Giving more insight for automatic risk prediction during pregnancy with interpretable machine learning

Muhammad Irfan, Setio Basuki, Yufis Azhar
Informatics Engineering, Universitas Muhammadiyah Malang (UMM), Indonesia

Article Info

Article history:

Received Mar 8, 2020
Revised May 24, 2020
Accepted Mar 7, 2021

Keywords:

C5.0 algorithm
Correlation-based feature selection
Interpretability
Maternal mortality rate
Poedji Rochjati screening card
Risk of pregnancy

ABSTRACT

Maternal mortality rate (MMR) in Indonesia intercensal population survey (SUPAS) was considered high. For pregnancy risk detection, the public health center (puskesmas) applies a Poedji Rochjati screening card (KSPR) demonstrating 20 features. In addition to KSPR, pregnancy risk monitoring has been assisted with a pregnancy control card. Because of the differences in the number of features between the two control cards, it is necessary to make agreements between them. Our objectives are determining the most influential features, exploring the links among features on the KSPR and pregnancy control cards, and building a machine learning model for predicting pregnancy risk. For the first objective, we use correlation-based feature selection (CFS) and C5.0 algorithm. The next objective was answered by the union operation in the features produced by the two techniques. By performing the machine learning experiment on these features, the accuracy of the XGBoost algorithm demonstrated the hightest results of 94% followed by random forest, Naïve Bayes, and k-Nearest neighbor algorithms, 87%, 66%, and 60% respectively. Interpretability aspects are implemented with SHAP and LIME to provide more insight for classification model. In conclusion, the similarity feature generated in the two interpretation approaches confirmed that Cesar was dominant in determining pregnancy risk.

Corresponding Author:

Setio Basuki
Informatics Engineering
Universitas Muhammadiyah Malang
Jalan Raya Tlogomas No. 246, Malang, Indonesia
Email: setio_basuki@umm.ac.id

1. INTRODUCTION

Based on the data in [1], there has been approximately 44% decrease in the maternal mortality rate (MMR) or maternal mortality ration (MMR) globally over the past 25 years, from 1990 to 2015. This reduction becomes part of the millennium development goals (MDGs) program) which was initiated in 2000 with a reduction target based on maternal death indicators per 100,000 births. To maintain the sustainability of the program, sustainable development goals (SDGs) were formed to reduce MMR globally by less than 70 for every 100,000 live births. The decline in MMR values needs to be continued due to 830 female deaths worldwide related to pregnancy and childbirth [2]. In addition, it is revealed that 99% of deaths occured in developing countries. Environmental factors play crucial roles, where an increased risk occurs if women lived in rural areas within poor communities.

Indonesia, as a developing country, has preventive efforts to reduce the value of MMR, such as by conducting the 2015 intercensal population survey revealing that in every 100,000 birth, the MMR value declined to 305 nationally [3]. In fact, the increase in MMR was due to lack of insight in pregnant women leading to the risk of pregnancy. In carrying out the first level of health activities, public health center *(*puskesmas) carries out preventive and promotive actions to sustain public health. To monitor the status of pregnancy, based on Poedji Rochjati screening card (KSPR), Puskesmas can currently detect the status of risks and disorders in pregnant women [4]. In addition, there is also a pregnancy control card as a risk monitoring tool which is completed during pregnancy examination as the first attempt, and followed by the next examination procedures.

In its implementation, there are differences in the number of features in KSPR only amounting to 20 attributes from a total of 117 features available on pregnancy control cards. The list of features on the pregnancy control card is categorized into four categories, which include: pregnancy history, childbirth and family planning, current pregnancy history, general, physical and obstetric examination, laboratory examination. This difference raises questions related to the role of each attribute both on the KSPR and pregnancy control card. The next question lies on whether the 20 features on KSPR are representative for all factors to determine the risk of pregnancy with similar attribute on both cards?

In more specific cases of pregnancy health, the machine learning algorithm indicates promising performance which is beneficial for the detection of high-risk pregnancy [5], [6] and pregnancy health services in general [7]. Furthermore, machine learning algorithm is also increasingly popular, implemented in various tasks with various data sources in the world of healthcare [8]-[10]. However, the resulting model has limitations because it is difficult to be interpreted by experts. The Interpretability aspect holds a key role to provide insight into why certain prediction was made by the model for the patient's condition. In addition, the existence of this interpretative aspect serves as a means of transparency of an intelligent system in predicting the risk of pregnancy. With this transparency, experts such as doctors can validate the output of intelligent systems to avoid the potential result for biased datasets.

Interpretability in the context of artificial intelligence (AI) is defined as a degree where humans (experts) can understand the causes of a decision [11]. With a similar concept, interpretability can also mean that situations where humans can consistently predict the results of machine learning model [12]. This aspect has also become increasingly popular to be combined with powerful methods such as deep learning and ensemble models providing high accuracy with less interpretability [13].

However, recently there has been no research to build a system to predict the risk of pregnancy by utilizing visualization techniques to provide more insight to the user. This facility is worth to appear in AI-based medical applications for further analysis and guarantees prediction accuracy. Furthermore, the urgency of this interpretability aspect also arises because the agreement is required in determining the most significant features of the two monitoring cards. Thus, this study aims to build a system for predicting the risk of pregnancy based on machine learning techniques which is more detailed represented in the three research questions as shown in:

(RQ. 1) Which features are most influential to determine the risk of pregnancy;
(RQ. 2) Are there links in the form of intersecting features on KSPR and pregnancy control cards;
(RQ. 3) How to build interpretable machine learning models for predicting pregnancy risk;

In this research, we propose to utilize the two interpretable methods, which are: local interpretable model-agnostic explanations (LIME) and SHapley Additive EXPlanation (SHAP). LIME works are based on a local learning interpretable model (local surrogate model) focusing on individual prediction to explain individual prediction rather than training the global surrogate model [14]. SHAP works are identified by assigning a value to each feature for a prediction task [15]. This study implements both interpretability models as an expert verification medium for machine learning models in addressing the pregnancy risk cases.

