ELINVO (Electronics, Informatics, and Vocational Education), *November 2023; vol 8(2):248-254*
ISSN 2580-6424 (printed), ISSN 2477-2399 (online,) DOI: https://doi.org/10.21831/elinvo.v8i2.55702

Visitor Decision System in Selection of Tourist Sites Based on Hybrid of Chi-Square And K-NN Methods

Devie Rosa Anamisa[1*], Fifin Ayu Mufarroha[2], Achmad Jauhari[3]
[1,2,3] Departemen of Informatics, Faculty of Engineering, Trunojoyo Madura University, Bangkalan, Indonesia
[*]devros_gress@trunojoyo.ac.id

ABSTRACT

Madura Island is one of the islands with a lot of tourism spread over four districts, such as natural, religious, and cultural tourism. And every year, various visitors visit various tourist sites in Madura, so an increase in the number of visitors has been found in multiple places. This is influenced in addition to the type of tourist attraction but also changes in tourist behavior in making decisions to visit tourist objects. Most of the researchers have applied the right decision-making with intelligence-based measurement. However, the accuracy obtained has not yet reached the optimal solution. Therefore, this study uses the Chi-Square and K-Nearest Neighbors (K-NN) methods to recommend tourist attraction locations based on visitor characteristics to increase visitor attractiveness in tourist attractions scattered in Bangkalan, Madura. Chi-Square is used to select features that affect tourist attraction visitor factors by testing the relationship between the variables involved. Meanwhile, K-NN is a method of classifying potential visitor attractions based on their characteristics by using the closest membership calculation, which is the largest from the test data. The calculation is carried out by the square of the Euclidian distance from each object, then sorted from the smallest to the largest value and looking for the value of k as the result of the decision. There are ten features used in the classification, such as tourism type, management services, facilities, gender, age, occupation, education, visitor status, ticket prices, and sales trends. There are three classes classified: low, medium, and high visitor attractiveness. The contribution of this study is to analyze the effect of the characteristics of tourist attraction visitors on increasing visitor attractiveness using the chi-square and K-NN methods. Based on the results of system testing using K-Fold Cross Validation with five folds from 315 datasets, it produces the highest accuracy at k-fold = 3 worth 84.12% with eight selected features.

Keywords*:* Decision System; Selection of Tourist Sites; Method; Chi-Square; K-Nearest Neighbors;

INTRODUCTION

Madura Island is one of the islands with a lot of tourism. Tourism is one sector that can improve the economy of the Indonesian people. Various essential components can improve the economy, one of which is visitors[1]. There are two types of tourists, namely foreign tourists and domestic tourists[2]. Tourists are one of the essential factors that can improve the regional economy with the number of visits to various tourist attractions[3]. In addition, tourists visiting tourist attractions have several purposes, such as the purpose of recreation, business, or other purposes. There are four regencies, namely Bangkalan, Sampang, Pamekasan, and Sumenep. Bangkalan Regency has around sixteen tourist attractions scattered, including: Sembilang Beach, Syaikhona Kholil Tomb, Tretan Swimming Pool, Jaddih Hill, et.al.

The number of tourists who come to various tourist attractions in Bangkalan, Madura, increases daily. This is in line with the behavioral model of visitors who come to the tourist attraction. Tourist behavior models[4] have been influenced by various visiting decision factors, including the desire to travel[5][6], because this requires a tourist profile and tourism awareness so that it triggers the search for information about the tourist attraction to get alternative trips before making a decision. From these problems, this study analyzes the influence of visitor characteristics to make the right decisions in determining the destination tourist attraction.

