Exploratory Study of Some Machine Learning Techniques to Classify the Patient Treatment

Mujiono Sadikin[1], Ida Nurhaida[2]
Faculty of Computer Science
Universitas Mercu Buana, Jakarta, Indonesia

Ria Puspita Sari[3]
Siloam Hospital
Jakarta, Indonesia

Abstract—Numerous studies have been carried out on computation and its applications to medical data with proven benefits for improving the quality of public health. However, not all research results or practical applications can be applied to all conditions but must be in accordance with the various contexts such as community culture, geographical, or citizen behaviors. Unfortunately, the use of digital data in Indonesia is still very limited. The study objective is to assess various data mining techniques to utilize data from laboratory test results collected from a private hospital in Indonesia in predicting the next patient treatment. Furthermore, various machine learning classification techniques were explored for the purpose. Based on the experiments, it was concluded that XGBoost with hyperparameter tuning produced the best accuracy level at 0.7579, compared to other classifiers. A better level of accuracy can be obtained by enriching the type of dataset used, such as the patient's medical record history.

Keywords—Electronic health record; XGBoost; patient treatment; patient laboratory test data

I. Introduction

The success of medical treatment services is dependent on the quality of health services and the information precision related to the medical aspects [1]. Unfortunately, the access performed to the relevant medical information is increasingly difficult due to the rapid growth in data volume and its heterogeneous format as well. Health care is one of the most complex industries which includes many stakeholders, various tools, and technologies as well [2]. The new techniques are always needed to assist the dealing with this type of data with the computational technique used to address the problem related to medical information.

Various techniques have been carried out on medically related datasets for many purposes. Due to the broad scope of the medical and health fields, various research topics are found which ranges from diagnosis [3]–[5], diseases [3], [4], [6]–[11], patient's condition [12]–[17], prescription and medication [1], [18]–[21], to genomics [22]–[24].

The development of the health system, its problems, and challenges tightly relates to multi factors and contexts such as geographic location, local regulation, community-style demographics, wealth level, etc. The contextual factors are the most important part used to develop the health-medical researches endorsed by the Agency for Healthcare Research and Quality USA [25]. These factors are following the World Health Organization, which supports the achievement level of best health services quality, and medical devices operation base

on the contextual context [26]. The studies regarding the contextual factors in developing the Primary Health Center (PHC) [39], showed that many factors such as social models, an institutional context that promotes risk-averseness and patient care, infrastructure, community expectation, and doctors' disinterest in primary care roles need to be considered.

Unfortunately for our local context i.e. Indonesia, compared to a very large population and a very large area of the country, the studies regarding the medical records or electronical health are very limited. Some of the studies which focus on the local context are published in [11] and [15]. The first article presents the study results of early dengue disease detection with the dataset captured from some public health (PUSKESMAS). The study overcame the problem associated with physical detection methods in detecting the patient's symptoms by comparing some conventional classifiers with the ELM technique. In the second study, the authors performed the toddler's nutritional status identification using the clustering method, which is categorized into 5 clusters: good, moderate, malnutrition, over, and obesity. The other study is the enrichment of ontology in tuberculosis epidemiology domain use the pulmonary TB (Tuberculosis) scientific documents [27].

Considering the importance of the right context in conducting health-medical research, this study was conducted to utilize the value of patient medical data from the results of laboratory tests taken from one of a private hospital in Indonesia. This dataset is the main consideration factor used by doctors to determine the next course of action for patients, whether they need to be hospitalized (in-patient care) or not (out-patient care). Various literary studies show that AI and Datascience-based tools are proven to be able to improve the quality of health services. According to authors' knowledge, in the field of health care in Indonesia, there are no AI-based tools available. This research is an attempt to contribute to this field.

In this study, we elaborate on some of the machine learning techniques used to classify these patient treatments based on the laboratory test results data. Compared to the other technique, the XGBoost with Grid Search hyperparameters optimization performance is outperform.

In addition, this research utilized EHR data from the local context to determine the characteristics of patients' treatment, with similar pattern distribution. Therefore, the machine learning technique was proposed by the authors to handle this problem. The article is organized as follows: the first section

describes the background and justification of the research. This is followed by section two; the material and method of study are used to analyze the dataset and research methodology. The next section presents the results and discussion regarding the experiments. And in the last section, we summarize our research findings as a study conclusion.

II. RELATED WORK

Artificial intelligence (AI) – Machine Learning (ML) is one of the most powerful techniques used to address any problem in the health and medical sector. The study carried out on paediatric diseases, as published in [5], the research showed that the Machine Learning Classifiers (MLCs) handles 101.6 million data points from a total of 1,362,559 pediatric patients. The hypothetic-deductive reasoning used by physicians and unearth associations are difficult to be found by the conventional statistical methods. D. S. Kermany et al. utilized the deep learning algorithm to identify medical diagnoses and treatable diseases [4]. The approach performance in classifying age-related macular degeneration and diabetic macular edema is comparable to human experts. The Machine Learning approach supports the decisions related to patient diseases and based on the imbalance classes proposed by [28] Previous studies defined the ML approach as the CCOA-RA used to overcome the imbalance negative and positive labels of the dataset.

