J-Icon : Jurnal Informatika dan Komputer
Vol. 13 No. 2, October 2025, pp. 60~67
DOI: 10.35508/jicon.v13i2.23729

CONSTRUCTING A DATASET FOR INFECTIOUS DISEASE PREDICTION AND SPATIAL CLUSTER ANALYSIS

Husni Iskandar Pohan[1*]

[1*]School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
Email[*]: husni.pohan@binus.ac.id

ABSTRACT

This study presents a structured methodology for developing a custom dataset from patient visit records collected between January 1, 2019, and December 31, 2021, at a healthcare facility in Bandung Regency, Indonesia. The raw medical records were transformed into a machine learning–ready dataset through processes such as feature extraction, labeling, and geospatial enrichment using longitude and latitude coordinates. Personally identifiable information was removed, and clinical symptoms were standardized into structured variables to support both supervised and unsupervised learning tasks, including disease classification, referral prediction, and spatial cluster detection. The final dataset consisted of 1,015 Covid cases (COV), 1,356 Dengue cases (DHF), and 308 Varicella cases (VAR). It has been applied in advanced experiments involving feature importance analysis with SHAP and LIME, geospatial clustering, and synthetic data generation to address privacy and data availability concerns. This methodology is designed to support future research in healthcare analytics and the development of decision support systems and public health planning tools. However, since the dataset was constructed using records from a single healthcare facility in Bandung, the findings and patterns identified may not be generalizable to other regions that could exhibit different disease trends or healthcare-seeking behaviors.

Keywords: Covid, Dengue, Varicella, Dataset, Cluster

1. INTRODUCTION

The Department of Health places special attention on diseases whose transmission is based on geographic clustering. Unlike heart disease and diabetes, which are typically experienced individually, cluster-based diseases can spread rapidly among patients located in close geographic proximity. One of the challenges in working with such datasets is that the available data is not always in a ready-to-use format. Therefore, it is necessary to develop an approach to understand the data and transform it into a form suitable for processing with machine learning algorithms. This requires a solid understanding of the characteristics of the diseases involved—in this case, Covid (hereafter called COV), Dengue (hereafter called DHF), and Varicella (hereafter called VAR). The objective of this study is to develop a structured methodology for transforming raw patient visit records into an anonymized, machine learning–ready dataset enriched with spatial attributes. By focusing on three infectious diseases—COV, DHF, and VAR—this research aims to support predictive modeling and geospatial cluster analysis in a clinical context.

The contribution of this study lies in the creation of a publicly reproducible framework for preparing healthcare datasets with both clinical and spatial features. This includes processes such as disease labeling, referral prediction, and geolocation tagging—key components often omitted or only partially addressed in prior works. Furthermore, the resulting dataset can facilitate subsequent research involving model interpretability techniques such as SHAP and LIME, as well as synthetic data generation using CTGAN, thus expanding its applicability in advanced analytical and privacy-preserving contexts.

Research Gap

This study addresses a notable research gap: the scarcity of integrated datasets and methodologies that combine symptom-level clinical data with geographic coordinates for infectious disease modelling in low- and middle-income country settings. Most existing works either focus solely on classification without spatial awareness or require pre-cleaned datasets. In contrast, this research provides end-to-end guidance—from raw data ingestion to usable analytic datasets—while highlighting technical and ethical considerations in real-world implementation.

***) Corresponding Author**

Submitted : July 15, 2025
Accepted : August 5, 2025
Published : August 12, 2025

The cause of COV is the SARS (Severe Acute Respiratory Syndrome) virus, which spreads through droplets from coughing or sneezing. Transmission can also occur through direct interaction with infected individuals or by touching contaminated surfaces. Common symptoms include fever, dry cough, and shortness of breath. In some cases, it can also lead to fatigue, loss of smell, and diarrhea.

DHF is caused by a virus transmitted through the bite of the Aedes Aegypti mosquito. Symptoms may include high fever, headache, pain behind the eyes, and joint or muscle pain. In more severe cases, it can cause bleeding in the nose, gums, or internal organs, which may lead to fatal outcomes. Based on its transmission pattern, DHF cases typically increase during the rainy season. The Figure 1 illustrates the virus transmission cycle between humans and mosquitoes.

