[1] **Husni Iskandar Pohan***

[2] **Sutoto**

[3] **Yaya Heryadi**

[4] **Harjanto Prabowo**

The Effect of Combined Synthetic Tabular Data Generated Using CTGAN Model with Actual Data on Performance of DHF, Varicella, and COVID-19 Recognition Model

Journal of Electrical Systems

Abstract: - There are several quickly spreading illnesses such as DHFs spread by mosquitoes, COVID-19 spreads through respiratory droplets and contact with contaminated surfaces, and Varicella spreads by direct touch. The transmission rate of these diseases can be reduced if medical services can identify them early. However, the performance of the prediction model based on the machine learning approach is limited by the availability of labelled patient datasets. This study showed some empirical evidence of the use of synthetic data generated using actual medical records as the basis to improve the performance of the prediction model. The empirical results showed that the Decision Tree algorithm which is trained using a mixed synthetic and actual dataset can achieve 91.98% average accuracy which is higher than model performance which is trained using real dataset only. The results of model interpretation using Shapley Additive Explanations have the advantage of measuring the overall dominant features and indicating that the top five most important features are Thrombocyte, Temp, Cough, Spot, and Nauseous .

Keywords: Decision Tree, CTGAN, SHapley Additive explanations.

I. INTRODUCTION

The pandemic disaster, especially COVID-19, in recent years has led to the emergence of various solutions from different fields. In the healthcare sector, for instance, solutions include vaccines and test kits (antigen, PCR). In the pharmaceutical field, there are antiviral medications, and in the manufacturing industry, healthcare equipment such as odor detectors and ventilators have been developed. Together with Varicella and DHF, these diseases typically have a geographical proximity in terms of transmission impact [1].

Anticipating contagious diseases based on geographic clusters requires real carefulness in high-density locations. There are several quickly spreading illnesses within the same geographic areas such as DHFs spread by mosquitoes, COVID-19 spreads through respiratory droplets and contact with contaminated surfaces, and Varicella spreads by direct touch. The transmission rate of these diseases can be reduced if medical services are able to identify them early. For example, patients may be referred to healthcare facilities that have greater resources available to them, and geographic isolation measures may be put in place right away if needed. Despite the importance of early prediction, the availability of label patient datasets is typically limited which affects the performance of the prediction model based on the machine learning approach.

This study showed some empirical evidence of the use of synthetic data generated using actual medical records as the basis to improve the performance of the prediction model. The remainder of this paper is organized as follows. Section 2 describes several studies related to this work. Next, Section 3 describes the research model used in this study followed by Section 4 which explains the experimentation findings. Finally, Section 5 concludes this paper.

II. RELATED WORKS

In an attempt to prevent the spread of transmitted diseases, many governments have implemented several regulations including strict control of incoming and outgoing human traffic, health screening procedures at airports, and the establishment of quarantine facilities. This also encompasses health protocols enforced in public places such as schools, malls, and cinemas.

The solutions from the field of information technology that recently emerged involve developing various prediction models using machine learning technology. Researchers are interested in exploring potential solutions in the form of early detection, specifically to help mitigate the potential spread of cluster-based diseases. In general, these solutions can assist policymakers in the healthcare domain. In particular, the advent of machine learning technology in the past decade has motivated many researchers to apply machine learning models to

[1]Computer Science Department, BINUS Graduate Program-Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480, husni.pohan@binus.ac.id
[2]Indonesian Hospital Accreditation Commission, Jakarta, Indonesia Jakarta, Indonesia 12960, sutoto@kars.or.id
[3]Computer Science Department, BINUS Graduate Program-Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia, 11480, yayaheryadi@binus.edu
[4]Management Department, BINUS Business School Doctor of Research in Management Jakarta, Indonesia 11480, harprabowo@binus.edu
*Corresponding Author: husni.pohan@binus.ac.id

develop robust classification models as early screening tools to help hospitals and clinics reduce the risk of health workers being infected by transmitted diseases from patients. However, developing classification models requires a large amount of labeled medical records which are hardly obtained for various reasons. The primary reason for such label data scarcity is the patient data privacy protection policy implemented by hospitals and health care. Hence, developing a robust classification model to recognize the mentioned diseases is challenging.

