregnancy. To address this issue, *Puskesmas* uses Pregnancy Control Card (PCC) at the beginning of pregnancy [2]. The PCC contains 118 attributes to represent the pregnancy condition. The PCC consists of four categories, i.e., (a) the history of pregnancy, (b) the family planning and childbirth, (c) the history of current pregnancy, (d) the physical and obstetric status, and (d) the laboratory check. *Puskesmas* uses these attributes to detect the risk of pregnancy.

The automatic system of pregnancy risk detection has gained much attention. Moreover, many studies employ Artificial Intelligence and Machine Learning (ML) in supporting patient care during pregnancy [3]. Presently, there are several works on automatic pregnancy risk detection. Akbulut, Ertugrul, and Topcu [4] developed an assistive system of e-Health application that achieved the best performance 89.5% using the Decision Forest model. The next telemedicine platform for the prenatal case was proposed by Bautista, Quiwa, and Reyes [5] for the Philippines case study. This research reached the best performance of test scores by 90%. Following this, Davidson and Boland [6] discussed the gap and possible future Artificial Intelligence applications for maternal health. The next research to build an application for pregnancy risk monitoring is by Sarhaddi et al. [7] which developed the Internet of Things (IoT) platform for maternal health monitoring.

Research by Moreira et al. [8] used the Random Forest algorithm to observe the hypertensive disorder during pregnancy which utilized the obstetricians' experience data based on 25 pregnant women. The categorization of the hypertensive diseases of pregnancy is under the standardized guidelines of the 10th Revision (ICD-10) of the International Statistical Classification of Diseases and Related Health Problems. The categories are (a) the pre-existing hypertension of pregnancy, childbirth, and chronic hypertension (CH), (b) the preeclampsia on chronic hypertension (PS); (c) gestational hypertension (GH); and (d) preeclampsia (PE). Similarly, another research by Moreira et al [9] conducted research for fetal birth weight estimation on High-Risk Pregnancy by employing several ML algorithms. Research by Tahir, Badriyah, and Syarif [10] implemented the Neural Network algorithm for preeclampsia detection by applying data from 2016 to 2017 consisting of 17 features, 239 medical records from Surabaya Hajj Hospital, Indonesia.

In research by Chu et al. [11], predictions of adverse events in pregnant women were performed by using several classical ML algorithms, including Support Vector Machine (SVM), Random Forest (RF), AdaBoost, Decision Tree (DT), k-Nearest Neighbor (kNN), Multilayer Perceptron (MLP), and Naïve Bayes (NB). Following this, Purwanti, Preswari, and Ernawati [12] proposed to apply Artificial Neural Network to detect preeclampsia in pregnant women using 11 risk factors as classification features. Besides, the application of such classical ML technique was performed in [13] to predict the presence of preeclampsia by employing the Logistic Regression (LR), DT, Artificial Neural Network (ANN), RF, SVM, and Ensemble Algorithm which applied on the National Health Insurance Dataset in Indonesia. Another study was conducted to predict the model of hypertensive disorder with the SVM algorithm [14]. Research [15] used several ML algorithms, i.e., Regularized LR, DT, RF, XGBoost, and MLP for the stillbirth prediction.

Our literature review reveals that there are several limitations in existing works.

- First, in response to high observed MMR, it is surprising that there is few available research on automatic detection of pregnancy risk in Indonesia.
- Second, the existing works addressing this issue use a limited size **of** pregnancy datasets and propose only a few classification features.
- Third, since all classification features are treated equally, there is no information about the most significant features which affect the prediction results. To address these issues, this paper proposes an automatic system of pregnancy risk detection which implements three key methods, namely ML, data augmentation, and Feature Selection (FS). Data augmentation is a method to increase the number of instances in the dataset by creating new synthetic data or adding a modified copy of existing instances. While adding more data through data augmentation makes better classification models, the FS enables the models to deliver the most influential features during experiments.

This paper uses 400 instances, each consisting of 118 attributes (features) of pregnancy records obtained from Puskesmas Cipto Mulyo, from 2016 to June 2017, in Malang City, Indonesia [2]. We employ three classes, i.e., the Very High-Risk Pregnancy (VHRP), the High-Risk Pregnancy (HRP), and the Low-Risk Pregnancy (LRP). Note that, the risk level is assigned by using the following rule: the score >= 12 for VHRP, the score between 6 and 10 for HRP, and the score equal to 2 for LRP which is manually assigned by the staff of the *Puskesmas* using the PCC as defined in [2].

To be more specific, this paper combines ML, data augmentation, and FS. To be more precise, this paper addresses several research challenges:

- How to implement two types of data augmentation methods, i.e., Conditional Generative Adversarial Network (CTGAN) and Gaussian Mixture Model (GMM), to generate a new synthetic dataset?
- How to adopt three types of FS methods, i.e., Recursive Feature Elimination (RFE), Random Forest (RF), and Chi-Square to select the most influential features?
- How to apply five ML methods, i.e., XGBoost, Random Forest, Naïve Bayes, Decision Tree, and Logistic Regression for pregnancy risks detection?
- Given the datasets (original and augmented) and FS methods, what is the optimal number of features for achieving the highest classification accuracy?
- Moreover, what features are considered as most influential in achieving better prediction accuracies on each scenario?