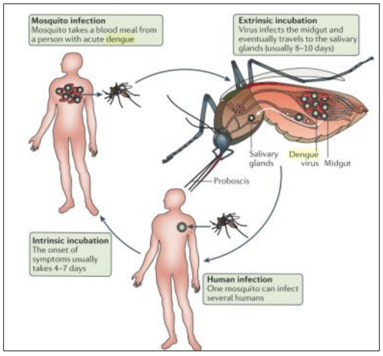
Mosquito infection
Mosquito takes a blood meal from a person with acute dengue
Extrinsic incubation
Virus infects the midgut and eventually travels to the salivary glands (usually 8–10 days)
Salivary glands
Dengue virus
Midgut
Proboscis
Intrinsic incubation
The onset of symptoms usually takes 4–7 days
Human infection
One mosquito can infect several humans

Figure 1. DHF Transmission Cycle [1]

VAR is caused by the Varicella-Zoster virus and is transmitted through fluid from skin lesions or blisters. Early symptoms may include fever, fatigue, loss of appetite, skin rashes, and red spots, which later develop into fluid-filled itchy vesicles that eventually dry and form scabs. It can be fatal for individuals with weakened immune systems. VAR has received relatively little attention as a contagious disease from both regional and global public health perspectives, especially in low- to middle-income countries.

In Indonesia—a developing country with limited public awareness and low vaccination coverage for VAR—cases of VAR remain common. In contrast to countries like the United States and China, where detailed statistical data on VAR is available, Indonesia still lacks sufficient epidemiological data on VAR [2]. The Characteristics of The Three Diseases is shown in the Table 1.

Table 1. Characteristics of The Three Diseases.

Aspect	**DHF** [3]	**VAR** [2]	**COV** [4]
Etiology	Dengue Virus	Varicella-Zoster Virus	Coronavirus
Transmission Medium	Mosquito (Aedes aegypti)	Close Contact	Close Contact
Anamnesis *(Nurse)*	Fatigue	Fever	Cough
	Muscle pain	Skin damage	Fever
	Red skin spots	Malaise	Shortness of breath
	Sudden high fever		Sore throat
	Nosebleeds		Diarrhea and vomiting
Physical Exam *(Doctor)*	Epistaxis (nosebleed)	Scabs (scaldhead)	Low oxygen saturation
	Black stool (melena)	Tear drop lesions	High fever (febrile)
	Upper abdominal pain/nausea	Vesicular skin eruption	Rapid breathing (tachypnea)
	Bloody urine (hematuria)		Diminished breath sounds
	Facial flushing		Red eyes (conjunctivitis)
	Petechiae (skin rash)		Rattling breath sounds
			Moderate to severe condition
			Epigastric tenderness
			Nose/throat infection
Supporting Tests *(Lab)*	NS1 Test	PCR Test	PCR Swab Test

	Platelet Count Hematocrit Test Hemoglobin/Leukocy Test Serological Test Leukocyte Type Count	Tzanck Test Serological Test	Antigen Swab Test

After identifying the characteristics of the three diseases, a total of 68,666 patient records in SQL Server format were first evaluated. The original data format is shown in the Table 2.