Having been motivated by the scarcity of labeled data and the protection of data privacy have increased researcher interest in many fields toward synthetic data generating methods, particularly on machine learning model development to address tasks in the medical field. Many studies showed evidence that synthetic data can accelerate the development of machine learning models for solving numerous downstream tasks such as classification and regression.

According to El Emam [2] , conceptually synthetic data are data that has been made from actual data with the same statistical features as the original data, but it is not genuine data itself. This implies that, as a measure of utility, an analyst working with a synthetic dataset should obtain analysis results that are comparable to those obtained with actual data.

Despite many prominent studies having been reported on this topic particularly to recognize DHF, Varicella, and COVID-19 diseases from variables observed or measured by patients, the aim of this study has two folds namely developing a robust model to recognize each of these diseases and overcoming the limitation of available medical records.

The general benefits of this research are to provide insights into dominant features related to early detection solutions for cluster-based communicable diseases for healthcare workers. The identification of these dominant features is expected to reduce the risk of transmission by minimizing contact between patients with communicable diseases and non-communicable diseases. Additionally, the overall benefit is to assist the government in minimizing risks for healthcare workers at entry-level healthcare facilities, which can have an impact on the paralysis of national healthcare services.

III. RESEARCH METHODOLOGY

A. Research Frameworks

The framework of this study can be summarized in Figure 1. As can be seen, this study comprises several steps. First, feature extraction. Data input for this study is collected from the medical record database with permission from the clinic owners based on agreed terms and conditions.

Second, data preprocessing comprises several steps aimed at making sure the data used for analysis or machine learning is valid, accurate, dependable, and appropriate for the intended use known as data validation in the field of data science. This step helps avoid biases and mistakes in further data analysis and is an essential stage in the data pretreatment workflow. Several methods are used in this study including

- Ensuring that the data types of each column match the expected types,
- Checking that numerical values fall within an expected range,
- Verifying that the data sample is consistent across different columns or datasets,
- Ensuring that there is no missing value in essential columns,
- Identifying anomalies or outliers in the data that may be errors or significant observations, and
- Addressing data quality issues by cleaning, imputing missing values, correcting errors, and handling outliers as needed.

Third, feature selection which is categorized into predictor and target features. The predictor and target features are summarized in Table 1. The predictor features are selected among data features that have strong importance in predicting the target feature.

Finally, model training and optimization, model performance evaluation, and model interpretation. In this study, the classification models that are explored are decision tree models. Predictions for a specific label are determined based on the value in the leaf node, while estimates for numeric values are obtained by averaging the values at the leaf nodes [3],[4]. The model selection is based on the following criteria:

- Interpretability: Healthcare services that use the model can understand which features contribute the most to the model's predictions and justify the prediction results based on the input data provided,
- Good Performance and Accuracy: The model's prediction results do not have significant deviations from the actual categories (actual categories), and

- No Overfitting: the model can generalize well from patterns in the training data, enabling it to recognize patterns in new data (test data). Whilst, model interpretation methods used in this study are SHAP methods.

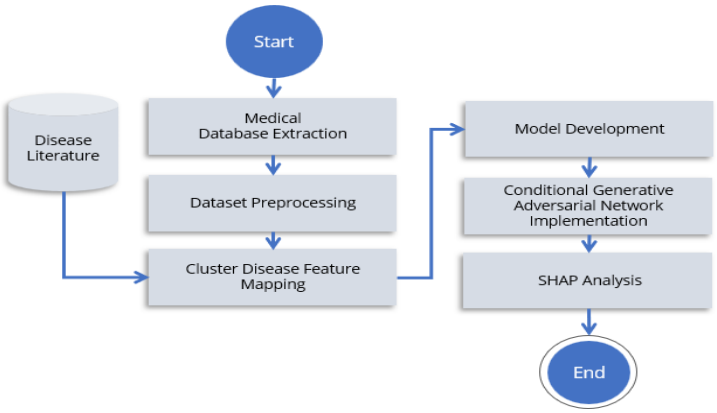

Figure 1. Research Framework

B. Input Dataset

Input data for this study is a set of medical records from January 1st, 2019 until December 31st, 2021 from a healthcare clinic in Bandung, Indonesia, used with permission. The input data comprises of 68,666 records in which 2,678 records comprise of patient data which have been infected with Covid-19 patient records, Varicella, or Dengue. The data features can be summarized using Table 1 as follows.