Since its inception [29], the Extreme Gradient Boost (XGBoost) has been the favourite technique used to address the challenge of classification prediction in the medical area. By utilizing XGBoost, many researchers addressed various sub-topic in the health/medical field such as diagnoses [5], [8], [30]–[34], related diseases[4][6], [9], [34], [35], medical treatment [36], [37], patient status[12], [14], [16], [38], or event genomic [39].

A research conducted an experimental study using XGBoost Classifiers with some scenarios such as transformation, resampling, clustering, and ensemble learning to predict the diagnosis of second primary cancers (SPCs) [30] The resampling and clustering strategies were used to determine the best method used to identify some important risk factors associated with SPCs in patients with breast cancer. The combination of the XGBoost and Clustering analysis approach was also proposed by [8] to predict the hypertension-related symptoms from 531 hypertensive patients data in a hospital in Beijing. These combination techniques showed that there are significant differences in symptomatic entropy between patients with type II and type I hypertension. The experimental study of various classifiers techniques, such as XGBoost, was conducted by [32] to predict falls – non falls of Parkinson Diseases. Therefore, clinical, demographic, and neuroimaging data used in this study were obtained from Medical Centres, University of Michigan, and Tel Aviv Sourasky Medical Center. The research finding has a prediction accuracy value of 70 % - 80 %, which is used to provide a more reliable clinical outcome forecast of falls in Parkinson's patients. The superiority of XGBoost used to predict and diagnose Alzheimer-type dementia in the blood is divided into two categories, namely Alzheimer's Disease (AD) and cognitively normal (CN), according to [5]. In this study,

experiments were conducted using some Classifier technique applied to 883 patient's data. The experiment finding shows that XGBoost gave the best performance. The high emergency diagnostics error rate is also commonly found in Urinary tract infection (UTI) due to clinical or physical symptoms. Machine learning based on the XGBoost technique has been demonstrated as the powerful tools used to overcome the challenge published in [32]. According to previous studies, the UTI prediction consists of six machine learning algorithms, with medical and social information. Therefore, the authors claim that XGBoost accurately diagnosed positive urine culture results.

III. MATERIAL AND METHOD

A. Dataset

Data were collected from a total of 80,000 patient's laboratory test daily recorded by the Hospital Information System (HIS) in January 2019 as shown in Table I. The privacy records such as PATIENT_ID, PATIENT_NAME, and CLINICIAN were not presented in the table. Out of eight attributes, only four were used in the experiments, namely, TEST, AGE, SEX, RESULTS, and FLAG. The indicators of RESULT are High, Low, or Normal

TABLE I. ATTRIBUTES AND VALUE SAMPLE OF RAW DATA

No.	Attribute	Value Sample
1.	Date	1/1/19
2.	Test	Number of Leucocyte
3.	Lab No.	19000001
4.	Age	15Y
5.	Sex	M
6.	Source	UGD
7.	Result	15.6
8.	Flag	H

B. Method

1) Research stages: The research stages globally consist of 3 blocks activities, namely data processing, modelling, and evaluation, as shown as the flow diagram in Fig. 1. The data processing stage comprises of collection and pre-processing activities. Some classifier techniques are used in the modelling stage, which is described in the next sections. In addition, the accuracy and AUC-ROC parameters were used to evaluate each technique performance.

Data were collected from 80.000 records of patients' laboratory test in HIS operated by a private hospital in Jakarta, Indonesia. All records were taken from transactions that were generated in January 2019.

This was followed by the processing step, which includes attribute removal, ignoring record with missing value, transforming rows to columns, value scaling normalization, and labelling. Some removable attributes are related to privacy such as patient and clinician's name as well as those attributes with no meaning to the study such is PATIENT_ID and LAB_ID. However, after row-column transformation and

ignoring the records with missing value, the dataset finally contains 4412 instants. Furthermore, labelling was manually performed in accordance with the patients' medical records. One of two instant data label is either 1 or 0, which represents inpatient and outpatient care. After the pre-processing step, each instant consists of 8 (eight) attributes, namely HAEMATOCRIT, HAEMOGLOBINS, ERYTHROCYTE, LEUCOCYTE, THROMBOCYTE, MCH, MCHC, MCV, AGE, SEX. The processed dataset is published in http://bit.ly/3kbK3Wn, and some of the instant samples, as presented in Table II.

2) Experimental scenario: Two experiments were carried out in accordance with the final data presentations. The first was conducted using the original value as shown in Table II, and labelled as Format 1:Lab Test Results, while the second scenario used the value transformed/coding of attributes as Format 2: Laboratory test result formatted attributes. In Format 1 the original value was the normalization scaling of 0 to 1 for all numeric value attributes. Meanwhile, the Format 2 dataset, used a rule regarding the laboratory test component obtained from some medical references[40], [41]. The rule is categorized into three levels, namely Low, Normal, and high, base on gender and age of the patient, as shown in Table III. For the attribute value transformation, 0, 1, and 2 were used to represent low, normal, and high, respectively.

In the first step of the modelling phase, to both of format data representation above we apply some classifier techniques such as the decision tree, Gaussian naïve Bayes, random forest, Adaboost, and XGBoost as shown in Fig. 1. The better format data representation is then used in the second step modelling phase, which also chooses the two best performances of techniques from the first step. In the first modelling step, the cross-validation scheme was chosen as the data testing splitting scenario since it is more representative compared to the random splitting process. Accuracy parameter was also used as the performance evaluation criteria, whereas in the second step, the AUC-ROC parameter was utilized.

