Implementation of K-Nearest Neighbor to Predict the Chances of COVID-19 Patients' Recovery

Salma Nurzaqiah
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
salmanurzaqiah99@gmail.com

Ichsan Taufik
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
ichsan@uinsgd.ac.id
https://orcid.org/0000-0001-5052-0635

Dian Sa'adillah Maylawati
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
diansm@uinsgd.ac.id
https://orcid.org/0000-0002-1193-3370

Wildan Budiawan Zulfikar
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
wildan.b@uinsgd.ac.id
https://orcid.org/0000-0002-7387-7300

Popon Dauni
Department of Information System
Universitas Kebangsaan
Bandung, Indonesia
popon.dauni@universitaskebangsaan.ac.id
https://orcid.org/0000-0002-2857-4722

***Abstract*— Coronavirus Disease 2019 (COVID-19) is a new disease discovered in 2019 in Wuhan, China, and then spread worldwide. Many victims have confirmed varying positive levels of infection based on the patient's immunity. This study aimed to predict the chances of COVID-19 patients' recovery based on the patient's symptoms and conditions. The method used is the K-Nearest Neighbor (KNN) algorithm. KNN produces two classes of predictions: the chance of recovering or the possibility of dying. Based on the experimental results on 496 data from patients who were confirmed positive for COVID-19, KNN predicted the chances of recovery for patients with confirmed COVID-19 with an average accuracy of 88.16%. A prediction system for the chance of recovery for COVID-19 patients is constructed by choosing the best model from five test scenarios based on the given k value. The best model is at a value of k equal to 4, with an accuracy value of 88.8%.**

Keywords—COVID-19, k-nearest neighbor, machine learning, prediction

I. Introduction

Coronavirus disease 2019 (COVID-19), which was discovered in 2019, contains the Sars-CoV-2 virus and spreads between animals and humans (zoonosis) rapidly [1]. Several studies have shown that SARS was transmitted from ferrets to humans, while MERS was transmitted from camels to humans. However, it is not yet known which animal is the main source of the spread of COVID-19. According to World Health Organization (WHO) data, as of June 3, 2022, 528,816,317 people worldwide had been confirmed positive for COVID-19 [2]. As of June 1, 2022, 6,055,341 residents in Indonesia had been certified positive for COVID-19 [3].

Immunity is the main factor that affects the chances of recovery for COVID-19 patients [4]. In addition to a healthy lifestyle such as adequate rest, namely 6-8 hours of quality sleep, washing hands, maintaining distance, isolating, and wearing masks, several factors affect the body's immunity, such as: age, comorbidities (chronic kidney disease, chronic kidney disease, chronic kidney disease, obstructive pulmonary disease, diabetes), fatigue, shortness of breath, altered mental status, and critical condition. These conditions can be used as indicators for the COVID-19 patient to have a chance to recover or allow the patient to pass away.

Machine learning technology can predict the chances of recovering COVID-19 patients based on symptoms, conditions, and historical data from patients confirmed positive for COVID-19. The prediction results are undoubtedly useful for preventing unwanted things from COVID-19 patients. Several related studies than utilized machine learning to predict condition during COVID-19 pandemic, including: (1) Prediction of recovery for COVID-19 patients in Indonesia through therapy using the Naive Bayes method [5]; (2) Using the Exponential Smoothing Method, an application for predicting the recovery rate of COVID-19 in Jakarta, Indonesia was developed [6]; (3) Exponential regression prediction of COVID-19 vaccination target achievement in Indonesia [7]; (4) Using temporal deep learning to predict COVID-19 disease progression and patient outcomes [8]; (5) Convolutional Neural Network and Grad-Cam combination for COVID-19 disease prediction and visual explanation [9]; (6) Classification of COVID-19 disease symptoms using several machine learning methods, including K-Nearest Neighbor (KNN), Neural Network (NN), Random Forest (RF), and Naive Bayes [10]. In contrast to some of these studies, this study utilizes primary data for cases of positive COVID-19 patients in Indonesia.