2. PREGNANCY RISK PREDICTION

To answer the three research questions, there are six stages to be implemented, namely: (1) pregnancy risk monitoring to identify features on the KSPR and pregnancy control card; (2) data acquisition to obtain patient data from both monitoring cards; (3) preprocessing data to process raw patient data that is ready to be used on the next stage; (4) feature selection and classification aim to determine the most significant features of pregnancy risk; (5) implementation of interpretable machine learning techniques to provide more insight for classification results. In more detail, each stage is described in the section below.

2.1. Poedji Rochjati screening card (KSPR)

Generally, every pregnant woman who checks her pregnancy to puskesmas will get a copy of the mother and baby health (MCH) book along with KSPR. The card has been arranged to facilitate health workers in screening potential risks for pregnant women. The screening results are utilized to classify mothers into categories such as: low risk pregnancy (LRP) group, high risk pregnancy (HRP) group, and very high risk pregnancy (VHRP) group. Thus, proper actions are easily performed by medical personnel to minimize the potential risks that might arise. Information to complete the KSPR is obtained when a pregnant woman visits the health center and checks her condition. Risks in KSPR are symbolized by numbers, such as: LRP with a score of 2, HRP with a score of 6-10, and VHRP with a score of >=12. Lists of attributes on KSPR, include: Too many children, 4/more; Too young, pregnant in 1≤16 years old; Too old, pregnant in 1≥35 years old; Too old, age of ≥35 years old; Ever failed pregnancy; (stillbirth) baby died in utero; Ever administered the cesarean section; too soon pregnancy (<2 years); too long pregnancy (≥10 years); Swelling on the face/legs and high blood pressure; Bleeding during pregnancy; Location of breech; Location of oblig; and low blood supply. Table 1 shows the list of KSPR attributes, some of these attributes are accompanied by a list of indicators.

Table 1. List of KSPR attributes

No.	Attribute Names
1	Too young, pregnant in 1≤16 years old
2	a. Too late pregnancy 1, get married ≥ 4 years. b. Too old pregnancy 1≥35 years old
3	Too soon pregnancy (<2 yrs.)
4	Re-expecting/pregnant (≥10 th)
5	Too many children 4/more
6	Too old, aged ≥35 th
7	Too short ≤145 cm
8	Failed pregnancy
9	Delivering baby, under the following conditions: a. Vacuum b. Enforced Urinal track c. Transfusion
10	Ever administered the caesarean
11	Illness during pregnancy: a. Low blood supply b. malaria c. TBC d. Cardiac failure e. Diabetes f. STD
12	Face/leg swollen
13	Twin/more babies
14	Hydramnion
15	Stillbirth
16	Post term pregnancy
17	Breech position
18	Oblig position
19	Pregnancy bleeding
20	Chronic pre-eclampsia/eizure

2.2. Pregnancy control card

In this study, polymer data of 400 pregnancy control cards were involved. The data was obtained from the research partner of Cipto Mulyo Malang Public Health Center from 2016 to June 2017. The data of pregnant women cards were in the form of physical files; thus, in collecting data the researchers moved them manually in the form of command separated value (CSV) format. The number of features on the pregnancy control card is 117. On the pregnant mother's card, there are 4 examinations performed, such as: a history of pregnancy, childbirth and birth control, current pregnancy history, general, physical and obstetric examinations, and laboratory examinations. Explanations of each examination are as shown in:

- Pregnancy history, delivery and family planning/birth control

An examination of a history of pregnancy, childbirth and birth control needs are necessary especially for women who are pregnant for more than once. When suspecting the complications in a previous pregnancy, for example a pregnant woman with a history of abortion or miscarriage in a previous pregnancy, then there will be an indication that it can reoccur in the current pregnancy. The examination indicators are shown in Table 2.

- Current pregnancy history

Examination carried out in pregnancy is now aimed at finding out whether there are dangerous indications for the mother's ongoing pregnancy. For example, if pregnant women suffer from hypertension in the current pregnancy, it will be at risk of the mother exposed to pre-eclampsia. Details of pregnancy history are shown in Table 3.

Table 2. Of pregnancy history, childbirth and birth control

History Criteria	Details of History
Pregnancy	pregnancy frequency complication (APB HT)
Childbirth/delivery	Abortus, I/P, IUFD, Normal, Breech, Tool, SC
Delivery sites	Hospital, PKM, BPS, Homebirth, others
Delivery complication	Complicated delivery, Infection, HPP
Medical aids	Obgyn/doctor, midwife, others
Baby weight condition	P/L, BBL(gr), healthy, ill, dead
Current child condition	alive (thn), dead
Family planning/birth control	Yes, No

Table 3. Current history of pegnancy

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

– General, physical and midwifery examinations

This examination is carried out to determine the physical condition of the pregnant woman and the state of the fetus. Table 4 shows the types of examinations that may be carried out to determine general conditions.

Table 4. General, physical and obstetric examination

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

– Laboratory examination

Laboratory tests are performed to determine hemoglobin levels and protein content in the urine. Low hemoglobin levels can cause anemia. Whereas, high urine protein content indicates preeclampsia. The screening criteria include: (a) Hemoglobin (gr); (b) Urine Albumin; (c) Reduction urine; and (d) other Indications.

2.3. Data preprocessing

In many cases, the dataset, both training data and test data, requires further processing to prepare the dataset with good format that fits the classification requirements. The process for preparing this dataset is called data preprocessing. The stages of preprocessing that will be carried out in this study are as shown in:

- Missing value replacement
  The dataset used in this study still contains missing values. Thus, it is necessary to do a substitution technique based on centrality tendency. Substitution is performed by filling in the blank data with the mean value and filling in the blank data with the mode value [16].
- Data transformation
  Data transformation includes the process of converting a dataset structure into another form or structure [16]. The data used in this study has 2 types of attribute data, which are: nominal and numeric to form a good data format for data processing, thus the nominal data type will be transformed into a numeric data type.
- Data normalization
  In some cases, there are data conditions that are quite far apart, requiring data normalization based handling to scale attributes with numeric types. One way to normalize is to use the min-max formula [16].