For this decision making several methods have been done in previous studies. But have not yet made the right decision like the research conducted by [7] for the classification and selection of features based on the bee colony algorithm where the effectiveness of the proposed method produces a classification accuracy of 57.17% of 8 components with six selected attributes, then developed by [8] regarding the search, categorization, and planning of tourism with an e-tourism approach. The method proposed in this paper is about searching for tours for users, and depending on the category of user tours, and then manually scheduled once. And the result of this method is the ability to plan trip features according to the user's requirements. However, in 2019 it was developed [9], The feature selection method, namely chi-square[10], can group fingerprint databases by reducing the negative impact of various indoor environments and ignoring the relative distance between RSS vectors in multiple locations. In addition, the research[11] regarding the classification of Twitter content using Chi-Square and KNN[12][13] using a dataset obtained from UCI in 2014 received an accuracy of 65% with the nearest neighbor distance of k=3. The classification method is a method used for grouping data into predetermined classes[14]. However, minimal effort has been made in the classification of tourism, which focuses specifically on the characteristics of visitors to attractions And several methods have been developed in previous research, including: Decision Tree (C4.5), Naive Bayes (NB), Neural Network (NN), and Support Vector Machines (SVM). However, it has not been able to increase the value of accuracy, as has been developed in [15] regarding the method of classification of instance-based learning patterns (IBL), which requires the deal of 'closeness' or 'similarity between instances and high classification effectiveness. Furthermore, in research[16] regarding C4.5 for predicting creditworthiness in banks. The C4.5 algorithm itself has good accuracy. However, C4.5 still has weaknesses in handling high-dimensional data. In a study conducted by the automatic classification of Chinese music emotions using the Naive Bayes method and produces an accuracy value of about 68%. Meanwhile, in the research that has been developed by[17] classify internet traffic based on machine learning-based transaction protocols, namely Naive Bayes and K-Nearest Neighbors. And value of accuracy achieved by K-NN is higher than that of NB. This study applies the KNN method for the classification stage from several previous studies. The KNN algorithm is also lazy learning, which means it does not use training data points to create models. All training data is used at the testing stage. This makes the training process faster and the testing phase slower. However, to speed up the classification process, a feature selection stage is needed[18]. Feature selection is the preprocessing of data. This is intended for data improvement so that learning algorithms become more effective in computing data and can also eliminate irrelevant and redundant data attributes. The process of learning algorithms on data is increasing[19]. In addition to weaknesses, there are also advantages of KNN, such as: being able to determine where new data should be placed with the character of KNN being "bird of a feather".

Therefore, to solve the problem in this study, it has proposed a model to increase accuracy by classifying the characteristics of tourist attraction visitors using the KNN method based on feature selection of chi-square to tourist location recommendations so that it can increase the attractiveness of visitors based on visitor characteristics.

METHODS

In this research, a system design consists of several well-structured and systematic stages to achieve the expected goals, which can be seen in Fig 1. Based on this design, the feature selection stage uses the chi-square method. The Chi-Square method selects ten features of the characteristics of the chosen visitors, such as: Type of tourism, tourism management services, facilities, gender, age, education, type of visitor work, status, ticket prices, and the presence or absence of ticket discounts. This stage serves to reduce unnecessary features to speed up the work of the classification process. While at the classification stage with KNN, at this stage, the classification of visitor characteristics is carried out based on training data based on the closest distance. Chi-Square uses statistical theory to test the independence of terms with their categories. Based on statistical theory, there are two events including the occurrence of features and occurrences of categories, which then each term value is sorted from the highest by forming a contingency table and determining the value of degrees of freedom (df) from the table. The degrees of freedom are used to compare the Chi Square results calculated with the Chi Square table. if the chi-square result is smaller than the chi-square table, the feature is selected. While at the classification stage with KNN, at this stage the classification of visitor characteristics is carried out by dividing training data and testing data. The classification process is done by calculating the closest distance. After generating alternative tourist attraction recommendations, to measure the accuracy of this hybrid method, a testing phase is carried out to find out how well the system is in classifying and measuring the accuracy of the KNN method and feature selection based on the k-fold Cross-Validation calculation.