Of the many methods and algorithms that can be used in machine learning technology, the K-Nearest Neighbor (KNN) algorithm is one of the algorithms that has good performance. KNN has been used in several prediction systems, such as prediction of heart disease [11], [12], detection of cervical cancer [13], [14], prediction of diphtheria [15], chronic kidney disease prediction [16], diabetes mellitus [17], [18], dengue fever, malaria, and typhoid [19]. Therefore, this study aims to detect the chances of recovering COVID-19 patients by using the KNN algorithm.

II. Research Methods

A. Cross-Industry Standard Process for Data Mining (CRISP-DM)

This study took a data science approach, employing the CRISP-DM (Cross-Industry Standard Process for Data Mining) data science methodology [20], [21]. Business understanding, data understanding, data preparation, modelling, model evaluation, deployment, and model management are all activities in the CRISP-DM Data Science methodology. Model management activity is not carried out in this study.

978-1-6654-5122-2/22/$31.00 ©2022 IEEE

Business understanding is identifying and investigating an organization's needs, establishing business objectives, technical data science objectives, and project planning for data science. ***Data understanding*** is the process of comprehending and reviewing the data requirements that will be used to solve the identified business challenges. This data understanding activity collects data for additional examination and validation. ***Data preparation*** is the preliminary procedure before constructing the model. Sorting, cleaning, building, integrating, and labeling data are examples of activities performed. ***Modeling*** is the primary method for creating models to solve data science problems. At this level of modeling, we design test scenarios for the model in addition to building the required model. Next, ***evaluation*** is carried out in order to choose the optimal model and ensure that it is well-executed and successfully solves business problems. Last, ***deployment*** is implementing a model in the form of an application or software that end-users can access. At this stage, the activities carried out include making a deployment model plan, carrying out the deployment model process, making a maintenance plan and performing maintenance on the model and application.

B. K-Nearest Neighbor

The K-Nearest Neighbor (KNN) algorithm, often known as the KNN algorithm, is a classification technique frequently used to carry out the classification process of an item based on learning data that is closest to the object. KNN is an algorithm based on the proximity of a data's location (distance) to other data [22]. KNN is a non-parametric method used for classification and regression [23], [24]. The input consists of the k closest training examples in the data set in both cases. The output depends on whether KNN is used for classification or regression. In KNN **classification**, the output is class membership. An object is classified by the plurality of its neighbors, with the object assigned to the most common class among its k nearest neighbors (k is a positive integer, usually small). If k = 1, then the object is assigned to that single closest neighbor class. Meanwhile, in KNN **regression**, the output is the object property value. This value is the average value of k nearest neighbors.

For the prediction system, the KNN used is a classification type where the function is only approximated locally and all computations are deferred until the evaluation of the function. Since this algorithm relies on distance for classification, if the features represent different physical units or have very different scales, normalizing the training data can dramatically improve their accuracy [25], [26]. For both classification and regression, a useful technique can be to assign weights to neighboring contributions so that closer neighbors contribute more to the average than farther away. For example, a general weighting scheme consists of assigning a weight of 1/d to each neighbor, where d is the distance to the neighbor. Neighbors are taken from a set of objects whose class (for KNN classification) or object property values (for KNN regression) are known. This can be thought of as a training set for the algorithm, although no explicit training step is required. In general, the KNN process consists of: (1) Training set includes classes; (2) Examine k items near the item to be classified; (3) New item placed in a class with the most number of close items; and (4) Iterate until each tuple in the size of the training set to be classified.

C. Confusion Matrix

The confusion matrix is not a metric, but it is useful to see the distribution of validity over an experiment. The characteristics include actual data axes and predictive data axes, each class is mapped to each other, and valid experiments are on the main diagonal. The confusion matrix measurement used in this research is accuracy, precision, and recall which are described in Table I [31]–[33].