This paper is organized as follows. Section II demonstrates how our proposed system will be developed. This section consists of three stages, i.e., data acquisition and preprocessing feature selection and data augmentation, and risks classifications. Following this, section III shows the experiment results containing the observed behavior of each prediction scenario. Finally, the conclusion is presented in section IV.

II. Material and Method

The proposed system of pregnancy risks detection is developed through the three stages, i.e., data acquisition and preprocessing, data augmentation, and pregnancy risks classification. In the first stage, we present obtained data source and its pregnancy indicators. Following this, several data preprocessing techniques are used to adjust the data for classification. Next, this paper proposes to construct a synthetic dataset using data augmentation methods based on CTGAN and GMM. These two new datasets combined with the original dataset are used to conduct the classification. In the last stage, we compare five ML algorithms on all datasets. Furthermore, three FS methods are implemented to obtain the most influential features from each scenario. Fig 1 depicts how our proposed system of pregnancy risk prediction is developed.

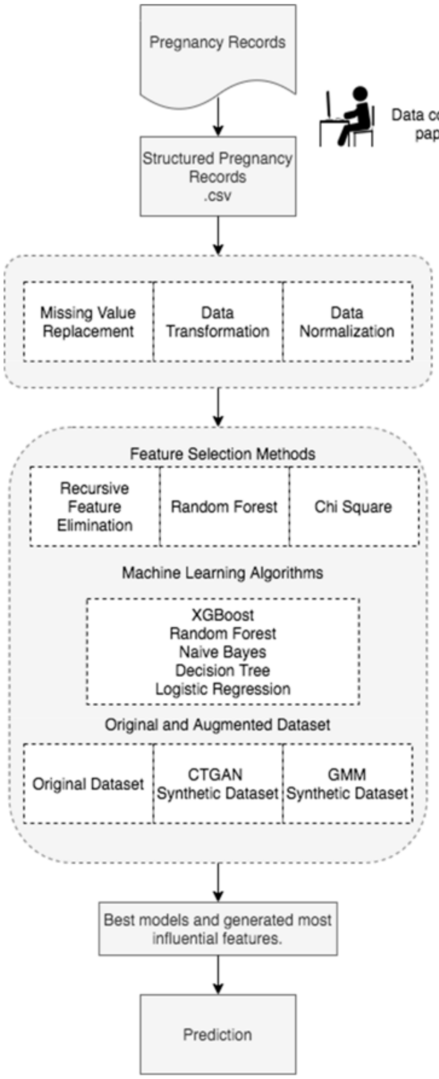
Pregnancy Records
Data convertion from papers to .csv
Structured Pregnancy Records .csv
Missing Value Replacement
Data Transformation
Data Normalization
Feature Selection Methods
Recursive Feature Elimination
Random Forest
Chi Square
Machine Learning Algorithms
XGBoost
Random Forest
Naive Bayes
Decision Tree
Logistic Regression
Original and Augmented Dataset
Original Dataset
CTGAN Synthetic Dataset
GMM Synthetic Dataset
Best models and generated most influential features.
Prediction

Fig. 1 The proposed pregnancy risk detection system.

A. Data Acquisition and Preprocessing

The original dataset which is extracted from PCC, consists of 400 instances and 118 attributes. The dataset was obtained from *Puskesmas* Cipto Mulyo, starting from 2016 to June 2017, in Malang, Indonesia. Since the pregnancy records were available in physical paper files, we manually converted the data to CSV format. The records are already split into four categories. The first category is Pregnancy History (Table I), containing indicators of previous pregnancy. The second category is related to Current Pregnancy indicators (Table II), covering both the pregnant women's condition and the family illness history. The third category represents general aspects that cover Physical Indicators (Table III). The last category (Table IV) is obtained through the Laboratory Check comprising several examination indicators. The distribution of original data is shown in Fig 2. containing LRP containing 149 instances, HRP containing 183 instances, and VHRP containing 68 instances.

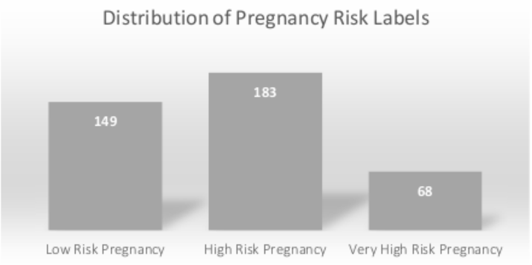

Fig. 2 Original distribution of pregnancy risk dataset

TABLE I
PREGNANCY HISTORY INDICATORS

History Criteria	Details of History
Pregnancy	Pregnancy frequency complication
The childbirth method	SC, Tool, Breech, Normal, IUFD, I/P, Abortus
The childbirth sites	Homebirth, other, BPS, Public Health Center, Hospital
The childbirth complication	HPP, Infection, Complicated delivery
Medical aids	Obgyn/doctor, midwife, other aids
The baby condition and weight	Dead, healthy, P/L, ill, BBL(gr)
The condition of current child	dead, alive
Birth control	No, Yes