For developing classification model, two datasets are prepared namely: dataset-A contains 5,356 records (samples) of duplicated real data, and dataset-B contains 5,356 records (samples) comprises of real and synthetic data with proportion 50:50.

In this study, decision tree is trained as a classification model using hold-out cross-validation techniques. Each of the dataset is used to train the model to measure the effect of the synthetic data to performance of the classification model. In the model training, the input dataset was split randomly into training and testing dataset with proportion 80:20. Two variables were used as target variable namely: Code and Referral.

Table 1. Extracted Variables From Patient Data.

Field	Scale	Type	Source	Sample
Referral	Nominal	Categorical	Diagnosis	REFERRAL/NON
Code	Nominal	Categorical	Diagnosis	COV/DHF/VAR
Spot	Nominal	Categorical	Anamnesis	Y/N
Red	Nominal	Categorical	Anamnesis	Y/N
Congested	Nominal	Categorical	Anamnesis	Y/N
Cough	Nominal	Categorical	Anamnesis	Y/N
Flu	Nominal	Categorical	Anamnesis	Y/N
Feverish	Nominal	Categorical	Anamnesis	Y/N
Stomach	Nominal	Categorical	Anamnesis	Y/N
Nauseous	Nominal	Categorical	Anamnesis	Y/N
Vomit	Nominal	Categorical	Anamnesis	Y/N
Dizzy	Nominal	Categorical	Anamnesis	Y/N
Itchy	Nominal	Categorical	Anamnesis	Y/N
Swallow	Nominal	Categorical	Anamnesis	Y/N
Blister	Nominal	Categorical	Anamnesis	Y/N
Sore	Nominal	Categorical	Anamnesis	Y/N
Weak	Nominal	Categorical	Anamnesis	Y/N
Rheumatic	Nominal	Categorical	Anamnesis	Y/N
Pain	Nominal	Categorical	Anamnesis	Y/N
Cold	Nominal	Categorical	Anamnesis	Y/N
Fever	Interval	Numerical	Med Examination	37.2
Temp	Interval	Numerical	Med Examination	110,000
Thrombocyte				

Source: Prepared by the author, (2024)

All features were chosen with the consideration that their results can be obtained within one hour to ensure that the disease prediction process can minimize contact between patients and others.

The Code feature represents the diagnosis results performed by healthcare personnel and serves as the label in this training data. The Referral feature represents the doctor decision based on medical examination whether a respective patient can be treated in internal clinic or will be referred to a referenced hospital.

According to [5], the distinctive characteristics of the three diseases chosen as the research objects are used as references for determining keywords that will be utilized in the medical history data. These characteristics are summarized in Table 3.

Following [6], the task of synthetic data generation process can be formulated as follows. Given a tabel T as input data, a data synthesizer G is trained to learn from the table T and then using G to generate a synthetic table where each column is considered to be a random variable. These random variables follow an unknown joint distribution A sample represented by a row is one observation from the joint distribution. To train the G model, the input table T is split into two random subsets: training dataset and testing dataset. Using the trained model G, synthetic data T_syn is constructed by independently sampling rows using G.

C. Conditional Generative Adversarial Network

Conditional Generative Adversarial Model (CTGAN) is a state-of-the-art tabular data modeling proposed by [7] in which the data consists of structured data organized into rows and columns, such as spreadsheets or relational databases. This model can be viewed as a specialized type of the Generative Adversarial Network (GAN) model proposed by [8]. The CTGAN model is designed to address the task of modeling the probability distribution of rows in tabular data and generating realistic synthetic data.