Table 2. The Original Data Format

No	Description	Data Type	Example
1	Visit ID	Varchar(15)	KJ3100251616
2	Visit Date	Datetime	2021-12-31 13:58:28.000
3	Diagnosis	Varchar(100)	Others
4	Patient ID	Varchar(9)	00-061228
5	Patient Name	Varchar(50)	Johny Iskandar
6	Patient Address	Varchar(100)	PBB IV C-71
7	Date of Birth	Datetime	2019-06-05 08:09:06.000
8	Gender	Varchar(9)	Male
9	Anamnesis	Varchar(500)	Fever 3 days, Cough, Flu
10	Physical Examination	Varchar(250)	Temp 38.9 ISPA BP DHF
11	Body Weight	Varchar(50)	11
12	Hemoglobin	Varchar(50)	13,7
13	Leukocytes	Varchar(50)	2,100
14	Platelets	Varchar(50)	70,000
15	Cholesterol	Varchar(50)	220
16	Triglycerides	Varchar(50)	180
17	Uric Acid	Varchar(50)	6.8
18	Fasting Blood Glucose	Varchar(50)	400
19	2-Hour Postprandial Glucose	Varchar(50)	100
20	Random Blood Glucose	Varchar(50)	331
21	Other Information	Varchar(250)	HT:37
22	Radiology Result	Varchar(50)	Atelaktasis
23	ECG Result	Varchar(50)	DBN
24	ECHO Result	Varchar(50)	Nephrolithiasis Dexta
25	Therapy	Varchar(150)	Lab Norages ZenirexErfasal
26	Wound Treatment (Yes/No)	Bit	0/1
27	Sutures (Yes/No)	Bit	0/1
28	Physiotherapy (Yes/No)	Bit	0/1
29	Nebulizer Treatment(Yes/No)	Bit	0/1
30	Additional Lab Info (Yes/No)	Bit	0/1
31	Additional Lab Notes	Varchar(150)	Desc
32	Active Visit Entry (Yes/No)	Bit	0/1
33	Age at Visit	Varchar(50)	2 Years, 6 Month, 28 Days
34	Action Description	Varchar(50)	Others

The conversion process must remove any patient-identifying information to comply with medical confidentiality regulations. Once the conversion is completed, the structure of the prepared final dataset for storing transactional data is organized as shown in the following Table 3.

Table 3. Final Dataset

Field	Description	Scale	Type	Source	Sample
VisitDate	Visit Date	Interval	Numerical	Registration	01/12/2021
Longitude	Coordinate	Interval	Numerical	Registration	10.767.125
Latitude	Coordinate	Interval	Numerical	Registration	-697.496
vCode	Disease Code	Nominal	Categorical	Diagnosis	COV/DHF/ VAR
vReferral	Referred or Not	Nominal	Categorical	Therapy	Y/N
vSpot	Presence of Spots	Nominal	Categorical	Anamnesis	Y/N
vRed	Redness Signs	Nominal	Categorical	Anamnesis	Y/N
vCongested	Shortness of Breath	Nominal	Categorical	Anamnesis	Y/N
vCough	Presence of Cough	Nominal	Categorical	Anamnesis	Y/N
vFlu	Presence of Flu	Nominal	Categorical	Anamnesis	Y/N
vFeverish	Feverish Sensation	Nominal	Categorical	Anamnesis	Y/N
vStomach	Stomach Issues	Nominal	Categorical	Anamnesis	Y/N
vNauseous	Nausea	Nominal	Categorical	Anamnesis	Y/N
vVomit	Vomiting	Nominal	Categorical	Anamnesis	Y/N
vDizzy	Dizziness	Nominal	Categorical	Anamnesis	Y/N

vItchy	Itching	Nominal	Categorical	Anamnesis	Y/N
vSwallow	Difficulty Swallowing	Nominal	Categorical	Anamnesis	Y/N
vBlister	Blisters or Lesions	Nominal	Categorical	Anamnesis	Y/N
vSore	Body Aches	Nominal	Categorical	Anamnesis	Y/N
vWeak	Weakness	Nominal	Categorical	Anamnesis	Y/N
vRheumaticPain	Joint/Muscle Pain	Nominal	Categorical	Anamnesis	Y/N
vCold	Runny Nose	Nominal	Categorical	Anamnesis	Y/N
vFever	Presence of Fever	Nominal	Categorical	Anamnesis	Y/N
vTemp	Body Temperature	Interval	Numerical	Examination	37.2
vThrombocyte	Thrombocyte Count	Interval	Numerical	Examination	110.000

2. MATERIAL AND METHODS

Unlike other attributes that are generally well-known in the medical domain, longitude and latitude require a more specific explanation. Every location on Earth is defined by two values: latitude, which determines the horizontal axis, and longitude, which determines the vertical axis. Latitude has been recognized since ancient times by civilizations such as the Greeks and Romans. Eratosthenes (276–194 BC), a Greek scholar, used latitude to estimate the Earth's circumference. Similarly, another Greek scholar, Ptolemy (2nd century AD), produced a monumental work titled Geographia to map the Earth [5]. Unlike latitude, which can be determined based on the sun's position, longitude is more difficult to determine, as it requires time as a reference parameter. In 1884, the Meridian Conference established the Prime Meridian at zero degrees longitude, passing through the Greenwich Observatory in London—creating a standardized global coordinate system that remains in use today.