2.4. Features selection: C5.0 algorithm and correlation-based features selection

The total features used in this study amounted to 120 features, consisting of 117 features of the pregnancy control card, and 3 attributes of KSPR that were not found within the pregnancy control card (not intersecting). From the total features, features are then selected to determine the most influential feature indicating health risk of pregnancy. Feature selection is conducted by using 2 methods of: Correlation-based features selection (CFS) and C5.0 algorithm based on information gain. CFS was selected because it is considered as one of the most stable feature selection methods. This technique considers the use of individual features for class label estimated with the level of intercorrelation among other features. In addition, several studies in the field of medical diagnostics using the CFS method have shown satisfactory results [17], [18]. Whereas, C5.0 algorithm was selected because the feature selection model was based on information gain. This method is a fairly a simple method and is widely utilized in classification cases [19]-[21]. Information gain can help reduce noise due to less relevant features. Information gain detects features presenting the most information by class. Compared to its predecessor, the C4.5 method, the C5.0 method is claimed to be able to produce better accuracy with less memory usage [22].

2.5. Interpretable model for pregnancy risk classification

Increased use of predictive statistical models such as the linear model, rule-based model, classification, and many others, dives to the machine learning model of accountability, transparency, and interpretability. In the case of healthcare, the need for interpretability, fidelity and performance models is considered higher than other domains [23]-[25]. This requirement is due to a large risk in the case of misclassification by the machine learning model. The process of model transparency allows users to understand, audit, and even correct decisions made by the model on the healthcare decision support system.

Every system developed based on machine learning certainly expects a high performance model. In its implementation, there is a trade-off between the interpretability model and the performance model (precision, recall, F-score). The higher the performance of an algorithm, the less interpretability would appear as depicted in Figure 1. The more interpretable models such as decision trees and regression models will have smaller predictive performance when compared to less interpretable models such as boosting and deep learning models, and others [25].

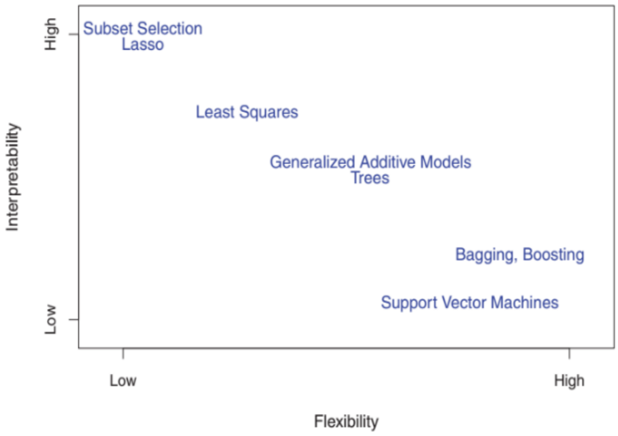


Figure 1. Trade-off antara prediction accuracy dengan model interpretability [26]

2.5.1. Local interpretable model-agnostic explanations

As a local surrogate model, LIME [14] is principled to provide an explanation of the reasons why a machine learning model creates a certain prediction. This process is carried out by observing a machine learning model towards the amount of data provided. LIME model tests on how the prediction process is performed by forming a dataset consisting of permuted samples and predictions as generated by the model. Based on this dataset, LIME then conducts training (decision tree, lasso). The resulting model becomes a good approximation of machine learning models prediction locally, but not globally. LIME as a local surrogate model can be formulated as shown in [27]:

$$explanation(x) = \arg\min_{g\,\epsilon\,G} L(f, g, \pi_{\pi}) + \Omega(g) \quad (1)$$

Explanation model in the (1) aims to minimize the L loss presenting the value of how close the explanation to the prediction from the original machine learning model f while maintaining the complexity model $\Omega(g)$ at a fairly low value. Further, G proves a list of potential explanations that might be generated and π_{π} defines how large the neighborhood is around instances x.

Figure 2 provides an illustration of how LIME works. Complex machine learning models are represented by pink and blue areas which are linearly inseparated. LIME will do sampling against instances, obtain predictive results on samples taken, and give weight to the instance based on the distance from the starting point (the closer the point is valued the more important). Further, some of these samples are uutilized to train the correct simple classifier locally (dotted line) [14].

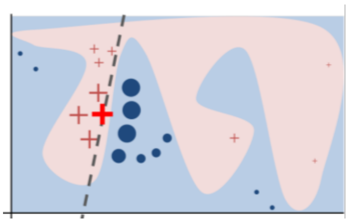


Figure 2. LIME interpretable method [14]

2.5.2. SHapley additive explanation

In interpretable machine learning, it is expected that the model can present an explanation of a classification or prediction result. In the case of prediction for pregnancy risk, a mother is curious to find out why she is predicted to have a high risk of pregnancy, despite uncompromissed condition. For this reason, the system is expected to be able to explain which features have the most influence on increasing the risk of pregnancy, including certain features which have effect on reducing the risk of pregnancy.

SHAP becomes a proper method for this purpose. SHAP will break down the model generated from the machine learning process to determine the effect of each feature on the prediction or classification results. The way SHAP works is by comparing the effect of the presence and absence of a feature on the predicted results. SHAP value can be defined through functions val in S [27].

$$\emptyset_j(p) = \sum_{S \subseteq \{x_1,\dots,x_p\}\setminus x_j} \frac{|S|!(p-|S|-1)!}{p!}\left(val\left(S \cup \{x_j\}\right) - val(S)\right) \quad (2)$$

In which: S is a subset of the feature set in the model, x x is defined as the vector of feature values of the instance being interpreted, dan p is the number of features. One of the strengths of SHAP is the simplicity of the method. SHAP will run on the model generated from the machine learning process with no effect on the model itself. SHAP also provides a pretty good interface to present the effect of a feature on the prediction results. In Figure 3, it is apparent that the blue arrow indicates certain features with a positive influence on the improvement of prediction results, while the red arrows indicate the opposite related to features with a negative influence on the improvement of prediction results. Meanwhile, the length of the arrow indicates the weight of each feature, in which longer and greater weight means that the feature has a greater influence on changes in prediction results.