TABLE II
CURRENT PREGNANCY INDICATORS

History Criteria	Details of History
Current Pregnancy History	Husband Sexual Couple, Wife Sexual Couple, Fluorbus Albus, Fluor albus, History of Immunization, Bleeding, Appetite, Edema, Fetal Motion, Abdominal Pain, Dizziness, Nausea / Vomiting, Long Menstruation, Menstrual Cycles,
Mother's illness history	STDs, decreased BB, Long cough, Heat, Prolonged diarrhea, Hypertension, Heart, Malaria, Kidney, Psychosis, Liver, Epilepsy, DM, Lung.
Father's illness history	Tumors, Hepatitis, STDs, HIV, diarrhea, Long cough, Tattoos, DM, Piercings
Family illness history	Psychosis, Hypertension, Lungs, Gemelli, DM, Heart
Mother habit	stomach massage, Smoking, herbal medicine, liquor, sedatives, narcotics

TABLE III
THE PHYSICAL INDICATORS

Examination Criteria	Details of Indicators
General	respiration, pulse, temperature, diastolic blood pressure, awareness, yellow, systolic blood pressure, height, body shape, weight before pregnancy, LILA, pale, weight
Physical	Reflexes, Limbs, Heart, Abdominal Mass, Surgical Injuries, Breast, Lung / Heart, Glandular Disorders, Teeth, Eye, Skin, Mouth,
Midwifery	Inspekulo, Heartbeat, Decreased Kep, Fetal Position> 36 Weeks, Fetal Position <36 weeks, uterus Shape, UK, TFU

TABLE IV
THE LABORATORY CHECK INDICATORS

Examination Criteria	Details of Indicators
Laboratory Examination	Reduction Urine, other Indications, Hemoglobin (gr), Urine Albumin,

Next, we performed data preprocessing to arrange the converted pregnancy for classification. Preprocessing plays a crucial role since the real-world data often lacks consistency, is incomplete, or contains many errors. For this purpose, this paper uses three data preprocessing as follows:

1) Missing Value Replacement. The first method is Missing Value Replacement, and this method replaces missing values with centrality tendency principles, such as using the attributes' mean value. Table V and Table VI below show how the missing values are replaced.

TABLE V
ORIGINAL DATA

Height	systolic blood pressure	diastolic blood pressure	Blood Type
150	100	70	B
146	90	70	O
?	100	60	?
151	110	?	O
153	120	80	A
160	?	70	AB

TABLE VI
MISSING VALUES REPLACEMENT RESULTS

Height	systolic blood pressure	diastolic blood pressure	Blood Type
150	100	70	B
146	90	70	O
152	100	60	**O**
151	110	**70**	O
153	120	80	A
160	**104**	70	AB

2) Data Transformation. The second method is *Data Transformation* to transform the nominal categorical data into numeric form. The transformations are shown in Table VII and Table VIII.

TABLE VII
ORIGINAL DATA

Nausea	Dizzy	Abdominal Pain	Fetal Motion
sometimes	no	yes	active
sometimes	sometimes	no	non-active
sometimes	sometimes	no	non-active
sometimes	no	no	active
always	always	no	non-active
sometimes	no	no	non-active
no	sometimes	yes	rarely
sometimes	no	no	non-active

TABLE VIII
AFTER DATA TRANSFORMATION

Nausea	Dizzy	Abdominal Pain	Fetal Motion
1	0	1	2
1	1	0	0
1	1	0	0
1	0	0	2
2	2	0	0
1	0	0	0
0	1	1	1
1	0	0	0

3) Data Normalization. The third technique is *Data Normalization* to scale the numeric attributes based on the *min-max* formula. Table IX and Table X below show the data normalization process.

TABLE IX
ORIGINAL DATA

Nausea	Dizzy	Mother Age
1	0	30
2	2	33
1	0	29
1	1	20
0	0	22

TABLE X
DATA NORMALIZATION RESULTS

Nausea	Dizzy	Mother Age
0.5	0	0.769
1	1	1
0.5	0	0.692
0	0.5	0
0	0	0.154

B. Feature Selection and Data Augmentation

1) Feature Selection (FS). This paper implemented three FS methods, i.e., Recursive Feature Elimination (RFE), Random Forest, and Chi-Square. The RFE is categorized as a wrapper-based FS method, which means the RFE wraps ML algorithms (estimators) as a core method to select the features. More specifically, given the external estimator, the RFE has smaller sets of features recursively. The estimator will be trained on the early set of features. After this, the least important features are dropped from the feature set. These steps are repeated until a pre-defined number of features. Here, we use the feature importance of Random Forest as the FS method. The feature importance is calculated based on the concept of Mean Decrease Impurity (MDI). The MDI is also defined as a total decrease in node impurity [16]. This value is weighted by the probability of reaching that node (which is approximated by the proportion of samples reaching that node) averaged over all trees of the ensemble. Another FS method is Chi-Square which is commonly used to test the independence of two events. Let's consider the scenario of determining the relationship between the independent category feature (predictor) and the dependent category feature (response). Chi-Square measures how expected count E and observed count O deviate from each other. For feature selection, we need to obtain features that are highly dependent on the response.