TABLE I. CONFUSION MATRIX

Actual Value	Prediction Value		
	Positive	**Negative**	**Formula**
Positive	True Positive (TP)	False Negative (FN)	Recall, Sensitivity, True Positive rate $\frac{TP}{TP+FN}$
Negative	False Positive (FP)	True Negative (TN)	Specificity, True Negative Rate $\frac{TN}{FP+TN}$ False Positive Rate $\frac{FP}{FP+TN}$
Formula	Precision $\frac{TP}{TP+FP}$	Negative Predictive Value $\frac{TN}{TN+FN}$	Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

True Positive denotes the number of times the model successfully classified a Positive sample as Positive. False Negative is the number of times the model wrongly classified a Positive sample as Negative. False Positive refers to the number of times the model mistakenly classified a Negative sample as Positive. On the other hand, True Negative refers to how many times the model correctly classified a Negative sample as Negative.

Accuracy is a statistic that describes how the model performs in general across all classes. It is helpful when all classes are equally important. It is determined by dividing the number of correct guesses by the total number of forecasts. The precision is calculated as the ratio of Positive samples accurately identified to the total number of Positive samples classified (either correctly or incorrectly). The precision of the model assesses its accuracy in categorizing a sample as positive. The recall is computed as the ratio of Positive samples that were correctly categorized as Positive to the total number of Positive samples. The recall of the model assesses its ability to detect positive samples. The more positive samples identified, the larger the recall.

III. RESULT AND DISCUSSION

This section presents the research result of the prediction system for the chance of COVID-19 patients' recovery. The explanation begins with business understanding, data understanding, data preparation, KNN modeling, KNN evaluation, and deployment process.

A. Business understanding

The increase in confirmed positive COVID-19 patients in Indonesia is quite troubling. The opportunity for patient recovery is also one of the problems that can be detected early by utilizing machine learning technology. Therefore, this study proposes a solution to predict the chances of a patient's recovery based on the symptoms and condition of a positive patient for COVID-19. This study used several attributes,

namely age, as well as symptoms such as headache, anosmia, fever, cough, loss of appetite, hoarseness, sore throat, chest pain, weakness, confusion, muscle pain, shortness of breath, diarrhea, abdominal pain, and comorbidities. The results of this system are in the form of prediction results for Covid-19 patients who fall into the category of likely to recover or possibly die. The prediction results are useful for making it easier for health workers to detect early opportunities for healing COVID-19 patients so that appropriate actions according to priorities can be taken to avoid unexpected things.

B. Data understanding

The data collected in this study is field data originating from a community health center in the Subang district, Indonesia. As of August 13, 2021, there were 544 patient data confirmed positive for COVID-19. The data obtained from the community health center is the initial data for further data collection following the research objectives. Because the data obtained only provides patient information in the form of age, address, gender, and recovery status, the researchers conducted interviews with patients and their families listed on the list. From 544 patient data, 272 data were collected with further identification of the patient's symptoms and conditions such as [27]: headache, loss of smell, loss of appetite, cough, fever, hoarseness, sore throat, chest pain, fatigue, dizziness, muscle pain, shortness of breath, diarrhea, abdominal pain and comorbidities.

TABLE II. DATASETS EXAMPLE

Symptoms	Patient I	Patient II	Patient III
Age	27	61	49
Headache	No	No	Yes
Loss of smell	Yes	Yes	Yes
Fever	Yes	Yes	Yes
Cough	Yes	No	No
Loss of appetite	Yes	Yes	Yes
Hoarseness	No	No	No
Sore throat	No	No	No
Chest pain	No	Yes	No
Fatigue	No	Yes	No
Dizziness	No	Yes	No
Muscle pain	No	Yes	Yes
Breathless	No	Yes	No
Diarrhea	No	No	No
Abdominal pain	No	No	No
Comorbidities	Yes	Yes	Yes
Status	Recovered	Passed away	Recovered

Table II presents the example of a dataset that was used as training and testing data in the processing using the KNN method to predict recovery from COVID-19 patients based on the history of symptoms felt by confirmed patients. The 15 symptoms used as variables came from 14 symptoms based on research conducted by researchers from King's College London. One other variable, namely comorbidities, was taken based on interviews with parties from the disease control section at the Subang District Health Office. From the data collected, there is a significant imbalance. Of the 272 data, 248 are patients with recovered status, while the remaining 48 have passed away. The solution to this problem is explained in the data preparation section.