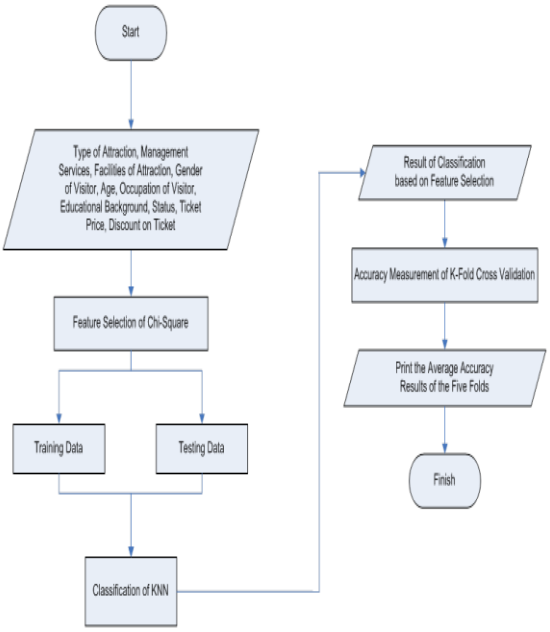

Start
Type of Attraction, Management Services, Facilities of Attraction, Gender of Visitor, Age, Occupation of Visitor, Educational Background, Status, Ticket Price, Discount on Ticket
Feature Selection of Chi-Square
Training Data
Testing Data
Classification of KNN
Result of Classification based on Feature Selection
Accuracy Measurement of K-Fold Cross Validation
Print the Average Accuracy Results of the Five Folds
Finish

Figure 1. System Diagram of Visitor KNN Method based on Feature Selection with Chi-Square

A. *Selection Features*

Feature selection is an optimization process to reduce a significant feature set from the original source, to obtain a relatively small and significant number of feature subsets to increase accuracy in the classification process[20]. So that feature selection can make the tools in the classification process better, more efficient, and effective by selecting the number of training data features, and determining suitable features to consider in the learning process. In this study, the Chi-Square method is used to select the characteristic features of tourist attraction visitors. The Chi-Square method is a feature selection used to calculate the level of dependency of a feature on a class. The Chi-Square method assigns a value to the features, which are then sorted and selected according to the percentage tested. The set features are used for the classification process. Based on statistical theory, which is based on two events, namely the occurrence of the element and the occurrence of its category, which is based on the calculation of equation (1) [21].

$$x^2(D,t,c) = \sum_{et\epsilon(1,0)} \quad \sum_{ec\epsilon(1,0)} \quad \frac{N_{et\,el}-E_{et\,el}}{E_{et\,el}}$$

(1)

B. *K-Nearest Neighbors Method*

The K-Nearest Neighbor (KNN) method is one method that is often used in the classification process based on machine learning[22]. The purpose of the KNN method is to classify objects into one of the classes that already exist in the sample data that has been previously defined. Based on the closest distance or similarity to the existing dataset or training data, the KNN algorithm assumes that something similar will exist in close proximity or neighbors. This means that data that tend to be similar will be close to each other. The new data is then assigned to the class where most of the neighboring data resides. The stages in the KNN process, including[23]:

- Determining the value of *k*
- Calculate the distance between the data to be classified against the label data
- Determine the smallest value of k
- Classify data by distance metric in calculating the proximity of the distance between the data can be done using the distance metric by calculating the Euclidean distance in equation (2).

$$(x_1, x_2) = \sqrt{\sum_{i=1}^{n} \quad (x_{1i} - x_{2i})^2} \qquad (2)$$

RESULT AND DISCUSSION

A. *Data Input*

Data mining is data that contains the search for patterns to assist decision-making in the future by using statistical techniques[24]. So, data mining is a series of processes. At the same time, the dataset is a collection of objects and attribute[25]. The dataset is divided into two parts[26], including: training data is data that previously existed according to facts and mathematics, and testing data is data used as test material. This data is used to measure the classification correctly. In this study, data has been obtained from the Bangkalan Tourism Office in 2021, with ten indicators with 21 tourist objects, which can be seen in Table 1 for criteria data, and Table 2 for 21 Bangkalan Tourist Attraction data, where the Bangkalan Wista object is located in Kwanyar District, Konang, Galis, Socah, Bangkalan, Arosbaya, Geger, Kokop, Tanjung Bumi, and Sepulu. Visualization of tourist attraction visitor data in Bangkalan in 2021, can be seen in Fig. 2.