2) Data Augmentation. Data augmentation is used to generate synthetic datasets. There are two methods used in this paper, i.e., the Gaussian Mixture Model (GMM) and Conditional Generative Adversarial Network (CTGAN). The GMM is a very popular unsupervised learning technique. This technique is popular to be used to form new synthetic data on small-sized datasets [17], [18], [19], [20]–[24]. Moreover, the GMM has been widely successfully applied in the prediction system on medical-related topics such as predicting COVID-19 cases [22], liver cancer detection [25], pancreatic cancer detection [26], medical image segmentation [27], and texture characterization of brain DTI image [28]. GMM consist of *N* Gaussian components combined using a linear combination to form a distribution *P (x)*, as follows (Eq. 1):

$$(x) = \sum_{i=1}^{N} \alpha_i N(x;\, \mu_i, \Sigma_i) \quad (1)$$

where x is d-dimensional random vector, the i-th gaussian component has mean $\mu_i \in R^d$ and covarian matrix $\Sigma_i \in R^{d\ x\ d}$ and α_i is the weight for the i-th gaussian component that satisfying condition $\sum_{i=1}^{N} \alpha_i = 1$ and $\forall_i, \alpha_i \geq 0$. Basically, learning GMM is to find the best parameters for μ_i, Σ_i dan α_i [29].

The CTGAN was proposed by [30] to generate a synthetic tabular dataset. The motivation of the CTGAN is to address the nature of tabular data, which contains continuous dan discrete columns. Xu argues that existing works are failed to model this characteristic of data properly.

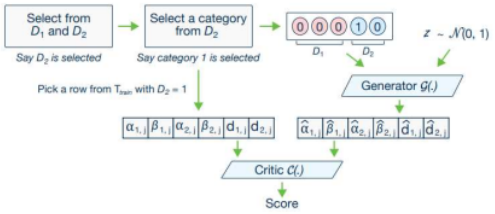

Fig. 3 CTGAN Model [30]

CTGAN, as depicted in Fig 3, works by applying conditional generator and training-by-sampling approaches. These two techniques are used to handle the imbalanced discrete column. CTGAN introduces the vector *cond* as the way to indicate the condition $(D_{i*} = k^*)$. Recall that all the discrete columns $D_1, \ldots, D_{Nd}$ end up as one-hot vectors $d_1, \ldots, d_{Nd}$ such that the i_{th} one-hot vector is di = $[d_i^{(k)}]$, for $k = 1, \ldots, |D_i|$. Let mi = $[m_i^{(k)}]$, for $k = 1, \ldots, |D_i|$ be the i_{th} mask vector associated to the i_{th} one-hot vector d_i. Hence, the condition can be expressed in terms of these mask vectors as (Eq. 2):

$$m_i^{(k)} = \begin{cases} 1\ if\ i = i^*\ and\ k = k^* \\ 0\ otherwise \end{cases} \quad (2)$$

The vector *cond* is defined as *cond* = m1 ⊕ . . . ⊕ m $_{Nd}$. The two discrete columns, D1 = {1, 2, 3} and D2 = {1, 2}, the condition (D2 = 1) is expressed by the mask vectors m1 = [0, 0, 0] and m2 = [1, 0]; thus, vector *cond* = [0, 0, 0, 1, 0].

C. *Classification Scenarios*

The pregnancy risk classification is performed using three datasets. The first dataset is an original version obtained after preprocessing, as shown in Fig 2. The next two datasets are augmented datasets based on GMM and CTGAN, respectively. Both GMM and CTGAN are used to generate 2,000 new synthetic instances. The original dataset is split into training and testing for building ML models with a 9:1 proportion. The proposed data augmentation techniques are applied only to the training set. Considering this scenario, the first dataset contains 360 training instances, the second and third datasets contain 2,360 training instances (360 original training instances + 2,000 synthetic instances). Based on the proposed prediction system in Fig. 1, this paper performs 45 classification experiments to combine three datasets, three FS methods, and five ML algorithms. The detailed proportion of each dataset is shown in Table XI.

We implemented five ML algorithms on each dataset, i.e., XGBoost, Random Forest, Naïve Bayes, Decision Tree, and Logistic Regression. Following this, we applied three Feature FS methods using RFE, RF, and Chi-Square. On each dataset and classifier, the FSs use incremental selection starting from one feature, two features until all 118 features. The models' performances are measured through the accuracies of the same testing instances on each selection. Through this strategy, we will answer the question of how many features need to achieve the best result on given datasets and classifiers.

III. Result and Discussion

As mentioned before, the data augmentation methods have been applied only to the training data, and the evaluations have been applied on the same testing instances, as shown in Table XI. This scenario is chosen to guarantee the evaluation fairness between the combination of the ML methods and FS methods. To be more specific, this paper constructed three datasets. The first **dataset** is the original dataset, the **second dataset** is the combination of the original dataset with the augmented dataset using GMM, and the **third dataset** is the combination of the original dataset with the augmented dataset using CTGAN.

Here, the main goal is to find the combination of methods that achieved the highest accuracies but using minimum features. By doing this, we are able to justify the most influence classification features to predict the risk of pregnancy. To give the analysis, this paper used the term *earliest number of features (ENoF)* as the fewest number of features that achieved the highest classification results (accuracies) in each scenario. While Table XII shows the ENoF of classification results, the whole classification accuracies are shown in Fig 4.