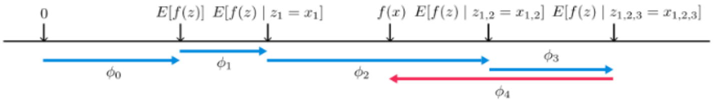


Figure 3. SHAP values [15]

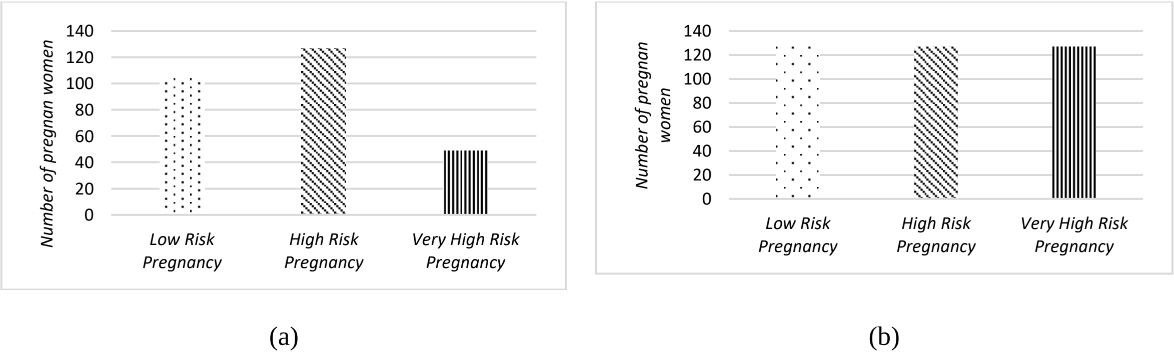
3. RESULTS AND DISCUSSION

In this section, the test results of the research question are presented by identifying a list of the most influential attributes by employing several stages of the scenario. The data in this test is preprocessing data. The results obtained from the two methods are then compared. CFS produces 14 attributes that are as the most important features and C5.0 considered algorithm produces 20 attributes. When compared, it turns out that there are 11 attributes intersecting each other from the list of attributes, as produced by the two methods and 12 other different attributes. Therefore, there are no missing attributes; in this study, twenty-three (23) attributes generated will be applied in the classification process. The list of attribute intersectionsis presented in the Table 5.

Table 5. Comparison for KSPR features and features selection results

CFS	C5.0 Algorithm	KSPR
Hamil_ke (Pregnancy frequency)	hamil_ke (pregnancy frequency)	Too many children, 4/more
UI (Mother age)	UI	Too young, pregnant in 1≤16 th; Too old, pregnant in 1≥35 th Too old, aged ≥35 th
Ab (abortus)	Ab	Failed pregnancy
IUFD (stillbirth)	IUFD	stillbirth
SC (Caesarean section)	SC	Ever administered caesarean
Pregnancy gap	pregnancy gap	Getting pregnant too soon (<2 years) Pregnant too long (≥10 years)
TDS (Systolic blood pressure)	-	Face/leg swollen and hypertension
Bleeding	bleeding	Bleeding during pregnancy
Fetal position ≥ 36weeks	Fetal position ≥ 36weeks	Breech position Oblig position
HB (hemoglobin)	Hb	Low blood pressure
Vomiting	Vomiting	-
Lukas_bekas_op (Surgical scar)	-	-
Hipertensi_PK (Family medical history Hypertension)	Hypertension	Chronic pre-eclampsia/seizure Face/leg swollen and hypertension
Smoking	-	-
-	fetal movement	-
-	abdominal pain	-
-	P/L	-
-	complication	Diseases in pregnant women: Lack of blood, malaria, pulmonary tuberculosis, heart trouble, diabetes, sexually transmitted diseases
-	blood type	-
-	TB	Diseases in pregnant women: Lack of blood, malaria, pulmonary tuberculosis, heart trouble, diabetes, sexually transmitted diseases
-	herb	-
-	tool	-
-	Paru_PI	Diseases in pregnant women: Lack of blood, malaria, pulmonary tuberculosis, heart trouble, diabetes, sexually transmitted diseases

The next step is performed to predict pregnancy risk by using a classification-based machine learning algorithm. The scenario of machine learning model development for prediction is completed by involving the 23 attributes with a number of instances of 400. However, there is an uneven distribution of data among: the severity of the risk level of LRP pregnancy (149 instances), HRP (183 instances), and VHRP (total 68 instances) as presented in Figure 4(a). Imbalance in data distribution affects the low quality of the resulting prediction models. Thusin this study, a synthetic minority over-sampling technique (SMOTE) algorithm is implemented to balance the distribution of data. The results obtained after the balancing process are as depicted in Figure 4(b) where all the risks of pregnancy have a number of instances of 127. To build a prediction model, this study utilized the 4 different algorithms such as: XGBoost, Random Forest, k Nearest Neighbor (kNN), and Naïve Bayes.



(a) (b)