C. Data preparation

There are two main data preparation activities: uniformity of data in numeric form and Synthetic Minority Over-sampling Technique (SMOTE) technique to solve the class imbalance. All of the data of the symptoms are changed into 1 and 0. 1 for "Yes" (experiencing symptoms) and 0 for "No" (no symptoms). For class label, 1 for "chance to recover" and 0 for "possibility of passed away." Table III shows the result after all of the data changed into numeric.

TABLE III. TRANSFORMATION OF DATASETS EXAMPLE

Symptoms	Patient I	Patient II	Patient III
Age	27	61	49
Headache	0	0	1
Loss of smell	1	1	1
Fever	1	1	1
Cough	1	0	0
Loss of appetite	1	1	1
Hoarseness	0	0	0
Sore throat	0	0	0
Chest pain	0	1	0
Fatigue	0	1	0
Dizziness	0	1	0
Muscle pain	0	1	1
Breathless	0	1	0
Diarrhea	0	0	0
Abdominal pain	0	0	0
Comorbidities	1	1	1
Status	1	0	1

Then, SMOTE is a popular strategy for dealing with class imbalances [28]–[30]. This technique creates a new instance of the minority class by generating a combinative convex of neighboring instances to create a new sample of the minority class to balance the dataset. Creating synthetic samples rather than copying them makes it possible to balance the dataset without becoming overfitting. Following the application of the SMOTE sampling procedure, 497 COVID-19 patient data were acquired with balance class results.

D. K-Nearest Neighbor Modeling

The K-Nearest Neighbor modeling to predict the chances of recovering COVID-19 patients was carried out with five scenarios to obtain which model was the best. The scenario considers the distribution of training data and testing data and variations in the value of k, which are 3, 4, 5, 6, and 7. The five scenarios for the development of the model include:

- Data distribution is 90% for training data and 10% for testing data (Scenario I).
- Data distribution is 80% for training data and 20% for testing data (Scenario II).
- Data distribution is 70% for training data and 30% for testing data (Scenario III).
- Data distribution is 60% for training data and 40% for testing data (Scenario IV).
- Data distribution is 50% for training data and 50% for testing data (Scenario V).

E. Evaluation result

Using a confusion matrix, an evaluation of the model results of KNN in predicting the chance COVID-19 patients' recovery. Table IV shows the result for each modeling scenario with various k values. The precision, recall, and accuracy result are measured using a formula that is available in Table I. The average precision, recall, and accuracy value of all modeling scenarios are 88.64%, 88.84%, and 88.16%.

TABLE IV. CONFUSIN MATRIX RESULT FOR KNN MODEL

Model	K Value	Precision	Recall	Accuracy
Scenario I	3	84.5	85	84
	4	89.5	89.5	88
	5	84.5	85	84
	6	87	87	86
	7	87	85	86
Scenario II	3	91.5	91.5	91
	4	91.5	91.5	91
	5	89.5	89.5	89
	6	89.5	89.5	89
	7	89.5	89.5	89
Scenario III	3	89.5	89.5	89
	4	89.5	90	89
	5	89	89	89
	6	88.5	89	88
	7	87.5	88	87
Scenario IV	3	90.5	91	90
	4	90	90	89
	5	89	89	88
	6	87.5	87.5	87
	7	87.5	87	87
Scenario V	3	90.5	91	90
	4	89.5	89	90
	5	90	90	88
	6	88.5	88	89
	7	90	90	87