The coordinate lookup function utilizes Nominatim, a geocoding service based on OpenStreetMap (OSM), to transform address strings into geographic coordinates. This process was implemented in Step 11 of the dataset preparation workflow to enrich the data with spatial attributes. This approach allows address-based location data to be converted into precise longitude and latitude values, enabling subsequent geospatial analyses such as clustering and disease mapping.

To ensure compliance with ethical standards and protect patient confidentiality, all personally identifiable information (PII) was removed from the dataset during preprocessing. This included patient names, identification numbers, exact residential addresses, and any other attributes that could directly or indirectly reveal an individual's identity. Additionally, sensitive fields were anonymized or transformed into categorical representations where appropriate. The final dataset only retains de-identified clinical and spatial attributes, such as symptom indicators, laboratory results, and approximate location coordinates (longitude and latitude), which are limited to facility-level granularity. These measures were implemented in accordance with data protection principles to minimize re-identification risks while maintaining the analytical value of the dataset for research purposes.

There are 11 steps involved in producing this dataset, starting from importing SQL data to completing the data with longitude and latitude (Figure 2).

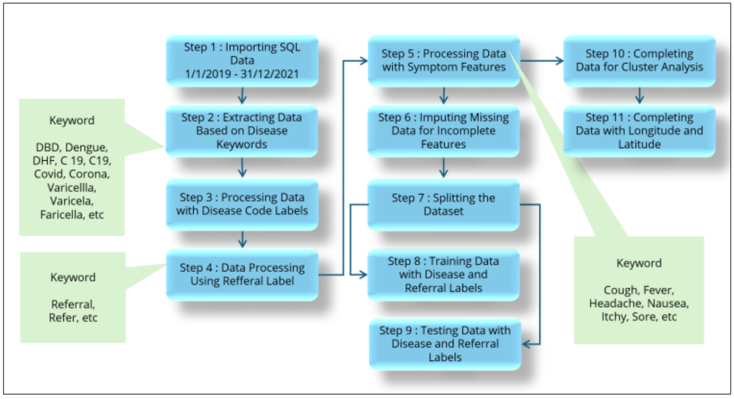
Step 1 : Importing SQL Data 1/1/2019 - 31/12/2021
Step 2 : Extracting Data Based on Disease Keywords
Step 3 : Processing Data with Disease Code Labels
Step 4 : Data Processing Using Refferal Label
Step 5 : Processing Data with Symptom Features
Step 6 : Imputing Missing Data for Incomplete Features
Step 7 : Splitting the Dataset
Step 8 : Training Data with Disease and Referral Labels
Step 9 : Testing Data with Disease and Referral Labels
Step 10 : Completing Data for Cluster Analysis
Step 11 : Completing Data with Longitude and Latitude
Keyword
DBD, Dengue, DHF, C 19, C19, Covid, Corona, Varicellla, Varicela, Faricella, etc
Keyword
Referral, Refer, etc
Keyword
Cough, Fever, Headache, Nausea, Itchy, Sore, etc

Figure 2. Dataset Preparation Process

Step 1: Importing SQL Data

The initial step in the dataset preparation process is importing data from Microsoft SQL Server, the primary storage for patient medical records. This involves establishing a secure connection to the server using ODBC or tools like SQL Server Management Studio (SSMS), Python (with pyodbc or sqlalchemy), or Power BI. The imported data includes key information such as patient ID, chief complaints, diagnoses, symptoms, and test results. During import, data quality must be ensured from the outset—this includes consistent date formats, valid diagnosis codes, and alignment between the extracted columns and analytical needs. This raw data serves as the foundation for all subsequent processes, so a secure connection and well-optimized queries are essential to avoid disrupting production systems.