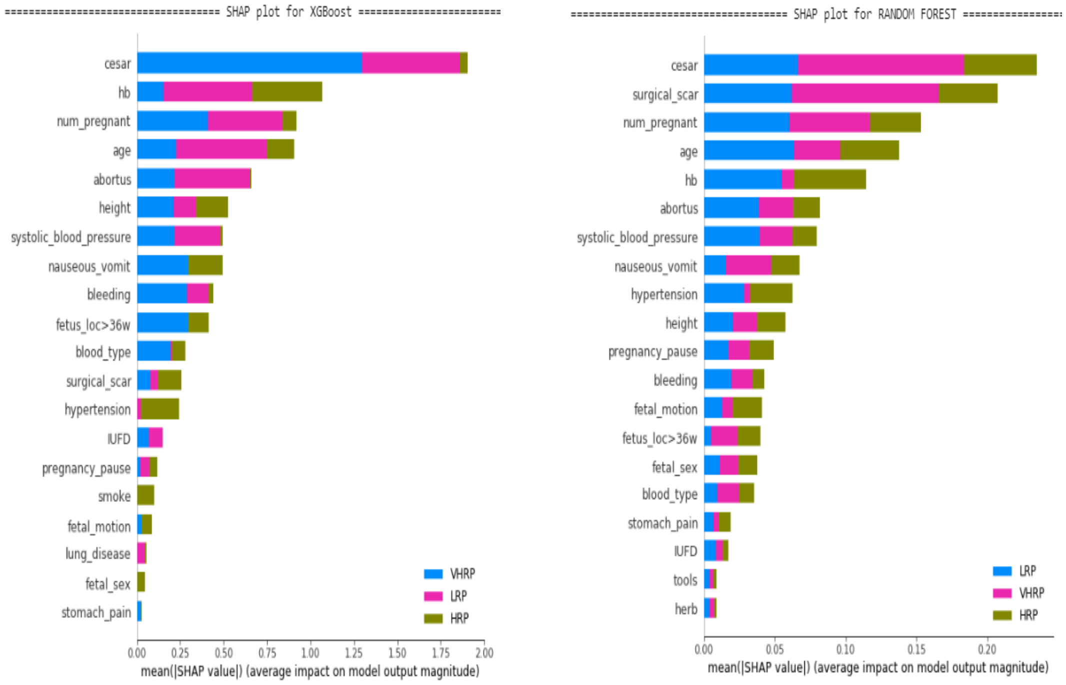
Figure 4. Imbalance pregnancy dataset before and after SMOTE, (a) Distribution before SMOTE, and (b) Distribution after SMOTE

3.1. SHAP visualization result

In Figure 6, the results of the SHAP-based Interpretable model are presented for all classifiers. To interpret it, it is first necessary to know how SHAP visualizes the resulting machine learning model. The y-axis is the name of a feature or variable that is displayed in respective order, based on aspects of its importance variable. The x-axis indicates SHAP value of the variable on the y-axis which is also ordered from the lowest value on the left to the highest value on the right. This x-axis value determines whether the value of the feature is caused by a higher or lower prediction. In the Figure 6, an interpretable model generated by using the SHAP multiclass is depicted in the 4 classifiers. In the plot, the distribution and the average SHAP values for the three classes are explained, which include: LRP, HRP, and VHRP.

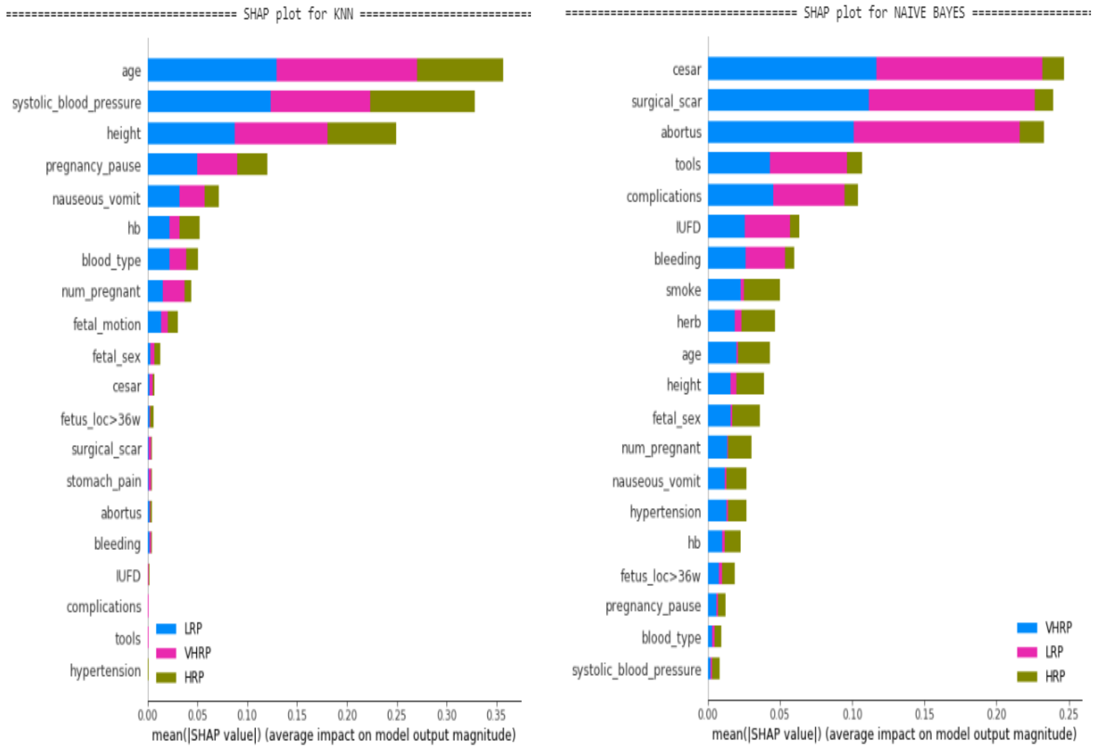
The interpretation process of Figure 6 is conducted by observing at the SHAP value for each class. For example, with the XGBoost algorithm, the average SHAP value generated for the VHRP class is around 1.25, 0.55 (1.80-1.25) for LRP, and 0.1 (1.90-1.80) for HRP. Thus, in general, it is apparent that the feature is highly dominant in influencing model prediction for all classifiers except kNN; however, there is a difference on how the cesar feature influences the model built. In XGBoost and the Random Forest, the Cesar influence predicting VHRP feature is compared to the other two classes such as VHRP and LRP with similar SHAP values.

Furthermore, Figure 6 apparently indicates that the kNN algorithm neglects more than half of the features (11 features) for prediction of LRP, HRP, and VHRP. Other algorithms conclude the resulting model will ignore these 11 features. However, because the kNN algorithm does not form a model, the learning mechanism in kNN algorithm indicates instances of these 11 features which do not show sufficient distance to affect the proximity to training data. A small average of SHAP value also occurs in other algorithms, which is not as small as the average of SHAP value in the kNN algorithm.