The second step of the modelling phase was conducted by choosing the best two classifiers applied to the better format of data representation which gives the better performance. The

hyperparameter optimization was performed on both classifiers to obtain best performance using the GridSearchCV which adapted from the skcit-learn library [42] In the second modelling we use the random splitting scenario of training – testing data selection, which the training data part is used in best hyperparameter searching and model training whereas the testing data part is for model validation (testing).
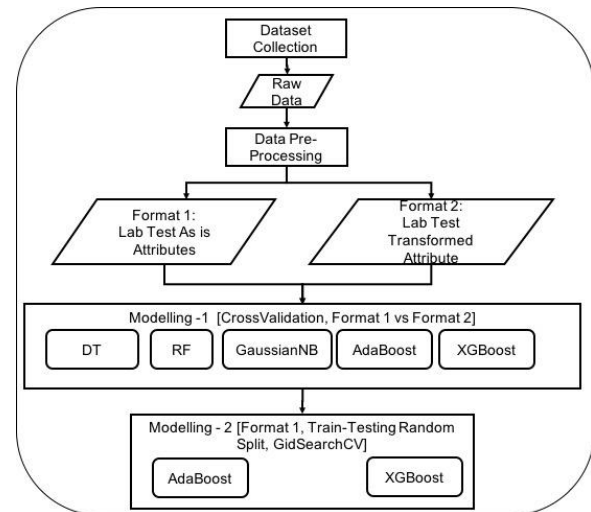

Fig. 1. Research Flow Diagram.

TABLE II. ATTRIBUTES AND VALUE SAMPLE OF RAW DATA

Attributes	Remark	Value Sample
HAEMATOCRIT	Continuous	35.1
HAEMOGLOBINS	Continuous	11.8
ERYTHROCYTE	Continuous	4.65
LEUCOCYTE	Continuous	6.3
THROMBOCYTE	Continuous	310
MCH	Continuous	25.4
MCHC	Continuous	33.6
MCV	Continuous	75.5
AGE	Continuous	12
SEX	Nominal - Binary	F

TABLE III. LABORATORY TEST RESULT FROM THE CATEGORIZATION RULE

	Low if less than*				High if more than*			
Gender	*Male*		*Female*		*Male*		*Female*	
Age	*Adult*	*Infant*	*Adult*	*Infant*	*Adult*	*Infant*	*Adult*	*Infant*
HAEMATOCRIT	38,8	33	34,9	33	50	38	44,5	38
HAEMOGLOBINS	13		12		17		15	
ERYTHROCYTE	4,7	4,1	4,2	4,1	6,1	5,5	5,4	5,5
LEUCOCYTE	5	5	5	5	10	10	10	10
THROMBOCYTE	150	150	150	150	400	400	400	400
MCH	27	27	27	27	33	33	33	33
MCHC	32	32	32	32	37	37	37	37
MCV	80	80	80	80	96	96	96	96

*Normal If the Value Is In Between

3) Evaluation criteria: This study used the commons criteria performance parameter in classification i.e. accuracy. The Accuracy computation is presented as formula (1) [43].

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad (1)$$

where:

TP   : True Positive

TN   : True Negative

FP   : False Positive

FN   : False Negative

The objective of the classification scenario is to achieve good quality performance of both classes. Therefore, the individual measures of both negative and positive classes are combined. The weakness of the accuracy parameters is none of them is adequate by itself. That's why we also use the other measure which is the common approach to unify those measures and produce an evaluation criterion such as the *Receiver Operating Characteristic (ROC)* graphic. ROC graphic represents the visualization of the trade-off between the benefits (*TPrate*) and costs (*FPrate*). The Area Under the ROC Curve (AUC) corresponds to the probability used in identifying the noise in the two stimuli [43]. Formula (2) presents the computation of AUC measure.

$$AUC = \frac{1+TPrate-FPrate}{2} \qquad (2)$$

where

$$TPrate = \frac{TP}{TP+FN}T \qquad (3)$$

$$FPrate = \frac{FP}{FP+TN} \qquad (4)$$

TPrate is the percentage of positive instances correctly classified is positive class.

FPrate is the percentage of positive instances misclassified.

In the first step of the modelling phase, some classifier techniques as depicted in Fig. 1: the decision tree (DT), gaussian naïve Bayes (GaussianNB), random forest (RF), AdaBoost, and XGBoost are applied to the data representation. The better format data representation is then used in the second step modelling modeling step, which also chooses the two best performances of techniques from the first step. The cross-validation scheme was selected as the training and testing data testing splitting scenario because it is assumed the cross-validation is more representative compared to random splitting. Furthermore, the accuracy parameter is used as the performance evaluation criteria in the first step of the modelling phase, whereas in the second step, the model was evaluated by using the AUC-ROC parameter.