Step 2: Extracting Data Based on Disease Keywords

Once the data is successfully imported, the next step is extracting entries based on relevant disease-related keywords. This helps filter records to focus only on data indicating specific diseases of interest—such as infectious, non-infectious, or chronic illnesses. The keywords may include disease names, medical terms, or ICD (International Classification of Diseases) codes. This extraction process is performed using text pattern matching on diagnosis or symptom columns, employing SQL's LIKE operator or regular expressions in Python. The result is a more targeted subset of data, allowing downstream processes to proceed more efficiently and aligned with the analysis objectives.

Step 3: Processing Data with Disease Code Labels

The filtered dataset is then labeled with disease codes, often referring to standardized classifications such as ICD-10. These labels serve as target variables for building disease classification models, ensuring each data entry has a clearly defined disease tag. Labeling can be automated using ICD reference tables or done semi-manually for unstructured entries. Maintaining consistency in labeling is critical to avoid biased analysis and to ensure the results are valid for research, prediction, or epidemiological reporting.

Step 4: Processing Data with Referral Code Labels

The next step involves labeling the data with referral codes, identifying whether a patient was treated at the current facility or referred elsewhere. The label can be binary (treated/referred) or more detailed based on the referral destination (e.g., general hospital, specialty hospital, another clinic). This label is crucial for building clinical decision-support systems and managing patient flow. By analyzing referral patterns in relation to symptoms and diagnoses, the system can better predict referral needs and help plan healthcare resources effectively.

Step 5: Processing Data with Symptom Features

Following labeling, the focus shifts to processing disease symptoms. These are typically captured as free-text entries or checkbox options selected by medical staff. This step involves extracting, normalizing, and converting symptoms into numerical formats suitable for machine learning. Techniques such as one-hot encoding or embeddings can be used depending on symptom complexity and variation. The result is a structured feature set representing a patient's clinical condition, which significantly influences model accuracy in disease and referral prediction.

Step 6: Imputing Missing Data for Incomplete Features

Medical datasets often contain missing or incomplete fields—such as unrecorded blood pressure or unfilled symptom checklists. To address this, imputation is applied, filling missing values using statistical or machine learning-based approaches. Common imputation methods include using the mean/mode for numerical data, or more advanced techniques like k-NN or regression-based imputation. The goal is to preserve data integrity and maximize the usable data pool for model training and evaluation.

Step 7: Splitting the Dataset

Once the data is cleaned and complete, it is divided into two sets: training and testing datasets. A typical ratio is 70:30 or 80:20 depending on dataset size and model complexity. The split is performed randomly, often using stratified sampling to maintain balanced label distributions across both sets. This ensures that the trained model can be objectively evaluated on previously unseen data.

Step 8: Training Data with Disease and Referral Labels

The training dataset is then used to build predictive models with two target labels: disease code and referral status. Depending on the approach, either two separate models or a single multi-label model may be developed. Algorithms like decision trees, random forests, or neural networks are commonly used. During training, the model learns to associate input features (symptoms, age, history) with output labels. The outcome is a model capable of predicting both the disease classification and referral likelihood for new patient records.

Step 9: Testing Data with Disease and Referral Labels

The testing dataset is used to evaluate the trained model's performance. Each model prediction is compared against the actual label in the test set. Evaluation is done using metrics like accuracy, precision, recall, F1-score, and AUC. This step is essential to assess the model's ability to generalize beyond the training data. The evaluation results help determine whether the model needs retuning, feature enhancement, or data re-imputation.

Step 10: Completing Data for Cluster Analysis

Beyond classification, cluster analysis is conducted to group patient data based on shared features. This is done using unsupervised learning algorithms like K-Means, DBSCAN, or Hierarchical Clustering.

Clustering helps reveal hidden patterns in the data—such as patient groups with similar symptoms or high referral rates. These insights can guide the development of targeted health interventions for specific clusters.

Step 11: Completing Data with Longitude and Latitude

Finally, the dataset is enriched with geographic information—longitude and latitude coordinates. These may come from healthcare facility addresses or administrative-level patient locations. Adding spatial data enables geospatial analyses such as disease spread mapping, referral hotspot detection, or interactive dashboard development. By combining clinical and geographic perspectives, the analysis becomes more holistic, supporting better planning and delivery of equitable healthcare services.