TABLE XI
Laboratory Check Indicators

Distribution of Dataset	Original Dataset	The Second Dataset	The Third Dataset
Training	360	360 + 2000	360 + 2000
Testing	40	40	40

The results of classification experiments are divided into three parts. The first part is the experiments results on the original dataset without data augmentation. It is seen that the best classification accuracy achieved by all FS methods outperformed the best accuracy achieved by the model which used full features. Interestingly, the Decision Tree achieved the best accuracy on the FS settings by 95%. This accuracy was reported on 19 features (RFE), 51 features (RF), and 76 features (Chi-Square). Here, we reported that the FS methods are effective in increasing the models' performances.

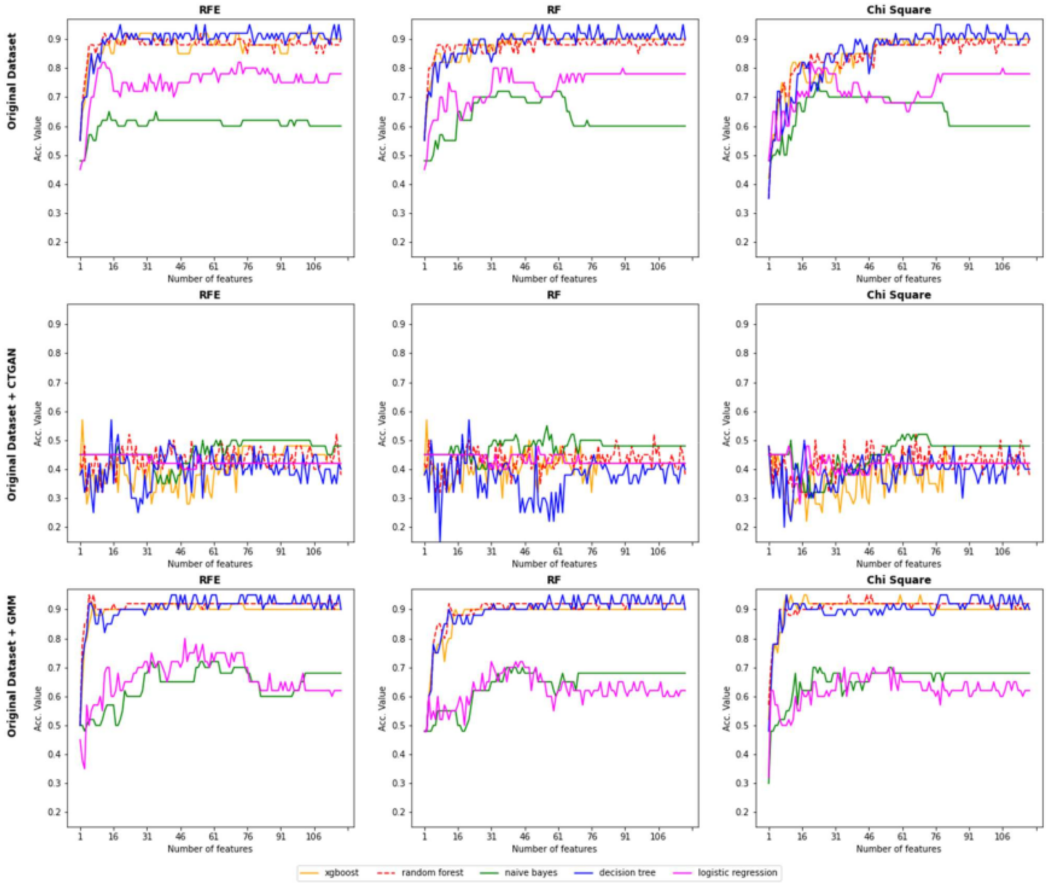

Fig. 4 Detailed accuracies comparison of all scenarios. The *x* axis is used to show the number of features and the *y* axis is used to depict the classification accuracies.

TABLE XII
CLASSIFICATION EXPERIMENT RESULTS ON THREE DIFFERENT DATASETS

Classifiers	Full Features	FS Method: RFE		FS Method: RF		FS Method: Chi Square	
	Acc.	Max. Acc.	ENoF	Max. Acc.	ENoF	Max. Acc.	ENoF
				Original Dataset			
XGBoost	**0.9**	0.92	22	0.92	41	0.9	49
Random Forest	**0.9**	0.92	12	0.9	34	0.9	55
Naïve Bayes	0.6	0.65	14	0.72	34	0.75	23
Decision Tree	**0.9**	**0.95**	19	**0.95**	51	**0.95**	76
Logistic Regression	0.78	0.82	11	0.8	32	0.82	20
				Original Data + GMM			
XGBoost	0.42	0.92	7	0.92	32	**0.95**	9
Random Forest	0.38	**0.95**	5	0.92	12	**0.95**	37
Naïve Bayes	**0.48**	0.72	33	0.7	38	0.7	21
Decision Tree	0.4	**0.95**	42	**0.95**	57	**0.95**	9
Logistic Regression	0.42	0.8	48	0.72	31	0.7	35
				Original Data + CTGAN			
XGBoost	0.9	**0.57**	2	**0.57**	2	0.48	1
Random Forest	**0.92**	0.52	23	0.52	104	**0.52**	67
Naïve Bayes	0.68	0.5	56	0.55	56	**0.52**	61
Decision Tree	0.9	**0.57**	15	**0.57**	21	0.5	17
Logistic Regression	0.62	0.45	1	0.48	52	0.48	10

** ENoF: earliest number of features reaching the best accuracy*

The second experiment was conducted on the second dataset consisting of the original dataset combined with the GMM-based augmented dataset. Surprisingly, classification accuracies by using all features demonstrated the lowest results. The best accuracy obtained in this setting is only 48% by using Naïve Bayes. However, applying FS methods demonstrated a significant improvement. It is worth mentioning that by only using five features, the combination RFE and Random Forest showed 95% accuracy, even other FS methods shared the same accuracy by employing more features.