4) Overview of machine learning techniques: The five algorithms used to explore the experiments are presented in this section as follows:

a) Decision Tree: Decision tree (DT) is one of the techniques widely-used for classification purposes [44] The decision tree built is similar to a flowchart [45] and acts as a predictive model that contains a mapping between object values in the tree and the data attributes. The classifier is represented as the decision node of each instant data attributes, whereas the tree branches correspond to different prediction output. Each leaf node represents a possible output of the final presentation of the DT construction phase, which is expressed as the construction and pruning phases. In the construction phase, the overall planning of the DT main structure is completed, whereas in the pruning phase, a more precise pruning process is performed. The main advantage of DT compared to other methods is that it is very interpretable. In a certain field, such as healthcare, the interpretability is often preferred rather than the higher accuracy and relatively uninterruptable [44].

b) Random Forest: Random Forest (RF) is the combination of supervised and unsupervised learning algorithms capable of increasing the accuracy of machine learning classifiers [46]. As a multi-class classifier, it is resistant to noise, fast in training and classification, and has powerful classification capabilities [47]. Furthermore, the ensemble learning technique is based on a decision tree and widely used in various areas with almost ideal prediction [48]. Y. Mishina, R. Murata, Y. Yamauchi, T. Yamashita, and H. Fujiyoshi claim that RF is more robust than other famous models and have been utilized in many fields such as, computer visions and pattern recognition [49]. The common weakness in using RF is the processing needs more time when applied to large amounts of data because it has to build many tree models [50]. A large number of trees also require significant memory capacity [49]. The summarization of the main process to construct the RF is referred to [51].

c) Gaussian Naïve Bayes: The Gaussian Naïve Bayes (GaussianNB) algorithm is categorized as the supervised learning method [52] It is extended to real-valued attributes through a Gaussian distribution network. GaussianNB algorithm assumes that the probability of each attribute belonging to a given class value doesn't depend on all other attributes. When the value of the attribute is identified, the probability is called conditional. Data instances probability is computed by multiplying all conditional attributes. The prediction is formulated by computing each class instance and by selecting the highest probability class value [53].

d) AdaBoost: AdaBoost (Adaptive Boosting) is known as the most famous Boosting algorithms, as proposed by [54]. It is able to self-adjust the weak classifiers after learning, and it is sensitive to noise data and outliers. AdaBoost has the ability to avoid overfitting some tasks efficiently and boosts weak learners to converge and become stronger classifiers [55]. The improved performance is achieved with different types of algorithms with the outputs obtained from the combination of a weighted sum.

The sampling to train data is used to replace the random sampling, which places attention on training data that are difficult to process. Weak classifiers are combined by replacing the average voting with a weighted mechanism. The effectiveness of the integrated weak classifiers is guaranteed by equipping the weak classifiers that are effective with a higher

weight and equipping those that are ineffective with a lower weight. It is not necessary to determine the learning performance of the weak algorithm in advance since the classification accuracy of the integrated strong classifier depends on the accuracy of all weak classifiers. The Algorithm that interprets the main idea of AdaBoost can be referred in [56].

e) XGBoost: XGBoost is a scalable machine learning system for tree boosting, which is proposed as an alternative method for predicting a response variable using certain covariates [26]. The system is available as an open-source package and widely recognized in many challenges of machine learning and data mining. Based on a study conducted by [8], some of the advantages of XGBoost are as follows, supports linear classifier, regularization to control the model complexity, capability in using the second order of Taylor expansion, and some more.

IV. RESULTS AND DISCUSSION

A. Data Analysis

Data analysis was conducted after some pre-processing methods such as attributes restructuring and instant with missing value removal applied to the dataset, which contains 4412 instants comprising of 1784 INPATIENT class and 2628 OUTPATIENT class. Some analysis exercises were conducted to determine the characteristics of the dataset. Univariate analysis is performed to determine the distribution pattern of each variable based on instant classes. Fig. 2 depicted the distribution histogram of HAEMATOCRIT features of each class, whereas Fig. 3 presents the MCH distribution pattern. These figures show that, in general, the distribution of HAEMOTOCRIT and MCH features are relatively similar between IN_PATIENT and OUT_PATIENT classes. The HAEMATOCRIT feature of both classes is normally distributed with the dominant value between 25 to 50 and the MCH of 25 to 32/33.

The remaining seven features also present the most similar pattern shown by these four features. The detailed patterns of these features are presented in this article's appendix. The high degree of feature distribution patterns is similar for both classes because the identification class belongs to a certain instant. Therefore, the classification task for the laboratory test result dataset is challenging.

Another analysis applied to the data is a multivariate correlation, as shown in Fig. 4. The first three features, namely HAEMATOCRIT, HAEMOGLOBINS, and ERYTHROCYTE, are highly correlated. The other high correlation is presented by MCH vs. MCV feature, and from the class point of view, the correlation is roughly the same for both classes.

B. First Modeling Step

In the first modeling step, six classifier techniques, namely, Decision Tree (DT), Random Forest (RF), Gaussian Naïve Bayes, Ada Boost, Ada Boost with DT as the basic learner, and XGBoost were evaluated and applied to two kinds of data

representation. The Cross-Validation dataset splitting scenario was performed on the training and testing with KFold value of 10. The testing accuracy performance from the cross-validation experiments is shown in Tables IV and V. In general, for all techniques explored, the first format data represension presented better results compared to the second. This shows that the recoding value of the laboratory test made the data easier for human understanding but reduces accuracy.

For both formats used in data representation, XGBoost and AdaBoost obtained the best testing performance by average with slight differences. For the maximum value, the Random Forest had a testing accuracy of 0.7986, which is outperformed compared to AdaBoost with a maximum accuracy is 0.7964. XGBoost is still the best technique for the maximum results with a testing accuracy performance of 0.8009. Therefore, the second modeling step used both methods in the Format 1 data representation.