(a) (b)

Figure 6. Multiclass SHAP visualization result for four different classification algorithms, (a) SHAP plot for XGBOOST, (b) SHAP plot for random forest



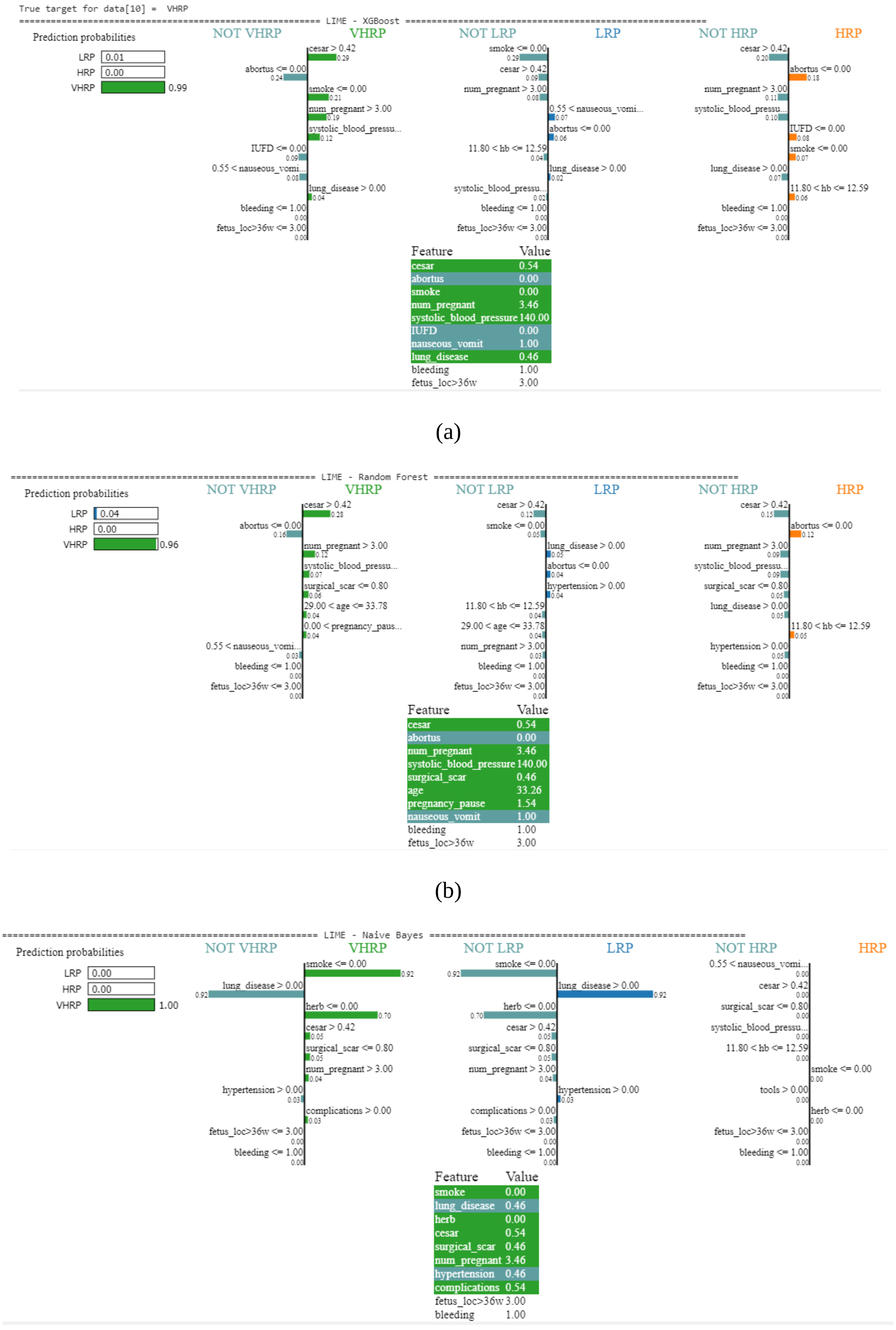
(c) (d)

Figure 6. Multiclass SHAP visualization result for four different classification algorithms, (c) SHAP plot for kNN, (d) SHAP plot for Naïve Bayes *(continue)*

3.2. LIME visualization result

LIME-based visualization model consists of three main parts. The first part contains the prediction probability (the far left part of the plot) which contains information about the probability distribution of the target classes, including: LRP, HRP, and VHRP. The second part contains the list of Importance Feat ures (to the right of the prediction probability section) that most contribute to the resulting model. The third part contains the actual values of the list from the most important features (at the bottom of the plot). In Figure 7, it is apparent that LIME only displays the list of features that most influence the model, which is different from the SHAP-based interpretable model displaying all features with the SHAP value for every single feature.

If observed, LIME visualization for all classifiers in Figure 7 presents several notable phenomena in the case of a multiclass LIME, which provides the negation option of a class. For example in the XGBoost model, it is indicated that VHRP class is supported by a feature collection and the negation (non) VHRP class is also supported by a feature collection. In this example, there is a Cesar feature of > 0.42 which means that this feature's value satisfies the criteria that support the VHRP class. If observing the four LIME Plots, it is apparent that the Cesar plot for XGBoost and Random Forest feature is very dominant in determining the three classes of pregnancy risk. In two other algorithms such as Naïve Bayes, LIME plot indicates that feature of smoking dominates in the formation of the model and LIME Plot in kNN is strongly influenced by feature such as blood pressure and age.



```
True target for data[10] =  VHRP
========================================================= LIME - XGBoost =========================================================
```

Feature	Value
cesar	0.54
abortus	0.00
smoke	0.00
num_pregnant	3.46
systolic_blood_pressure	140.00
IUFD	0.00
nauseous_vomit	1.00
lung_disease	0.46
bleeding	1.00
fetus_loc>36w	3.00

(a)

```
========================================================= LIME - Random Forest =========================================================
```