The last experiment was performed on the third dataset, consisting of the original dataset combined with the CTGAN-based augmented dataset. In contrast with the two previous settings, the best results were achieved by a classification model built using all features. In this setting, the Random Forest without FS method demonstrated 92% accuracy, which is much higher than the best accuracies when using the FSs method achieving 57% (RFE and RF) and 52% (Chi-Square).

From the perspective of the number of selected features, a combination of Random Forest and RFE on the second dataset requires only five features to achieve 95% accuracy. These features are *MotherAge, SC, WeightBefPreg, Weight*, and *SBP*. Even the Decision Tree combined with RFE reached the same 95% accuracy; it needs 19 features. Showing the lowest performance, the XGBoost combined with RFE or RF showed the same 57% accuracy by only two features, i.e., Weight and Hb. The detailed selected features of each setting are shown in Table XIII below.

TABLE XIII
SELECTED FEATURES ON EACH FS METHOD

Dataset	Classifiers	Acc.	NF	Selected Features
Original	Decision Tree + RFE	0.95	19	PregnantFreq, MotherAge, Abortus, IUFD, Tools, SC, BirthWeight, StillbirthWeight, PregnancyPause, MenstrualCycle, Dizzy, Bleeding, Lung-PI, Hypertension-PK, HerbalMed, WeightBefPreg, Height, SBP, Hb.
Original + GMM	Random Forest + RFE	0.95	5	MotherAge, SC, WeightBefPreg, Weight, SBP.
Original + CTGAN	XGBoost+ RFE*(or)*XGBoost +RF	0.57	2	Weight, Hb.

**NF: number of selected features*

IV. CONCLUSION

This paper has developed an automatic detection system for pregnancy risks in Indonesia. This paper aims to address the lack of existing works on this topic, a small-sized pregnancy risk dataset, and no information on which features influence most. We designed experiments to observe which scenario achieved the highest accuracy on the least number of features. The experiment shows that the classification performance can be improved using FS methods and data augmentation. The FSs selected the most important features in this paper to obtain higher accuracies. Furthermore, the GMM-based data augmentation enhanced the results by selecting fewer features for classification and still maintaining accuracy. Overall, the combination of FS methods and GMM achieved better classification performances.

ACKNOWLEDGMENT

This work was supported by the Grant from the Indonesian Ministry of Education (DIKTI) No. 015/SP2HL/L7/2019.

REFERENCES

[1] Kementerian Kesehatan Republik Indonesia, “Profil Kesehatan Indonesia 2015,” M. K. Dr. drh. Didik Budijanto, M.Kes;Yudianto, SKM, M.Si; Boga Hardhana, S.Si, MM ; drg. Titi Aryati Soenardi, Ed. Jakarta: Kementerian Kesehatan Republik Indonesia, 2016, p. 403.

[2] M. Irfan, S. Basuki, and Y. Azhar, "Giving more insight for automatic risk prediction during pregnancy with interpretable machine learning," *Bull. Electr. Eng. Informatics*, vol. 10, no. 3, pp. 1621–1633, 2021.

[3] L. Davidson and M. R. Boland, "Towards deep phenotyping pregnancy: a systematic review on artificial intelligence and machine learning methods to improve pregnancy outcomes," *Brief. Bioinform.*, vol. 22, no. 5, pp. 1–29, 2021.

[4] A. Akbulut, E. Ertugrul, and V. Topcu, "Fetal health status prediction based on maternal clinical history using machine learning techniques," *Comput. Methods Programs Biomed.*, vol. 163, pp. 87–100, 2018.

[5] J. M. Bautista, Q. A. I. Quiwa, and R. S. J. Reyes, "Machine learning analysis for remote prenatal care," *IEEE Reg. 10 Annu. Int. Conf. Proceedings/TENCON*, vol. 2020-Novem, pp. 397–402, 2020.

[6] L. Davidson and M. R. Boland, "Enabling pregnant women and their physicians to make informed medication decisions using artificial intelligence," *J. Pharmacokinet. Pharmacodyn.*, vol. 47, no. 4, pp. 305–318, 2020.

[7] F. Sarhaddi, I. Azimi, S. Labbaf, H. Niela-vilén, and N. Dutt, "Long-Term IoT-Based Maternal Monitoring: System Design and Evaluation," *MDPI Sensors*, vol. 21, pp. 1–21, 2021.

[8] M. W. L. Moreira, J. J. P. C. Rodrigues, A. M. B. Oliveira, K. Saleem, and A. J. V. Neto, "Predicting hypertensive disorders in high-risk pregnancy using the random forest approach," *IEEE Int. Conf. Commun.*, 2017.