Table 1. Criteria data from characteristics of tourist attractions

No	Criteria	Sub-Criteria	No	Criteria	Sub Criteria
1.	Type of Tour	Nature Artificial Religion	6.	Job	Farmers/ Fishermen Employee Government employees Other
2.	Manager Service	Low Medium High	7.	Graduate	Primary school Secondary school High School
3.	Facility	0≤ x ≤5 6≤ x ≤10 x ≥ 11	8.	Status	Not married yet Married
4.	Gender	Male Female	9.	Ticket Price	Rp 1.000-5.000 Rp 6.000-10.000 Rp 11.000-15.000 Rp 16.000-20.000 Rp 21.000-25.000
5.	Age	x ≤ 15 16 ≤ x ≤ 24 25 ≤ x ≤ 34 35 ≤ x ≤ 49 x ≥ 50	10	Ticket discount	No Yes

Table 2. Tourist attraction of data in bangkalan, madura

No	Name's of Tourism	No	Name's of Tourism	No	Name's of Tourism
1	Sembilang beach	8	Paseban Park	15	Bangkalan Grand Mosque
2	Shaykhona Kholil's Tomb	9	Friendly Lantern Hill	16	Gebang Beach
3	Tretan Swimming Pool	10	Bangkalan Square	17.	Aer Mata Ebu's Tomb
4	Jaddih Hill	11	Siring Kemuning Beach	18	Arosbaya Limestone Hill
5	Cakraningrat Museum	12	Bangkalan City Recreation Park	19	Tomb of Sultan R. Abdul Kadirun
6	Rongkang Beach	13	Goa Pote Swimming Pool	20	Tlagoh Beach
7	Labuhan Mangrove Education Park	14	Geger Hill	21	Tresna Art

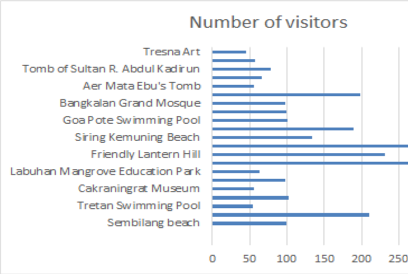
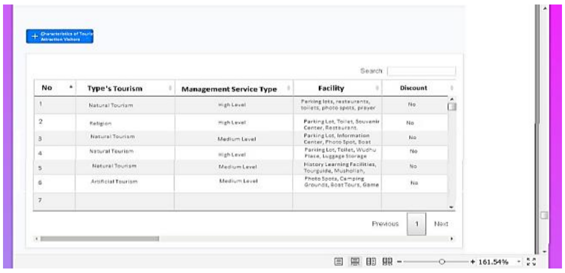


Figure 2. Visitor Data For Each Tourist Attraction in Bangkalan

Based on the stages to complete the classification of the characteristics of visitors to Bangkalan tourism objects using KNN based on feature selection with Chi-Square, these steps measure the level of accuracy by entering a dataset of 315 data.

B. Testing with Chi-Square and KNN Methods

Based on the classification process diagram design carried out in this study using the KNN and chi-square methods, the first step is to process 315 data with ten features, as shown in Fig. 3 for feature selection using the Chi-Square method, and the results can be seen in Fig. 4. The graph shows that if X2 is greater than x2table shows that the selected features, such as type of tour, management service, facility, gender, age, job, education, and status. Then the process is carried out using the KNN method for the classification process of tourist objects based on the characteristics of tourist attraction visitors to obtain alternative recommendations for suitable tourist objects from 21 tourist objects in Bangkalan. By calculating the Euclidean Distance based on equation 2, the resulting distance from each location will be searched for the shortest distance as an alternative solution for the right location for visitors according to their characteristics. The implementation of the system can be seen in Fig. 5.