Feature	Value
cesar	0.54
abortus	0.00
num_pregnant	3.46
systolic_blood_pressure	140.00
surgical_scar	0.46
age	33.26
pregnancy_pause	1.54
nauseous_vomit	1.00
bleeding	1.00
fetus_loc>36w	3.00

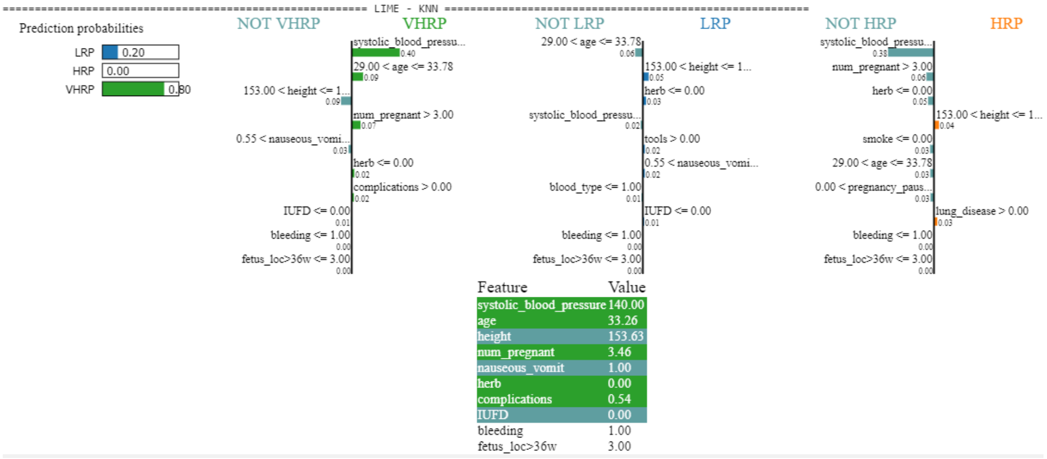
(b)

```
========================================================= LIME - Naive Bayes =========================================================
```

Feature	Value
smoke	0.00
lung_disease	0.46
herb	0.00
cesar	0.54
surgical_scar	0.46
num_pregnant	3.46
hypertension	0.46
complications	0.54
fetus_loc>36w	3.00
bleeding	1.00

(c)

Figure 7. Multiclass LIME visualization result for four different classification algorithm, (a) LIME plot for XGBOOST, (b) LIME plot for Random Forest, (c) LIME plot for Naïve Bayes



Feature	Value
systolic_blood_pressure	140.00
age	33.26
height	153.63
num_pregnant	3.46
nauseous_vomit	1.00
herb	0.00
complications	0.54
IUFD	0.00
bleeding	1.00
fetus_loc>36w	3.00

(d)

Figure 7. Multiclass LIME visualization result for four different classification algorithm, (d) LIME plot for kNN (*continue*)

3.3. Comparison of accuracy of pregnancy risk classification

The prediction model is established based on a dataset of three balanced classes. Classification accuracy testing is completed by forming the composition of training data and testing data with ratio of 80:20. The test produces a comparison of the accuracy value as depicted in the Table 6 The XGBoost algorithm has the highest accuracy value of 94%, followed by Random Forest, Naïve Bayes, and kNN of 87, 66, and 60 respectively.

Table 6. Comparison of accuracy value of pregnancy risk classification

	XGBOOST	kNN	Naïve Bayes	Random Forest
Classification Accuracy (%)	94	60	66	87

4. CONCLUSION

In this study, a pregnancy risk prediction system was established in Indonesia based on the features inherent in pregnant women. Pregnancy risk is divided into three which are: low risk pregnancy (LRP) group, high risk pregnancy (HRP) group, and very high risk pregnancy (VHRP) group. For this reason, a pregnancy dataset is required as a representation of the mother's condition during pregnancy. This study involved 400 pregnancy data. The data cannot be directly applied because there are problems with the format and consistency and even distribution of three pregnancy risk statuses. For this reason, data preprocessing, selection attributes, and data balancing have been carried out. The process of forming the model is conducted by using 4 machine learning algorithms, such as: XGBoost, Random Forest, Naïve Bayes, and kNN. The classification results demonstrated that the XGBoost algorithm presented the highest accuracy value of 94% which was followed by Random Forest, Naïve Bayes, and kNN, each of which was equal to 87%, 66%, 60%. Both SHAP and LIME-based plots indicated the suitability of feature importance in all classes and all applied algorithms because both of these Interpretable Machine Learning techniques interpret the same model. However, there is a difference between the two; in which LIME only displays the list of features that most influence the model, unlike the SHAP-based interpretable model that displays all features (along with the SHAP value for every single feature).

REFERENCES

[1] WHO, “Trends in Maternal Mortality : 1990 to 2015,” *United Nations Population Fund*, p. 98, 2015.
[2] WHO, “Maternal Mortality,” 2018.
[3] D. Budijanto;Yudianto, B. Hardhana, and T. A. Soenardi, “Indonesia Health Profile 2015 (in bahasa: Profil Kesehatan Indonesia 2015),” Kementerian Kesehatan Republik Indonesia, pp.1- 403, 2016.
[4] P. Rochjati, “Skrining Antenatal Pada Ibu Hamil (in bahasa: Antenatal Screening for Pregnant Women),” Airlangga University Press, vol. 2, p. 43, 2011.
[5] G. Guidi, G. Adembri, S. Vannuccini, and E. Iadanza, “Predictability of Some Pregnancy Outcomes Based on SVM

and Dichotomous Regression Techniques," in *International Workshop on Ambient Assisted Living IWAAL 2014*, Springer, Cham, 2014, vol. 8868, pp. 163-166, doi: 10.1007/978-3-319-13105-4_25.

[6] K. Paydar, S. R. Niakan, A. Sheikhtaherim and M. Akbarian, "A clinical decision support system for prediction of pregnancy outcome in pregnant women with systemic lupus erythematosus," *International Journal of Medical Informatics,* vol. 97, pp. 239-246, 2017, doi: 10.1016/j.ijmedinf.2016.10.018.