1) Second step of modelling: The GridsearchCV hyperparameter tuning of AdaBoost and XGBoost provided the performance results, as shown in Table VI. For both performance parameters, XGBoost achieved better results compared to Ada Boost. It also outperformed the training and testing steps with a different performance pattern. In the training stage, the difference between the accuracy and ROC-AUC performance parameter values between XGBoost and AdaBoost was 0.0671 and 0.0775, respectively, while at the testing phase, the difference was 0.0108 and 0.0166.
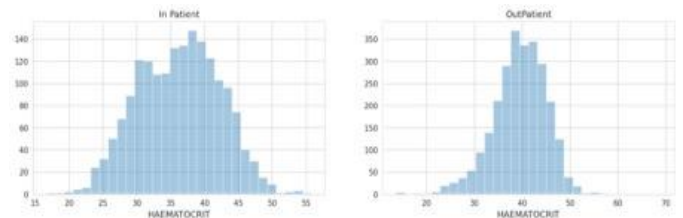

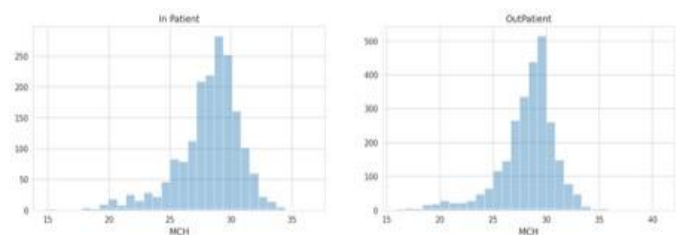Fig. 2. Hematocrits Feature Distribution for Two Classes.

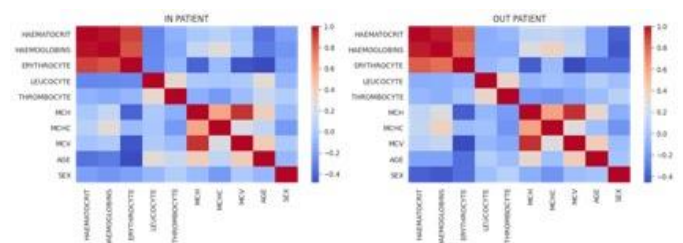Fig. 3. MCH Feature Distribution for Two Classes.

Fig. 4. Multivariate Analysis of Two Classes.

TABLE IV. TESTING ACCURACY PERFORMANCE OF FORMAT 1 DATA REPRESENTATION

	DT	RF	GaussianNB	AdaBoost	AdaBoost-DT	XGBoost
Max.	0.7014	0.7986	0.7392	0.7964	0.6825	0.8009
Min.	0.5646	0.6327	0.6131	0.6448	0.5692	0.6576
Ave.	0.6339	0.7158	0.6895	**0.7208**	0.6382	**0.7359**

TABLE V. TESTING ACCURACY PERFORMANCE OF FORMAT 2 DATA REPRESENTATION

	DT	RF	GaussianNB	AdaBoost	AdaBoost-DT	XGBoost
Max.	0.6281	0.6946	0.7308	0.7805	0.6719	0.7959
Min.	0.4751	0.5737	0.5737	0.4807	0.5510	0.6009
Ave.	0.5757	0.6335	0.6546	**0.6847**	0.6031	**0.6969**

TABLE VI. PERFORMANCE OF GRIDSEARCHCV HYPERPARAMETER TUNING

	Best Hyperparameter	Training		Testing	
		Accuracy	*ROC-AUC*	*Accuracy*	*ROC-AUC*
AdaBoost	n_estimators=230 learning_rate=0.1	0.7525	0.8145	0.7471	0.7936
XGBoost	learning_rate =0.1, n_estimators=1000, max_depth=5, min_child_weight=1, gamma=0, subsample=0.8, colsample_bytree=0.8, objective= 'binary:logistic', nthread=4, scale_pos_weight=1, seed=27)	**0.8196**	**0.8920**	**0.7579**	**0.8102**

Fig. 5 and Fig. 6 show the details of AdaBoost and XGBoost performance in ROC-AUC. The ROC curve shows that in any stage of specificity, XGBoost provides better AUC results compared to AdaBoost. Another information-insight shown by the curve is the different behaviour pattern of Training-Testing AUC for both classifiers. For XGBoost, Training – AUC is always on top of the Testing – AUC, whereas for AdaBoost in some part of curve Testing-AUC is the same.
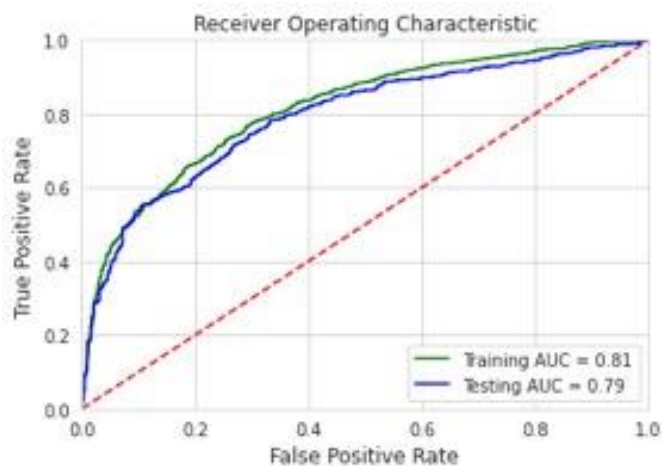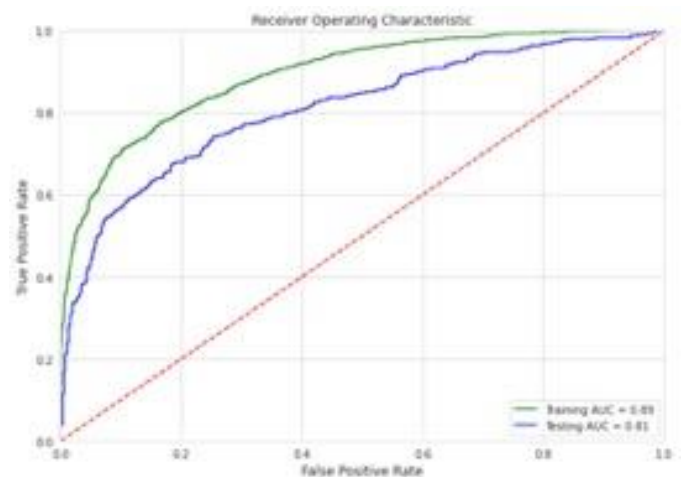