No	Type's Tourism	Management Service Type	Facility	Discount
1	Natural Tourism	High Level	Parking lots, restaurants, toilets, photo spots, prayer	No
2	Religion	High Level	Parking Lot, Toilet, Souvenir Center, Restaurant.	No
3	Natural Tourism	Medium Level	Parking Lot, Information Center, Photo Spot, Boat	No
4	Natural Tourism	High Level	Parking Lot, Toilet, Wudhu Place, Luggage Storage	No
5	Natural Tourism	Medium Level	History Learning Facilities, Tourguide, Musholla h,	No
6	Artificial Tourism	Medium Level	Photo Spots, Camping Grounds, Boat Tours, Game	No
7				

Figure 3. Input data training of tourist attraction visitors characteristics

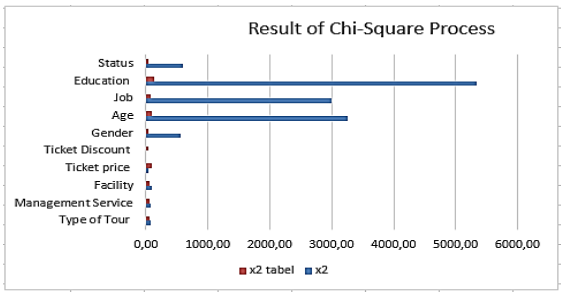
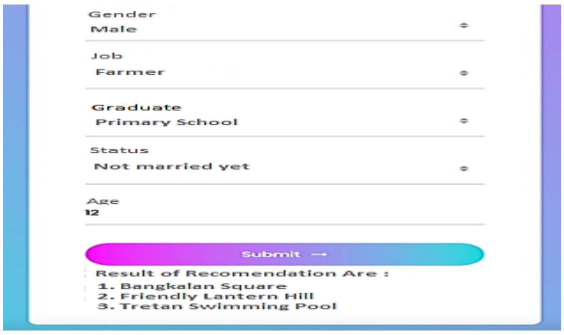


Figure 4. Result of chi-square process with ten fitur


Gender
Male

Job
Farmer

Graduate
Primary School

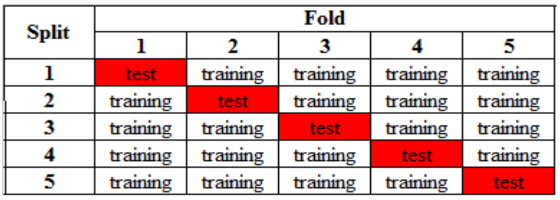
Status
Not married yet

Age
12

Result of Recomendation Are :
1. Bangkalan Square
2. Friendly Lantern Hill
3. Tretan Swimming Pool

Figure 5. Implementation of KNN and Chi-Square to classification of caracteristic of tourist attractions

The calculation of the measurement of classification performance in this study uses k-fold cross validation[28] from the model's predictions by breaking the data into k parts of data sets of the same size. Training and testing are carried out with k values from 1 to 5, as shown in Fig 6. And the measurement of this classification performance is carried out by comparing all test data that are classified correctly from many test data. The results of system accuracy can be seen in Table 3.