3. RESULT AND DISCUSSION

The preparation and processing of the dataset yielded a well-structured and anonymized collection of patient records related to three infectious diseases: COV, DHF, and VAR. After careful filtering, labeling, and feature engineering, the final dataset consisted of 1,015 COV cases, 1,356 DHF cases, and 308 VAR cases. These cases were extracted from a pool of over 68,000 visit records spanning three years (2019–2021), originating from a healthcare facility in Bandung Regency.

The first result was the successful classification of diseases using supervised learning models. Experiments using traditional classifiers such as Decision Trees, Random Forests, and XGBoost showed high accuracy in identifying disease types based on patient symptoms, examination data, and additional features. The disease code (vCode) became the primary label, and symptoms like fever, cough, shortness of breath, red spots, and nausea were among the most significant predictors. Feature importance analysis using SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations) confirmed these variables as the most influential in decision-making by the models.

In addition, a second predictive model focused on referral status (vReferral) — determining whether a patient required referral to a more advanced healthcare facility. Using similar features as inputs, the model successfully predicted referral decisions, aiding the potential development of clinical decision support systems (CDSS). This experiment is particularly valuable for healthcare systems with limited resources, enabling smarter patient routing.

Further analysis was conducted through cluster-based modeling. Using geospatial data (longitude and latitude), unsupervised algorithms like K-Means and DBSCAN were applied to identify patterns in the distribution of cases. The clustering revealed geographic hotspots of infection, supporting epidemiological surveillance and localized health interventions. This spatial understanding is crucial in managing outbreaks where disease transmission is closely linked to proximity and mobility patterns.

Another key outcome was the application of synthetic data generation to address the challenge of small sample sizes and privacy concerns in medical data. Using techniques such as CTGAN and basic generative models, the team was able to augment the dataset, preserving data structure without compromising patient confidentiality. This synthetic data supported further model training and experimentation.

Lastly, the inclusion of interpretable models and geospatial dimensions elevated the usefulness of the dataset. By combining clinical features with location data, this research not only supports predictive analytics but also opens the door to the development of public health dashboards, early warning systems, and policy planning tools. The methods and results shown here are expected to contribute to future research on disease prediction and outbreak control in developing regions. Figure 3 shows the disease quantity of the three disease types relative to the total number of cases obtained.

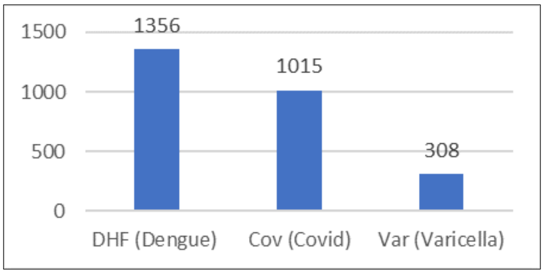

Figure 3. Disease Quantity

The results of this dataset development have been utilized in seven experiments, four of which have already been published in two journals and two conference proceedings. Table 4 describes several uses of the dataset in research, some of which have already been published in journals and conference proceedings.

ISSN: 2337-7631 (Printed)
ISSN: 2654-4091 (Online)

Table 4. Experiment

No	Experiment Name	Experiment Objective	Result
1	DHF Disease Prediction	Dataset trial for predicting Dengue Hemorrhagic Fever [6]	Achieved satisfactory results
2	Disease Type Prediction	Determining patient's disease type among three alternatives: COV, DHF, and VAR	Achieved satisfactory results
3	Referral Prediction	Predicting whether a patient should be referred to a facility equipped to handle infectious diseases	Achieved satisfactory results
4	Cluster Analysis	Using longitude and latitude to identify the number of cases in a specific geographic area, proximity, etc [7].	Achieved satisfactory results
5	Dominant Feature Prediction using LIME and SHAP	Utilizing model interpreters LIME and SHAP to identify the most dominant features in predictions	Achieved satisfactory results
6	Synthetic Data Generation [8]	To overcome limited medical data populations due to privacy regulations [9]	Achieved satisfactory results
7	Prediction with XGBoost and Feature Interpretation (SHAP)	XGBoost implementation with SHAP-based dominant feature identification [10] [11]	Achieved satisfactory results

4. CONCLUSION

This study potentially contributes to the development of decision support systems, although further validation with healthcare stakeholders is required. As part of this effort, it presents a structured approach to the development of a disease-specific dataset derived from electronic medical records, with an emphasis on three geographically transmissible infectious diseases: COV, DHF, and VAR. The methodology encompasses data extraction, transformation, anonymization, feature engineering, and geospatial enrichment—ensuring that the dataset adheres to both analytical rigor and ethical standards regarding patient confidentiality.