[9] M. W. L. Moreira, J. J. P. C. Rodrigues, V. Furtado, C. X. Mavromoustakis, N. Kumar, and I. Woungang, "Fetal Birth Weight Estimation in High-Risk Pregnancies Through Machine Learning Techniques," *IEEE Int. Conf. Commun.*, vol. 2019-May, pp. 1–6, 2019.

[10] M. Tahir, T. Badriyah, and I. Syarif, "Classification Algorithms of Maternal Risk Detection For Preeclampsia With Hypertension During Pregnancy Using Particle Swarm Optimization," *Emit. Int. J. Eng. Technol.*, vol. 6, no. 2, pp. 236–253, 2018.

[11] R. Chu *et al.*, "Predicting the Risk of Adverse Events in Pregnant Women With Congenital Heart Disease," *J. Am. Heart Assoc.*, vol. 9, no. 14, p. e016371, 2020.

[12] E. Purwanti, I. S. Preswari, and Ernawati, "Early risk detection of pre-eclampsia for pregnant women using artificial neural network," *Int. J. online Biomed. Eng.*, vol. 15, no. 2, pp. 71–80, 2019.

[13] H. Sufriyana, Y. W. Wu, and E. C. Y. Su, "Artificial intelligence-assisted prediction of preeclampsia: Development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia," *EBioMedicine*, vol. 54, 2020.

[14] L. Yang *et al.*, "Predictive models of hypertensive disorders in pregnancy based on support vector machine algorithm," *Technol. Heal. Care*, vol. 28, no. S1, pp. S181–S186, 2020.

[15] E. Malacova *et al.*, "Stillbirth risk prediction using machine learning for a large cohort of births from Western Australia, 1980–2015," *Sci. Rep.*, vol. 10, no. 1, pp. 1–8, 2020.

[16] S. Bhadra *et al.*, "Quantifying leaf chlorophyll concentration of sorghum from hyperspectral data using derivative calculus and machine learning," *Remote Sens.*, vol. 12, no. 13, 2020.

[17] P. W. Hatfield *et al.*, "Augmenting machine learning photometric redshifts with Gaussian mixture models," *Mon. Not. R. Astron. Soc.*, vol. 498, no. 4, pp. 5498–5510, 2020.

[18] D. A. B. Oliveira, "Augmenting Data Using Gaussian Mixture Embedding for Improving Land Cover Segmentation," *2020 IEEE Lat. Am. GRSS ISPRS Remote Sens. Conf. LAGIRS 2020 - Proc.*, pp. 333–338, 2020.

[19] A. Arora, N. Shoeibi, V. Sati, A. González-Briones, P. Chamoso, and E. Corchado, "Data augmentation using gaussian mixture model on csv files," *Adv. Intell. Syst. Comput.*, vol. 1237 AISC, no. January, pp. 258–265, 2021.

[20] M. Javeed, M. Gochoo, A. Jalal, and K. Kim, "Hf-sphr: Hybrid features for sustainable physical healthcare pattern recognition using deep belief networks," *Sustain.*, vol. 13, no. 4, pp. 1–27, 2021.

[21] H. Elmoaqet, J. Kim, D. Tilbury, S. K. Ramachandran, M. Ryalat, and C. H. Chu, "Gaussian mixture models for detecting sleep apnea events using single oronasal airflow record," *Appl. Sci.*, vol. 10, no. 21, pp. 1–15, 2020.

[22] A. Singhal, P. Singh, B. Lall, and S. D. Joshi, "Modeling and prediction of COVID-19 pandemic using Gaussian mixture model," *Chaos, Solitons and Fractals*, vol. 138, p. 110023, 2020.

[23] H. Zhang, L. Huang, C. Q. Wu, and Z. Li, "An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset," *Comput. Networks*, vol. 177, no. April, 2020.

[24] A. Saygılı, "Computer-Aided Detection of COVID-19 from CT Images Based on Gaussian Mixture Model and Kernel Support Vector Machines Classifier," *Arab. J. Sci. Eng.*, vol. 47, no. 2, pp. 2435–2453, 2022.

[25] A. Das, U. R. Acharya, S. S. Panda, and S. Sabut, "Deep learning based liver cancer detection using watershed transform and Gaussian mixture model techniques," *Cogn. Syst. Res.*, vol. 54, pp. 165–175, 2019.

[26] K. Sekaran, P. Chandana, N. M. Krishna, and S. Kadry, "Deep learning convolutional neural network (CNN) With Gaussian mixture model for predicting pancreatic cancer," *Multimed. Tools Appl.*, vol. 79, no. 15–16, pp. 10233–10247, 2020.

[27] F. Riaz *et al.*, "Gaussian Mixture Model Based Probabilistic Modeling of Images for Medical Image Segmentation," *IEEE Access*, vol. 8, pp. 16846–16856, 2020.

[28] L. Moraru *et al.*, "Gaussian mixture model for texture characterization with application to brain DTI images," *J. Adv. Res.*, vol. 16, pp. 15–28, 2019.

[29] Y. Yu and W. J. Zhou, "Mixture of GANs for clustering," *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2018-July, pp. 3047–3053, 2018.

[30] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional GAN," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, 2019.