Table 2. Characteristic Of The Disease

Desc	DHF	VAR	COV
Etiology	Dengue	Varicella Zooster	Corona Virus
Transmission medium	Mosquito	Close Contact	Close Contact
Anamnesis	Sluggish	Fever	Cough
	Muscle Ache	Skin Damage	Fever
	Skin Red Spot	Malaise	Out of Breath
	Sudden High Fever		Swallowing Pain
	Signs of Bleeding in The Nose		Diarrhea and Vomiting
Physical Examination	Epistaxis	Scaldhead	Saturation down
	Melena	Tear Drops	Febris
	Heartburn / Nausea	Vesicle Erythematous Papules	Takipnea
	Hematuria		Dull
	Flushing		Conjunctivitis
	Ptechiae		Ronchi
			Epigastrium
			Rhinofaringitis
Supporting Investigation	NS1 Test	PCR Test	PCR Swab Test
	Platelets Test	Tzanck Test	Antigen Swab Test
	Hematocrit Test	Serological Test	
	HB Leukocytes Test		
	Serological Test		
	Leukocytes Type Count Test		

Source: Prepared by the author, (2024)

The novelty of CTGAN model, among others, are involving several new techniques: augmenting the training procedure with mode-specific normalization, architectural changes, and addressing data imbalance by employing a conditional generator and training-by-sampling. With its ability to produce artificial data that is statistically close to actual tabular data, CTGAN is particularly useful for a range of applications, including data augmentation, privacy protection, and machine learning model testing.

The term ‘synthetic data’ refers to data that is created intentionally and imitates the traits and trends of real-world data, but is not based on genuine observations or measurements. It is produced without the presence of sensitive or personally identifiable information using a variety of statistical and computational techniques to mimic actual data. Synthetic data is frequently employed to overcome constraints like data sharing while adhering to privacy and regulatory requirements. The main purpose of synthetic data generating are as follows: privacy protection, data sharing, model development and testing, data augmentation, benchmarking, and simulation. The prominent methods for synthetic data generating is typically treating each column in a table as a random variable, modeling a joint multivariate probability distribution, and then sampling from that distribution. The quality of the generated synthetic data is significantly limited by the probability distribution models due to limitations in the types of distributions and computational problems. Unlike these prominent methods, CTGAN model (see Figure 2) uses novel approach to generating synthetic tabular data using conditional generation capabilities, specialized loss functions, and data preprocessing techniques.

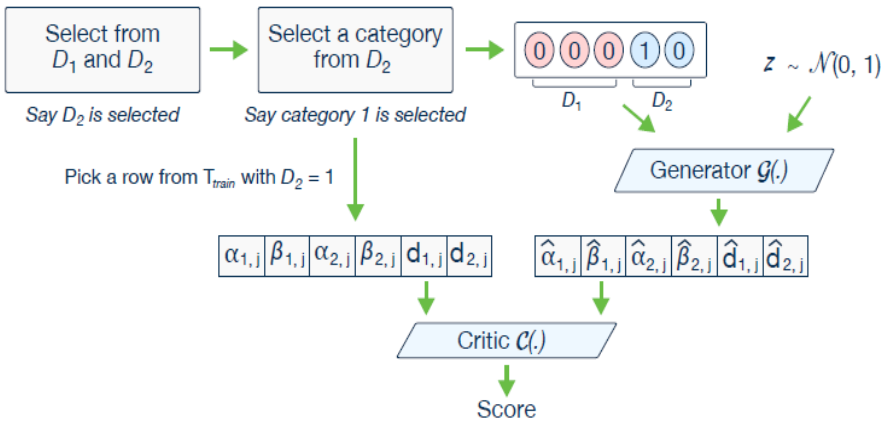

Figure 2. Characteristic Of The Disease

IV. RESULT AND DISCUSSION

A. Training and Testing Results

The results of model training using each input datasets can be summarized in the Tabel 3 and Table 4 as follows

Table 3.The Effect Of Synthetic Data To Performance Of Decision Tree Model w/ Code As Label