Split	Fold				
	1	2	3	4	5
1	test	training	training	training	training
2	training	test	training	training	training
3	training	training	test	training	training
4	training	training	training	test	training
5	training	training	training	training	test

Figure 6. Model 5-fold cross validation

Table 3. Accuracy results from Comparison of Recommendation Process with Hybrid KNN and chi-square methods and KNN without chi-square

k	Accuration (%)	
	KNN and chi-Square	KNN Without Chi-Square
1	77,78	77,61
2	80,95	75,89
3	84,12	74,95
4	79,37	65,78

5	77,78	56,88

CONCLUSION

From the results of research with calculations, the KNN method classifies the characteristics of tourist attraction visitors based on Chi-Square as a feature selection method, the average accuracy is 80% and produces the highest accuracy at fold = 3 worth 84.12% against 315 datasets with 21 attractions and eight indicators selected from the characteristics of visitors. Therefore, it can be concluded that based on the level of accuracy, the modeling of the KNN method and feature selection with Chi-Square is better in the classification process so that visitors can determine tourist objects according to their characteristics as alternative decision-making and can assist the government in developing tourism objects in Bangkalan, Madura based on visitor attraction.

REFERENCES

[1] L. Gorji, “Tourism Development in the Urban Side ’ s Promenades (Study Case : Barzok City) Abstract : Tourism phenomenon is rooted in Motion and displacement and it gives mobility to human life in impartibly way . Today tourism activity is one of the most dynamic ,” pp. 1–5, 2016.

[2] C. Torng, “The Effect of Experiential Agriculture Activities on the Tourism Image of Foreign Tourists to Dai-Dai Recreational Agriculture Area in Taiwan,” 2013, doi: 10.1109/CISIS.2013.125.

[3] S. Li, S. Takahashi, K. Yamada, M. Takagi, and J. Sasaki, “Analysis of SNS Photo Data Taken by Foreign Tourists to Japan and a Proposed Adaptive Tourism Recommendation System,” 2017.

[4] C. H. C. Hsu and S. S. Huang, “Formation of Tourist Behavioral Intention and Actual Behavior,” 2010.

[5] S. Sefidpour, J. Wang, and K. Srivastava, “Factors Affecting Traveling Wave Protection”.

[6] N. Penghu, “Exploring the Relationship of Travel Constraints , Destination Image , and Revisit Intention,” pp. 799–804, 2019, doi: 10.1109/IIAI-AAI.2019.00163.

[7] H. Hongnian, Y. Qian, Z. Shuang, and W. F. Zhu, “Selection Based On Improved Artificial Bee Colony Algorithm,” pp. 242–247, 2016.

[8] V. B. Joshi and R. H. Goudar, “Searching, categorizing and tour planning: A novel approach towards e-tourism,” *RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc.*, vol. 2018-Janua, pp. 1002–1005, 2017, doi: 10.1109/RTEICT.2017.8256749.

[9] S. Hu, C. Shen, K. Zhang, and X. Huang, “Improved Wknn Indoor Positioning Algorithm Based On C-Means And Chi-Square Distance,” *2019 Int. Conf. Robot. Intell. Syst.*, no. 2, pp. 432–435, 2019, doi: 10.1109/ICRIS.2019.00113.

[10] M. Göl, “A Modified Chi-Squares Test for Improved Bad Data Detection,” no. 1, pp. 1–5.

[11] Y. D. Setiyaningrum, “Classification of Twitter Contents using Chi-Square and K-Nearest Neighbour Algorithm,” *2019 Int. Semin. Appl. Technol. Inf. Commun.*, pp. 1–4, 2019, doi: 10.1109/ISEMANTIC.2019.8884290.

[12] S. Demirci, “KNN ile ˙ Istenmeyen E-posta Filtreleme : k gerinin Sınıflandırma Performansına Etkisinin Ara ¸ stırılması Spam Filtering with KNN : Investigation of the Effect of k Value on Classification Performance,” pp. 14–17, 2020.

[13] “Data Mining Algorithms (KNN & DT) Based Predictive Analysis on Selected Candidates in Academic Performance,” pp. 332–337, 2021.

[14] S. xia Chen, X. kang Wang, H. yu Zhang, J. qiang Wang, and J. juan Peng, “Customer purchase forecasting for online tourism: A data-driven method with multiplex behavior data,” *Tour. Manag.*, vol. 87, no. May, p. 104357, 2021, doi: 10.1016/j.tourman.2021.104357.