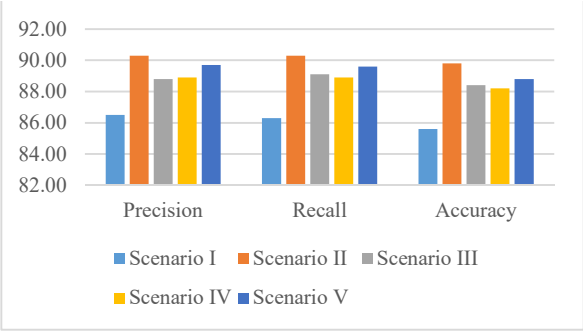

Fig. 1. The average result of confusion matrix for each modeling scenario

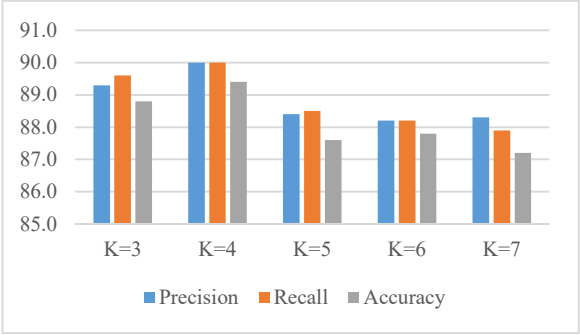

Fig. 2. The average confusion matrix result based on k value

Figure 1 shows the average value of precision, recall and accuracy of each KNN modeling scenario. From those results, modeling scenario II has the highest average precision, recall, and accuracy than other scenarios. Scenario II has average value of precision 90.3%, recall 90.30%, and 89.80% of accuracy. Meanwhile, the performance based on k-value that available in the Figure 2, k=4 has the highest score of precision, recall, and accuracy average value. The average value of precision, recall, and accuracy for k=4 is 90%, 90%, and 88.8%. Based on those result, the best modeling to predict chance of COVID-19 patients' recovery using KNN is when percentage of training and testing data is 80% and 20% with the k-value is 4. Therefore, the best model that resulted is used for the prediction system.

F. Deployment

A system to predict the chance of COVID-19 patients' recovery using the KNN algorithm is developed as a web-based application. This system has two main functions (available in Figure 3): identifying the COVID-19 symptoms and predicting the chance of recovery with the KNN model. Figure 4 describes the user interface of the prediction system available in the Indonesian language.

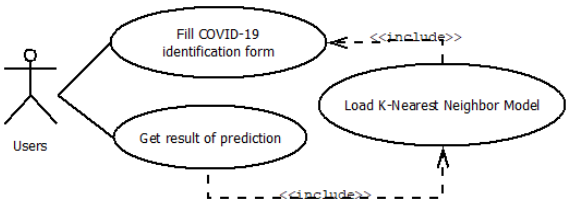

Fig. 3. Use case diagram for prediction system

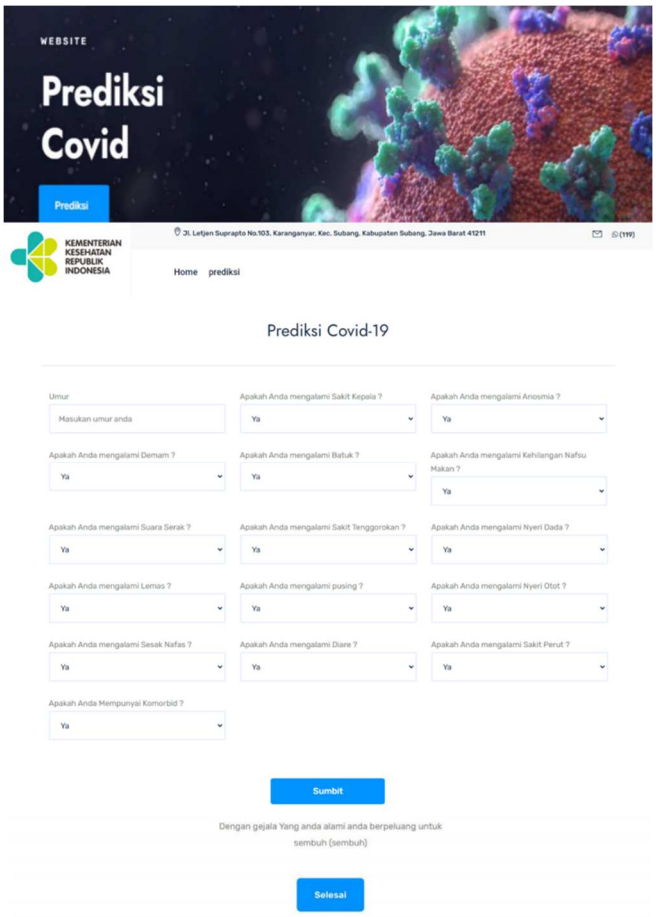

Fig. 4. User interface example of prediction system

CONCLUSION

This research uses machine learning technology to predict the chance of COVID-19 patients' recovery with the K-Nearest Neighbor (KNN) algorithm. Sixteen symptoms are identified to predict whether COVID-19 patients have an opportunity to recover or have the possibility of passed away. Experiments show that KNN can predict the chance of COVID-19 patients' recovery with a good performance. The

best model of KNN is also implemented into a web-based prediction system. Since this study lacks data, further research can collect more representative data to get more accurate prediction results. It also can use another machine learning algorithm besides KNN, such as the Deep Learning method, which is popular in machine learning applications.