Real Data w/ Code as Label				
Accuracy = 85.81%	true DHF	true VAR	true COV	class precision
pred. DHF	521	63	21	86.12%
pred. COV	20	335	39	85.03%
pred. VAR	1	8	63	87.50%
class recall	96.13%	82.51%	51.22%	
Synthetic Data w/ Code as Label				
Accuracy = 85.81%	true DHF	true VAR	true COV	class precision
pred. DHF	3	0	0	100.0%
pred. COV	432	552	84	51.69%
pred. VAR	0	0	0	0.00%
class recall	0.69%	100.00%	0.00%	

Table 4.The Effect Of Synthetic Data To Performance Of Decision Tree Model W/ Referral As Label

Real Data w/ Referral as Label			
Accuracy = 85.81%	true non referral	true referral	class precision
pred. Non Referral	917	150	85.94%
pred. Refferal	4	0	0.00%
class recall	99.57%	0.00%	

Synthetic Data w/ Referral as Label			
Accuracy = 85.81%	true non referral	true referral	class precision
pred. Non Refferal	986	86	91.98%
pred. Refferal	0	0	0.00%
class recall	100.00%	0.00%	

Source: Prepared by the author, (2024)

From the two tables above, it can be concluded that the use of synthetic data produced by the CTGAN model in the case of Referral as a label can improve the performance of the decision tree model as a classification model, but on the contrary, in the case of Code as a label, the model performance actually decreases.

B. Trained Model and Interpretation

In this study, model interpretation methods used SHAP. The SHAP (SHapley Additive exPlanations) a machine learning method widely used for model interpretability and explaining the predictions provided by complex models.

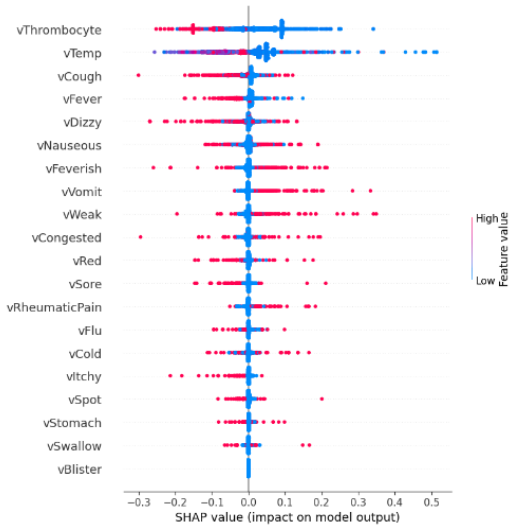

Figure 3. BEESWARM CODE AS LABEL

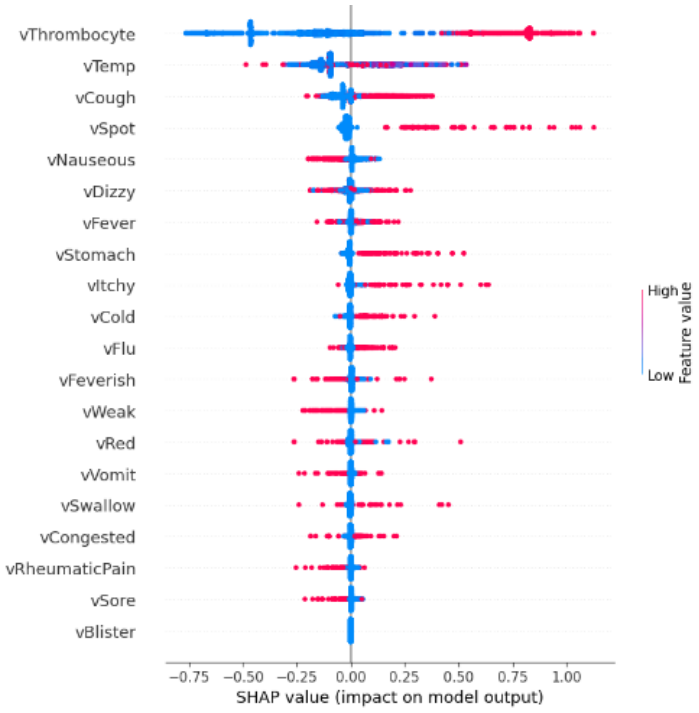

Figure 4. BEESWARM REFFERAL AS LABEL

It can be seen that there is a difference in the order of the dominant features of the two labels, especially for Fever and below, even though the first 3 features are the same. The role of Thrombocyte is more dominant in the Referral label. Meanwhile Temp has a more dominant role in the Code label.