The experimental results underscore the dataset's utility in supporting various machine learning tasks, including disease classification, referral prediction, and cluster-based spatial analysis. The integration of model interpretability techniques (SHAP and LIME) further enhances transparency and trust in predictive outputs. Moreover, the generation of synthetic data addresses constraints related to data availability and privacy, offering a viable path for model training and validation in sensitive healthcare contexts.

Overall, the findings affirm that the structured and geotagged dataset developed in this study holds significant potential for advancing data-driven epidemiological research and clinical decision support, particularly in settings with limited health data infrastructure. Future investigations may build upon this foundation by incorporating longitudinal analyses, real-time data integration, and broader population-level health indicators.

BIBLIOGRAPHY

[1] M. G. Guzman, D. J. Gubler, A. Izquierdo, E. Martinez, and S. B. Halstead, "Dengue infection," *Nat. Rev. Dis. Prim.*, vol. 2, 2016, doi: 10.1038/nrdp.2016.55.

[2] J. P. Utami, "Epidemiologi Varicella," *www.alomedika.com*, 2023. [online]. Available at: https://www.alomedika.com/penyakit/penyakit-infeksi/cacar-air/epidemiologi.

[3] T. Willy, "Pengertian Demam Berdarah," *Dokter.Tips*, 2014. [online]. Available at: https://www.alodokter.com/demam-berdarah.

[4] alodokter.com, "COVID-19," 2021.[online]. Available at: https://www.alodokter.com/covid-19.

[5] S. Sharma, H. K. Shakya, and V. Marriboyina, "A location based novel recommender framework of user interest through data categorization," *Mater. Today Proc.*, vol. 47, no. 19, pp. 7155–7161, 2020, doi: 10.1016/j.matpr.2021.06.325.

[6] H. I. Pohan, W. Suparta, Y. Heryadi, A. Wibowo, and L. Lukas, "Prediction of DHF (Dengue Hemorrhagic Fever) Severity Using Random Forest, KNN, Decision Tree and Naïve Bayes," *Proc. 2022 IEEE 7th Int. Conf. Inf. Technol. Digit. Appl. ICITDA 2022*, 2022, doi: 10.1109/ICITDA55840.2022.9971377.

[7] H. I. Pohan, "*Using Maps as a Factor to Increase The Accuracy of Collaborative Filtering in Providing Recommendations Regarding Cluster-Based Diseases Covid-19, Varicella and Dengue v2*," Educational Administration: Theory and Practice, Bandung, 2024. doi: 10.53555/kuey.v30i4.2460.

[8] J. Moon, S. Jung, S. Park, and E. Hwang, "Conditional tabular GaN-based two-stage data generation scheme for short-term load forecasting," *IEEE Access*, vol. 8, pp. 205327–205339, 2020, doi: 10.1109/ACCESS.2020.3037063.

[9] H. I. Pohan, "The Effect of Combined Synthetic Tabular Data Generated Using CTGAN Model with

Actual Data on Performance of DHF, Varicella, and COVID-19 Recognition Model," *J. Electr. Syst.*, vol. 20, no. 3, pp. 1867–1873, 2024, doi: 10.52783/jes.3797.

[10] H. I. Pohan, R. Rahmania, and A. I. Arrahmah, "Predicting Infectious Diseases Using XGBoost Algorithm and Discovering Dominant Features Using SHAP Model Interpreter," *2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI)*, pp. 479–484, 2025, doi: 10.1109/ICoCSETI63724.2025.11019611.

[11] S. R. Vadyala, S. N. Betgeri, E. A. Sherer, and A. Amritphale, "Prediction of the number of COVID-19 confirmed cases based on K-means-LSTM," *Array*, vol. 11, p. 100085, Sep. 2021, doi: 10.1016/j.array.2021.100085.