[7] I. Pan, *et al.*, "Machine Learning for Social Services : A Study of Prenatal Case Management in Illinois," *American journal of public health,* vol. 107, no. 6, pp. 938-944, 2017.

[8] S. Islam, *et al.*, "A Systematic Review on Healthcare Analytics : Application and Theoretical Perspective of Data Mining," *Healthcare,* vol. 6. no. 2, p. 54, 2018, doi: 10.3390/healthcare6020054.

[9] F. Jiang *et al.*, "Artificial intelligence in healthcare : past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, pp. 230- 243, 2017, doi: 10.1136/svn-2017-000101.

[10] S. Widodo, *et al.*, "Lung diseases detection caused by smoking using support vector machine," *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 17, no. 3, pp. 1256-1266, 2019, doi: 10.12928/telkomnika.v17i3.9799.

[11] T. Miller, "Explanation in Artificial Intelligence : Insights from the Social Sciences," *Artificial Intelligence*, vol. 267, pp. 1-38, 2018, doi: 10.1016/j.artint.2018.07.007.

[12] B. Kim, R. Khanna, and O. O. Koyejo, "Examples are not Enough, Learn to Criticize ! Criticism for Interpretability," *NIPS*, pp. 2280-2288, 2016.

[13] S. M. Lundberg *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering*, vol. 2, no. 10, pp. 749-760, 2018, doi: 10.1038/s41551-018-0304-0.

[14] M. T. Ribeiro and C. Guestrin, "Why Should I Trust You ?' Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144, 2016, doi: 10.1145/2939672.2939778.

[15] S. M. Lundberg and S. Lee, "A Unified Approach to Interpreting Model Predictions," *arXiv preprint arXiv:1705.07874*, vol. 1, no 2, pp. 1-10, 2017.

[16] J. P. Jiawei Han and Micheline Kamber, "Data Mining-Concepts & Techniques," Morgan Kaufmann Publishers is an imprint of Elsevier, 2011.

[17] M. Mursalin, *et al.*, "Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier," *Neurocomputing,* vol. 241, pp. 204-214, 2017, doi: 10.1016/j.neucom.2017.02.053.

[18] O. Cigdem *et al.*, "Diagnosis of Bipolar Disease Using Correlation-Based Feature Selection with Different Classification Methods," *2019 Medical Technologies Congress (TIPTEKNO),* Izmir, Turkey, 2019, pp. 1-4, doi: 10.1109/TIPTEKNO.2019.8895232.

[19] M. Hassoon, *et al.*, "Rule Optimization of Boosted C5.0 Classification Using Genetic Algorithm for Liver disease Prediction," *2017 International Conference on Computer and Applications (ICCA),* Doha, Qatar, 2017, pp. 299-305, doi: 10.1109/COMAPP.2017.8079783.

[20] M. Hassoon, *et al.*, "Rule Optimization of Boosted C5.0 Classification Using Genetic Algorithm for Liver disease Prediction," *2017 International Conference on Computer and Applications (ICCA),* Doha, Qatar, 2017, pp. 299-305, doi: 10.1109/COMAPP.2017.8079783.

[21] L. Z. Albances, *et al.*, "Application of C5.0 Algorithm to Flu Prediction Using Twitter Data," *2018 International Conference on Platform Technology and Service (PlatCon)*, Jeju, Korea (South), 2018, pp. 1-6, doi: 10.1109/PlatCon.2018.8472737.

[22] R. Revathy and R. Lawrance, "Comparative Analysis of C4. 5 and C5. 0 Algorithms on Crop Pest Data," in *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5, no. 1, pp. 50-58, 2017.

[23] U. Bhatt, B. Davis, and J. M. Moura, "Diagnostic Model Explanations: A Medical Narrative," *AAAI Spring Symposium: Interpretable AI for Well-being,* 2019.

[24] R. Elshawi, M. H. Al-Mallah, and S. Sakr, "On the interpretability of machine learning-based model for predicting hypertension," *BMC Medical Informatics and Decision Making,* vol. 19, no. 1, p. 146, 2019, doi: 10.1186/s12911-019-0874-0.

[25] M. A. Ahmad, C. Eckert, A. Teredesai, and G. Mckelvey, "Interpretable Machine Learning in Healthcare," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics,* pp. 559-560, 2018, doi: 10.1145/3233547.3233667.

[26] G. James, *et al.*, "Springer Texts in Statistics An Introduction to Statistical Learning," *Springer,* vol. 10, p. 978-1, 2013.

[27] C. Molnar, "Interpretable Machine Learning-A Guide for Making Black Box Models Explainable," *ChristophMolnar*, 2019

BIOGRAPHIES OF AUTHORS

Muhammad Irfan was born in Mojokerto, Indonesia in 1966. He graduated in 1991 with a bachelor's degree in Engineering, from the Department of Electrical Engineering, Universitas Brawijaya Malang, Master graduated in 2000 from the Department of Informatics, Sepuluh Nopember Institute of Technology (ITS), Surabaya. Currently, he is a senior lecturer at the Universitas Muhammadiyah Malang (UMM) and active in the Directorate of Research and Community Service. His areas of interest are Energy, Energy Audit, Digital Signal Processing, and Information Technology.

Setio Basuki is a faculty in the Information department of the Universitas Muhammadiyah Malang (UMM). He completed his bachelor from Telkom Institute of Technology (IT Telkom) Bandung in 2007. He completed his master’s degree at the Informatics Department of the Bandung Institute of Technology (ITB) in 2015. While taking his Masters, Setio Basuki took the Intelligent System option with a focus on Natural Language Processing. He is actively researching Machine Learning applications in the Health and Education fields. Now, he is continuing his Doctorate at Toyohashi University of Technology, Japan.

Yufis Azhar is a faculty in the Informatics department of the Universitas Muhammadiyah Malang (UMM). He completed his undergraduate studies at the same study program and university in 2009. In 2011, he continued his master's degree in the Informatics Engineering study program of the Ten November Institute of Technology and completed his studies in 2013. Currently, he is actively researching in the field of computer vision and artificial intelligence.