ACKNOWLEDGMENT

We would like to thank the Center for Research and Publications, Institute for Research and Community Service at UIN Sunan Gunung Djati Bandung for supporting and financing this publication.

REFERENCES

[1] M. Ciotti, M. Ciccozzi, A. Terrinoni, W.-C. Jiang, C.-B. Wang, and S. Bernardini, "The COVID-19 pandemic," *Crit. Rev. Clin. Lab. Sci.*, vol. 57, no. 6, pp. 365–388, 2020.

[2] World Health Organization, "WHO Coronavirus (COVID-19) Dashboard," 2022. https://covid19.who.int/ (accessed June 6, 2022).

[3] Satuan Tugas Penanganan COVID-19, "Data Sebaran COVID-19," *covid19.go.id*, 2022. https://covid19.go.id/ (accessed Nov. 08, 2021).

[4] S. Akbar, "Imunitas Faktor Utama Kesembuhan Pasien COVID-19," *probolinggokab.go.id*, 2020. https://probolinggokab.go.id/imunitas-faktor-utama-kesembuhan-pasien-covid-19/ (accessed Jun. 06, 2022).

[5] O. P. Barus and A. Tehja, "Prediksi Kesembuhan Pasien COVID-19 di Indonesia Melalui Terapi Menggunakan Metode Naive Bayes," *J. Inf. Syst. Dev.*, vol. 6, no. 2, pp. 59–66, 2021.

[6] A. Setiawan, "Aplikasi Prediksi Tingkat Kesembuhan Covid di DKI Jakarta Dengan Metode Exponensial Smoothing," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 4, pp. 2187–2197, 2021.

[7] T. E. E. Tju, D. S. Maylawati, G. Munawar, and S. Utomo, "Prediction of the COVID-19 Vaccination Target Achievement with Exponential Regression," *JISA (Jurnal Inform. dan Sains)*, vol. 4, no. 2, pp. 179–182, 2021.

[8] C. Sun, S. Hong, M. Song, H. Li, and Z. Wang, "Predicting COVID-19 disease progression and patient outcomes based on temporal deep learning," *BMC Med. Inform. Decis. Mak.*, vol. 21, no. 1, pp. 1–16, 2021.

[9] H. Moujahid *et al.*, "Combining CNN and Grad-Cam for COVID-19 Disease Prediction and Visual Explanation," *Intell. Autom. Soft Comput.*, vol. 32, no. 2, pp. 723–745, 2021.

[10] S. Anggraini, M. Akbar, A. Wijaya, H. Syaputra, and M. Sobri, "Klasifikasi Gejala Penyakit Coronavirus Disease 19 (COVID-19) Menggunakan Machine Learning," *J. Softw. Eng. Ampera*, vol. 2, no. 1, pp. 57–68, 2021.

[11] L. Andiani, S. Sukemi, and D. P. Rini, "Analisis Penyakit Jantung Menggunakan Metode KNN Dan Random Forest," in *Annual Research Seminar (ARS)*, 2020, vol. 5, no. 1, pp. 165–169.

[12] M. E. I. Lestari, "Penerapan algoritma klasifikasi Nearest Neighbor (K-NN) untuk mendeteksi penyakit jantung," *Fakt. Exacta*, vol. 7, no. 4, pp. 366–371, 2015.

[13] T. Praningki and I. Budi, "Sistem Prediksi Penyakit Kanker Serviks Menggunakan CART, Naive Bayes, dan k-NN," *Creat. Inf. Technol. J.*, vol. 4, no. 2, pp. 83–93, 2018.