Fig. 6. XGBoost ROC-AUC Graph.

XGBoost classifier shows that THROMBOCYTE, AGE, and LEUCOCYTE are the top three factors of patient laboratory test results. Conversely, SEX is the least important attribute of patients, which contributes to the next treatment. The feature importance of patient data is shown in Fig. 7.
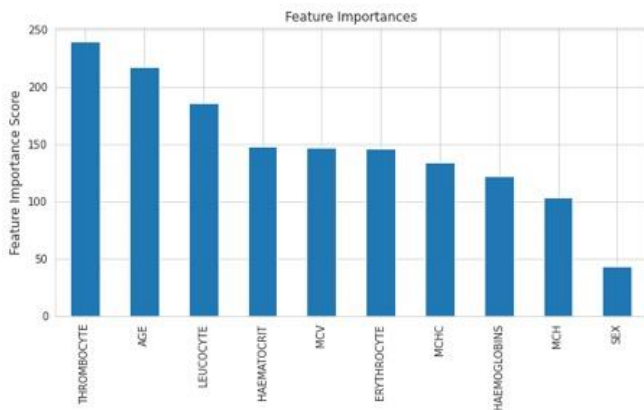

Fig. 5. AdaBoost ROC-AUC Graph.

Fig. 7.   Feature Importance by XGBoost Classifier.

V.   CONCLUSION

In this research, the EHR dataset was collected from a private hospital located in Jakarta Indonesia to predict patient treatment recommendations. The work is one of limited computational study on a health-medical domain performed based on Indonesia local context. Based on data analysis, it can be concluded that the instant data characteristics belong to each class can be used to determine the next patient treatment. However, since the characteristics are quite similar, this condition makes it difficult to identify and classify challenges manually. The study shows that the XGBoost technique provides the best performance in predicting the next treatment to patients based on their laboratory test results. Another experimental result showed that THROMBOCYTE, AGE, and LEUCOCYTE are the most dominant feature in determining the class of a certain instant data.

The best testing accuracy achieved in the experiment is 0.7579. However, this is not acceptable in the health-medical field, which is related to human life. Therefore, many studies need to be carried out to overcome the obstacles. The limited information utilized as the input of machine learning techniques is one of the barriers addressed. Therefore, the use of additional patients' data such as their medical record history has the ability to improve the quality of the model. Future studies need to be conducted with easy access to patient information.

REFERENCES

[1] M. Moscatelli et al., 'An infrastructure for precision medicine through analysis of big data', BMC Bioinformatics, vol. 19, no. Suppl 10, 2018. DOI: 10.1186/s12859-018-2300-5.

[2] A. Mavrogiorgou, A. Kiourtis, M. Touloupou, E. Kapassa, and D. Kyriazis, 'Internet of medical things (Iomt): Acquiring and transforming data into hl7 fhir through 5g network slicing', Emerg. Sci. J., vol. 3, no. 2, pp. 64–77, 2019.

[3] H.-Y. Liang et al., 'Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence', Nat. Med., vol. 25, 2019. DOI: 10.1038/s41591-018-0335-9.

[4] D. S. Kermany et al., 'Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning', Cell, vol. 172, no. 5, pp. 1122-1131.e9, 2018.

[5] D. Stamate et al., 'A metabolite-based machine learning approach to diagnose Alzheimer-type dementia in blood: Results from the European Medical Information Framework for Alzheimer disease biomarker discovery cohort', Alzheimer's Dement. Transl. Res. Clin. Interv., vol. 5, pp. 933–938, 2019.

[6] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, 'A data-driven approach to predicting diabetes and cardiovascular disease with machine learning', BMC Med. Inform. Decis. Mak., vol. 19, no. 1, pp. 1–15, 2019.

[7] M. Sharma and H. Aggarwal, 'Mobile based application for prediction of diabetes mellitus: FHIR Standard', Int. J. Eng. Technol., vol. 7, no. 2, pp. 117–120, 2018.

[8] W. Chang et al., 'Probability analysis of hypertension-related symptoms based on XGboost and clustering algorithm', Appl. Sci., vol. 9, no. 6, 2019. DOI: 10.3390/app9061215.