[15] C. C. Huang and H. Y. Chang, “A novel SVM-based reduced NN classification method,” *Proc. - 2015 11th Int. Conf. Comput. Intell. Secur. CIS 2015*, pp. 62–65, 2016, doi: 10.1109/CIS.2015.23.

[16] N. Iriadi and N. Nuraeni, “Kajian

Penerapan Metode Klasifikasi Data Kelayakan Kredit Pada Bank," *J. Tek. Komput. AMIK BSI*, vol. II, no. 1, pp. 132–137, 2016.

[17] M. Dixit, R. Sharma, S. Shaikh, and K. Muley, "Internet traffic detection using naïve bayes and K-Nearest neighbors (KNN) algorithm," *2019 Int. Conf. Intell. Comput. Control Syst. ICCS 2019*, no. Iciccs, pp. 1153–1157, 2019, doi: 10.1109/ICCS45141.2019.9065655.

[18] H. Elmunsyah, I. A. E. Zaeni, F. A. Dwiyanto, and T. Widiyaningtyas, "Classification of Employee Mental Health Disorder Treatment With K-Nearest Neighbor Algorithm," pp. 4–8, 2019.

[19] E. S. Gualberto, R. T. D. E. Sousa, S. Member, and C. G. Duque, "The Answer Is in the Text : Multi-Stage Methods for Phishing Detection Based on Feature Engineering," pp. 223529–223547, 2020, doi: 10.1109/ACCESS.2020.3043396.

[20] K. Sriporn and C. F. Tsai, "Predicting Tourists' Behavior of Virtual Museum Using Support Vector Machine with Feature Selection Technique," *Proc. - Int. Conf. Mach. Learn. Cybern.*, vol. 2, pp. 433–438, 2018, doi: 10.1109/ICMLC.2018.8526959.

[21] B. Cheng, R. J. Stanley, S. Antani, and G. R. Thoma, "Graphical figure classification using data fusion for integrating text and image features," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 693–697, 2013, doi: 10.1109/ICDAR.2013.142.

[22] Y. Li and B. Cheng, "An improved k-nearest neighbor algorithm and its application to high resolution remote sensing image classification," *2009 17th Int. Conf. Geoinformatics, Geoinformatics 2009*, pp. 1–4, 2009, doi: 10.1109/GEOINFORMATICS.2009.5293389.

[23] G. A. Sandag, N. E. Tedry, and S. Lolong, "Classification of Lower Back Pain Using K-Nearest Neighbor Algorithm," *2018 6th Int. Conf. Cyber IT Serv. Manag. CITSM 2018*, no. Citsm, pp. 1–5, 2019, doi: 10.1109/CITSM.2018.8674361.

[24] S. J. Saleh, S. Q. Ali, and A. M. Zeki, "Random Forest vs. SVM vs. KNN in classifying Smartphone and Smartwatch sensor data using CRISP-DM," *2020 Int. Conf. Data Anal. Bus. Ind. W. Towar. a Sustain. Econ. ICDABI 2020*, pp. 28–31, 2020, doi: 10.1109/ICDABI51230.2020.9325607.

[25] Y. Peng and H. Biao, "KNN based outlier detection algorithm in large dataset," *2008 Int. Work. Educ. Technol. Train. 2008 Int. Work. Geosci. Remote Sensing, ETT GRS 2008*, vol. 1, pp. 611–613, 2008, doi: 10.1109/ETTandGRS.2008.306.

[26] R. Patra and B. Khuntia, "Predictive Analysis of Rapid Spread of Heart Disease with Data Mining," *Proc. 2019 3rd IEEE Int. Conf. Electr. Comput. Commun. Technol. ICECCT 2019*, pp. 1–4, 2019, doi: 10.1109/ICECCT.2019.8869194.