Impact of Feature Selection

ORIGINALITY REPORT

15%	12%	12%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	essay.utwente.nl Internet Source	1%
2	web.njit.edu Internet Source	1%
3	Enrico De Santis, Alessio Martino, Antonello Rizzi, Fabio Massimo Frattale Mascioli. "Dissimilarity Space Representations and Automatic Feature Selection for Protein Function Prediction", 2018 International Joint Conference on Neural Networks (IJCNN), 2018 Publication	1%
4	www.medrxiv.org Internet Source	1%
5	Submitted to Universitas Islam Indonesia Student Paper	1%
6	journals.riverpublishers.com Internet Source	1%
7	amauroboliveira.files.wordpress.com Internet Source	1%

8	www.techscience.com Internet Source	1 %
9	jnk.phb.ac.id Internet Source	1 %
10	elibrary.nusamandiri.ac.id Internet Source	<1 %
11	Guangman Song, Quan Wang, Jia Jin. "Fractional-Order Derivative Spectral Transformations Improved Partial Least Squares Regression Estimation of Photosynthetic Capacity From Hyperspectral Reflectance", IEEE Transactions on Geoscience and Remote Sensing, 2023 Publication	<1 %
12	Submitted to Sunway Education Group Student Paper	<1 %
13	Ruoqi Wei, Cesar Garcia, Ahmed El-Sayed, Viyaleta Peterson, Ausif Mahmood. "Variations in Variational Autoencoders - A Comparative Evaluation", IEEE Access, 2020 Publication	<1 %
14	d-scholarship.pitt.edu Internet Source	<1 %
15	Hannakaisa Niela-Vilen, Eeva Ekholm, Fatemeh Sarhaddi, Iman Azimi, Amir M. Rahmani, Pasi Liljeberg, Miko Pasanen, Anna	<1 %

Axelin. "Comparing prenatal and postpartum stress among women with previous adverse pregnancy outcomes and normal obstetric histories: A longitudinal cohort study", Sexual & Reproductive Healthcare, 2023

Publication

---

16 Shangkun Deng, Xiaoru Huang, Jiahui Wang, Zhaohui Qin, Zhe Fu, Aiming Wang, Tianxiang Yang. "A Decision Support System for Trading in Apple Futures Market Using Predictions Fusion", IEEE Access, 2021 <1%

Publication

---

17 Anthony U Adoghe, Etinosa Noma-Osaghae, Rimamchika Israel Yabkwa. "Photonic Crystal and its Application as a Biosensor for the Early Detection of Cancerous Cells", International Journal of Online and Biomedical Engineering (iJOE), 2020 <1%

Publication

---

18 Ibrahim Maamoun, Mostafa A. Rushdi, Omar Falyouna, Ramadan Eljamal, Osama Eljamal. "Insights into machine-learning modeling for Cr(VI) removal from contaminated water using nano-nickel hydroxide", Separation and Purification Technology, 2023 <1%

Publication

---

19 John Mark Bautista, Rosula S.J. Reyes. "Comparative Analysis of ML algorithms for <1%

	Predictive Prenatal Monitoring", 2022 IEEE 4th Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS), 2022 Publication	
20	doaj.org Internet Source	<1 %
21	Yang Xing, Chen Lv, Huaji Wang, Dongpu Cao, Efstathios Velenis. "An ensemble deep learning approach for driver lane change intention inference", Transportation Research Part C: Emerging Technologies, 2020 Publication	<1 %
22	ijaseit.insightsociety.org Internet Source	<1 %
23	jmoc.state.oh.us Internet Source	<1 %
24	koreamed.org Internet Source	<1 %
25	noa.gwlb.de Internet Source	<1 %
26	tailieu.vn Internet Source	<1 %
27	Snigdha Sen, Krishna Pratap Singh, Pavan Chakraborty. "Dealing with imbalanced regression problem for large dataset using	<1 %

	scalable Artificial Neural Network", New Astronomy, 2022 Publication	
28	Submitted to RMIT University Student Paper	<1%
29	joiv.org Internet Source	<1%
30	eprints.umm.ac.id Internet Source	<1%
31	hdl.handle.net Internet Source	<1%
32	journal.literasisains.id Internet Source	<1%
33	ojs.unud.ac.id Internet Source	<1%
34	Ketut Suarayasa, Elli Yane Bangkele, Sumarni Sumarni, Haerani Harun, Bohari Bohari. "The Effectiveness of M.D-Risti Application as an Alternative for Independent Early Detection of Risk of Pregnancy during the Pandemic COVID-19 in Palu City, Central Sulawesi, Indonesia", Open Access Macedonian Journal of Medical Sciences, 2021 Publication	<1%
35	Narendra Kumar Mishra, Pushpendra Singh, Shiv Dutt Joshi. "Automated detection of	<1%

COVID-19 from CT scan using convolutional neural network", Biocybernetics and Biomedical Engineering, 2021
Publication

36 Paminto Agung Christianto, Eko Sediyono, Irwan Sembiring. "Case-Based Reasoning Modifications for Intelligent Systems in Handling In Vitro Fertilization (IVF) Patients Post Embryo Transfer", 2020 International Seminar on Application for Technology of Information and Communication (iSemantic), 2020
Publication
<1%

37 thesai.org
Internet Source
<1%

38 www.biorxiv.org
Internet Source
<1%

39 www.mdpi.com
Internet Source
<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		