[9] X. Tian et al., 'Using machine learning algorithms to predict hepatitis B surface antigen seroclearance', Comput. Math. Methods Med., vol. 2019, 2019. DOI: 10.1155/2019/6915850.

[10] R. B. Lukmanto and E. Irwansyah, 'The Early Detection of Diabetes Mellitus (DM) Using Fuzzy Hierarchical Model', Procedia Comput. Sci., vol. 59, no. Iccsci, pp. 312–319, 2015.

[11] Suhaeri, N. M. Nawi, and M. Fathurahman, 'Early detection of Dengue disease using extreme learning machine', Int. J. Adv. Sci. Eng. Inf. Technol., vol. 8, no. 5, pp. 2219–2224, 2018.

[12] C. A. Hu et al., 'Using a machine learning approach to predict mortality in critically ill influenza patients: A cross-sectional retrospective multicentre study in Taiwan', BMJ Open, vol. 10, no. 2, 2020. DOI: 10.1136/bmjopen-2019-033898.

[13] X. Hu et al., 'Machine learning to predict rapid progression of carotid atherosclerosis in patients with impaired glucose tolerance', Eurasip J. Bioinforma. Syst. Biol., vol. 2016, no. 1, 2016. DOI: 10.1186/s13637-016-0049-6.C. Ye et al., 'Identification of elders at higher risk for fall with statewide electronic health records and a machine learning algorithm', Int. J. Med. Inform., vol. 137, no. February, 2020.

[14] S. Winiarti, H. Yuliansyah, and A. A. Purnama, 'Identification of Toddlers' nutritional status using data mining approach', Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 1, pp. 164–169, 2018.

[15] A. Shaham, G. Chodick, V. Shalev, and D. Yamin, 'Personal and social patterns predict influenza vaccination decision', BMC Public Health, vol. 20, no. 1, pp. 1–12, 2020.

[16] Kaneta Y, Tomida A, Tsuruo T, Nakamura Y, and Ohno R., 'Prediction of sensitivity to STI571 among chronic myeloid leukemia patients by genome-wide cDNA microarray analysis', Jpn J Cancer Res., vol. 93, no. 8, pp. 849–56., Aug. 2002.

[17] M. Sadikin, M. I. Fanany, and T. Basaruddin, 'A New Data Representation Based on Training Data Characteristics to Extract Drug Name Entity in Medical Text', vol. 2016, 2016.

[18] M. Skeppstedt, M. Kvist, G. H. Nilsson, and H. Dalianis, 'Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study.', J. Biomed. Inform., vol. 49, pp. 148–58, Jun. 2014.

[19] M. Sadikin, 'Mining relation extraction based on pattern learning approach', Indones. J. Electr. Eng. Comput. Sci., vol. 6, no. 1, 2017. DOI: 10.11591/ijeecs.v6.i1.pp50-57.

[20] L. Deléger, C. Grouin, and P. Zweigenbaum, 'Extracting medication information from French clinical texts', Stud. Health Technol. Inform., vol. 160, no. PART 1, pp. 949–953, 2010.

[21] Y.-F. Huang, H.-Y. Yeh, and V.-W. Soo, 'Network-based inferring drug-disease associations from chemical, genomic and phenotype data', 2012 IEEE Int. Conf. Bioinforma. Biomed., pp. 1–6, Oct. 2012.

[22] S. Saha and S. Rajasekaran, 'NRGC: A novel referential genome compression algorithm', Bioinformatics, vol. 32, no. 22, pp. 3405–3412, 2016.

[23] High-Throughput Variation Detection, and Genotyping Using Microarrays, David J. Cutler; Michael E. Zwick, and Minerva M. Carrasquillo,3, 'High-Throughput Variation Detection and Genotyping Using Microarrays', Genome Res, vol. 11, no. 11, pp. 1913–1925, 2001.

[24] P. R. M. Series, 'Contextual Factors: The importance of considering and reporting on the context in research on the patient centered medical home', 2013.

[25] F. Beenkens and P. Stolk, 'Context dependency of medical devices', 2010.

[26] D. Ramayanti et al., 'Tuberculosis Ontology Generation and Enrichment

Based Text Mining', in 2020 International Conference on Information Technology Systems and Innovation (ICITSI), 2020, pp. 429–434.

[27] H. Han, M. Huang, Y. Zhang, and J. Liu, 'Decision support system for medical diagnosis utilizing imbalanced clinical data', Appl. Sci., vol. 8, no. 9, 2018. DOI: 10.3390/app8091597.

[28] T. Chen and C. Guestrin, 'XGBoost: A scalable tree boosting system', Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-Augu, pp. 785–794, 2016.

[29] C. C. Chang and S. H. Chen, 'Developing a Novel Machine Learning-Based Classification Scheme for Predicting SPCs in Breast Cancer Survivors', Front. Genet., vol. 10, no. September, pp. 1–6, 2019.

[30] R. A. Taylor, C. L. Moore, K. H. Cheung, and C. Brandt, 'Predicting urinary tract infections in the emergency department with machine learning', PLoS One, vol. 13, no. 3, pp. 1–15, 2018.

[31] C. Gao et al., 'Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease', Sci. Rep., vol. 8, no. 1, pp. 1–21, 2018.

[32] Y. Wang, Z. Du, W. R. Lawrence, Y. Huang, Y. Deng, and Y. Hao, 'Predicting hepatitis b virus infection based on health examination data of community population', Int. J. Environ. Res. Public Health, vol. 16, no. 23, 2019. DOI: 10.3390/ijerph16234842.