In general, from the overall data:

- Thrombocyte is the most influential feature in diagnosing patients' diseases. The higher the SHAP value of the Thrombocyte feature, the greater its contribution to a specific model's prediction result.
- The contribution of the Spot feature is only significant if its SHAP value is large. If the SHAP value is low, the Spot feature is not dominant for the model's prediction.
- Other features (Temp, Dizzy, Nauseous, Fever, Rheumatic Pain) are generally not strong features in predicting a specific disease.

The conclusion from the analysis results with SHAP using Beeswarm visualization is as follows:

- Thrombocyte is the most significant feature (indicator) of DHF. However, it is sometimes accompanied by Red Spots.
- Temp and Cough are the most significant features (indicators) of Covid.
- Nauseous is a significant feature (indicator) of both DHF and Covid.
- Watery Spot is the most significant feature (indicator) of Varicella

V. CONCLUSIONS

From the experimental finding, is can be concluded that the CTGAN model for synthetic data can increase performance of the decision tree model as a classification model, but it does not apply to all cases.

The use of synthetic model for generated data is very crucial for developing predictive model especially in medical domain where availability of the large dataset for training model is limited. A mixture of synthetic and real data sets using referral as a label with a result of 91.98%, can predict the referral target variable which has practical utility for use in health services in protecting health workers and improving health services. However, this does not apply to Code as a label.

CONFLICTS OF INTEREST

The authors declares there is no conflict of interest.

FUNDING

None.

REFERENCES

[1] Dianbo, “Real-Time Forecasting of the COVID-19 outbreak in Chinese provinces: Machine learning approach using novel digital data and estimates from mechanistic models,” *J. Med. Internet Res*, vol. 22, no. 8, p. e20285, Aug. 2020, doi: 10.2196/20285.

[2] K. E. Emam, L. Mosquera, and R. Hoptroff, *Practical synthetic data generation: Balancing Privacy and the Broad Availability of Data*. “O’Reilly Media, Inc.,” 2020.

[3] M. G. Guzmán, D. J. Gubler, A. Izquierdo, É. Martinez, and S. B. Halstead, “Dengue infection,” *Nat. Rev.. Dis Prim*, vol. 2, no. 1, Aug. 2016, doi: 10.1038/nrdp.2016.55.

[4] M. Tayarani. N, “Applications of artificial intelligence in battling against covid-19: A literature review,” *Cha. Sol. Frac*, vol. 142, p. 110338, Jan. 2021, doi: 10.1016/j.chaos.2020.110338.

[5] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, “Modeling Tabular data using Conditional GAN,” *arXiv (Cornell University)*, Jan. 2019, doi: 10.48550/arxiv.1907.00503.

[6] Goodfellow *et al.*, “Generative adversarial networks,” *Adv. Neural Inf. Process. Syst.*, Jan. 2014, doi: 10.48550/arxiv.1406.2661.

[7] I. C. Sari and Y. Ruldeviyani, “Sentiment Analysis of the Covid-19 Virus Infection in Indonesian Public Transportation on Twitter Data: A Case Study of Commuter Line Passengers,” *Int. Work. Big Inf. Secur*, Oct. 2020, doi: 10.1109/iwbis50925.2020.9255531.

[8] D. Mienye, Y. Sun, and Z. Wang, “Prediction performance of improved decision tree-based algorithms: a review,” *Proc Manu*, vol. 35, pp. 698–703, Jan. 2019, doi: 10.1016/j.promfg.2019.06.011.