[14] E. S. Salim, H. Bindan, E. Pranoto, and A. Dharma, "Analisa Metode Random Forest Tree dan K-Nearest Neighbor dalam Mendeteksi Kanker Serviks," *J. Ilmu Komput. Dan Sist. Inf.*, vol. 3, no. 2, pp. 97–101, 2020.

[15] C. S. Fatoni and F. D. Noviandha, "Case Based Reasoning Diagnosis Penyakit Difteri dengan Algoritma K-Nearest Neighbor," *Creat. Inf. Technol. J.*, vol. 4, no. 3, pp. 220–232, 2018.

[16] W. Yunus, "Algoritma K-Nearest Neighbor Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Ginjal Kronik," *J. Cosphi*, vol. 2, no. 2, 2018.

[17] I. Indrayanti, D. Sugianti, and A. Al Karomi, "Optimasi Parameter K Pada Algoritma K-Nearest Neighbour Untuk Klasifikasi Penyakit Diabetes Mellitus," *Pros. SNATIF*, pp. 823–829, 2017.

[18] W. Wijanarto and R. Puspitasari, "Optimasi Algoritma Klasifikasi Biner dengan Tuning Parameter pada Penyakit Diabetes Mellitus," *J. Eksplora Inform.*, vol. 9, no. 1, pp. 50–59, 2019.

[19] E. N. Shofia, R. R. M. Putri, and A. Arwan, "Sistem Pakar Diagnosis Penyakit Demam: DBD, Malaria dan Tifoid Menggunakan Metode K-Nearest Neighbor - Certainty Factor," *J. Pengemb. Teknol. Inf. dan Ilmu Komputer, e-ISSN*, 2017.

[20] P. Chapman *et al.*, "Cross Industry Standard Process for Data Mining 1.0," *Step-by-step Data Min. Guid.*, 2000.

[21] P. Chapman *et al.*, "The CRISP-DM user guide," in *4th CRISP-DM SIG Workshop in Brussels in March*, 1999, vol. 1999.

[22] E. Prasetyo, "Data mining mengolah data menjadi informasi menggunakan matlab," *Yogyakarta Andi Offset*, 2014.

[23] E. Fix and J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *Int. Stat. Rev. / Rev. Int. Stat.*, vol. 57, no. 3, p. 238, 1989, doi: 10.2307/1403797.

[24] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Am. Stat.*, vol. 46, no. 3, pp. 175–185, 1992, doi: 10.1080/00031305.1992.10475879.

[25] S. M. Piryonesi and T. E. El-Diraby, "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems," *J. Transp. Eng. Part B Pavements*, vol. 146, no. 2, p. 04020022, 2020, doi: 10.1061/jpeodx.0000175.

[26] T. Hastie, R. Tibshirani, J. H. Friedman, and MyiLibrary., "The elements of statistical learning data mining, inference, and prediction : with 200 full-color illustrations," *Springer Ser. Stat.*, p. xvi, 533 p., 2001, [Online]. Available: http://www.myilibrary.com?id=18743.

[27] C. Sudre *et al.*, "Attributes and predictors of Long-COVID: analysis of COVID cases and their symptoms collected by the Covid Symptoms Study App," 2020.

[28] A. Fernandez, S. Garcia, F. Herrera, and N. V Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018.

[29] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm," in *International Conference on Neural Information Processing*, 2010, pp. 152–159.

[30] P. aw Skryjomski and B. Krawczyk, "Influence of minority class instance types on SMOTE imbalanced data oversampling," in *first international workshop on learning with imbalanced domains: theory and applications*, 2017, pp. 7–21.

[31] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artif. Intell. Rev.*, 2017, doi: 10.1007/s10462-016-9475-9.

[32] R. Rautray and R. C. Balabantaray, "An evolutionary framework for multi document summarization using Cuckoo search approach: MDSCSA," *Appl. Comput. informatics*, vol. 14, no. 2, pp. 134–144, 2018.

[33] P. Verma, S. Pal, and H. Om, "A Comparative Analysis on Hindi and English Extractive Text Summarization," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 18, no. 3, pp. 1–39, 2019.