[33] S. V. Murty and R. Kiran Kumar, 'Accurate liver disease prediction with extreme gradient boosting', Int. J. Eng. Adv. Technol., vol. 8, no. 6, pp. 2288–2295, 2019.

[34] J. Taninaga et al., 'Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study', Sci. Rep., vol. 9, no. 1, pp. 1–9, 2019.

[35] X. Mo et al., 'Early and accurate prediction of clinical response to methotrexate treatment in juvenile idiopathic arthritis using machine learning', Front. Pharmacol., vol. 10, no. October, pp. 1–11, 2019.

[36] Z. Qiao, N. Sun, X. Li, E. Xia, S. Zhao, and Y. Qin, 'Using machine learning approaches for emergency room visit prediction based on electronic health record data', Stud. Health Technol. Inform., vol. 247, pp. 111–115, 2018.

[37] P. Kumar, A. Nestsiarovich, S. J. Nelson, B. Kerner, D. J. Perkins, and C. G. Lambert, 'Imputation and characterization of uncoded self-harm in major mental illness using machine learning', J. Am. Med. Inform. Assoc., vol. 27, no. 1, pp. 136–146, 2020.

[38] G. N. Dimitrakopoulos, A. G. Vrahatis, K. Sgarbas, and V. Plagianakos, 'Pathway analysis using xgboost classification in biomedical data', in ACM International Conference Proceeding Series, 2018, no. July 2018. DOI: 10.1145/3200947.3201029.

[39] A. Mulisin, 'Hematokrit : Nilai Normal, Rendah, dan Tinggi', 2020. [Online]. Available: https://www.honestdocs.id/hematokrit. [Accessed: 18-Dec-2019].

[40] Y. Capriyanti, 'Bahayakah Bila Jumlah Leukosit Berlebihan - Tanya Alodokter Tanya Dokter Diskusi Terbaru', 2020. [Online]. Available: https://www.alodokter.com/komunitas/topic/lekosit. [Accessed: 18-Dec-2019].

[41] L. Buitinck et al., 'API design for machine learning software: experiences from the scikit-learn project', 2013.

[42] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, 'An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics', Inf. Sci. (Ny)., vol. 250, pp. 113–141, 2013.

[43] D. Bertsimas and J. Dunn, 'Optimal classification trees', Mach. Learn., vol. 106, no. 7, pp. 1039–1082, 2017.

[44] A. Cubero-Fernandez, F. J. Rodriguez-Lozano, R. Villatoro, J. Olivares, and J. M. Palomares, 'Efficient pavement crack detection and classification', Eurasip J. Image Video Process., vol. 2017, no. 1, 2017. DOI: 10.1186/s13640-017-0187-0.

[45] M. Al-Qatf, Y. Lasheng, M. Al-Habib, and K. Al-Sabahi, 'Deep Learning Approach Combining Sparse Autoencoder with SVM for Network Intrusion Detection', IEEE Access, vol. 6, pp. 52843–52856, 2018.

[46] A. M. Joshi and S. Prabhune, 'Random forest: A hybrid implementation for sarcasm detection in public opinion mining', Int. J. Innov. Technol. Explor. Eng., vol. 8, no. 12, pp. 5022–5025, 2019.

[47] Z. L. Li, H. X. Wang, Y. W. Zhang, and X. H. Zhao, 'Random forest–based feature selection and detection method for drunk driving recognition', Int. J. Distrib. Sens. Networks, vol. 16, no. 2, 2020. DOI: 10.1177/1550147720905234.

[48] Y. Mishina, R. Murata, Y. Yamauchi, T. Yamashita, and H. Fujiyoshi, 'Boosted random forest', IEICE Trans. Inf. Syst., vol. E98D, no. 9, pp. 1630–1636, 2015.

[49] N. Azizah, L. S. Riza, and Y. Wihardi, 'Implementation of random forest algorithm with parallel computing in R', J. Phys. Conf. Ser., vol. 1280, no. 2, 2019. DOI: 10.1088/1742-6596/1280/2/022028.

[50] W. Zhang, X. Zhao, and Z. Li, 'A Comprehensive Study of Smartphone-Based Indoor Activity Recognition via Xgboost', IEEE Access, vol. 7, pp. 80027–80042, 2019.

[51] M. C. Belavagi and B. Muniyal, 'Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection', in Twelfth International Multi-Conference on Information Processing-2016 (IMCIP-2016), 2016, vol. 89, pp. 117–123.

[52] A. Al Bataineh, 'A comparative analysis of nonlinear machine learning algorithms for breast cancer detection', Int. J. Mach. Learn. Comput., vol. 9, no. 3, pp. 248–254, 2019.

[53] Y. Freund and R. E. Schapire, 'A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting', J. Comput. Syst. Sci., vol. 55, no. 1, pp. 119–139, 1997.

[54] S. J. Lee, T. Chen, L. Yu, and C. H. Lai, 'Image Classification Based on the Boost Convolutional Neural Network', IEEE Access, vol. 6, pp. 12755–12768, 2018.

[55] Z. Yang, L. Xu, Z. Cai, and Z. Xu, 'Re-scale AdaBoost for attack detection in collaborative filtering recommender systems', Knowledge-Based Syst., vol. 100, pp. 74–88, 2016.