Recursive Journal of Informatics
Vol. 4, No. 1, March 2026
p-ISSN 2963-5551 https://journal.unnes.ac.id/journals/index.php/rji e-ISSN 2986-6588

Optimization K-Nearest Neighbor Algorithm Using Information Gain and Hyperparameter Tuning in Adult Male Fertility Classification

Muhammad Zaenal Muttaqin[1], Anggyi Trisnawan Putra[2]
[1]Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract. Male fertility plays an important role in reproductive capability and global population dynamics. Male infertility can be caused by lifestyle, health conditions, and sperm quality. This research develops a male fertility classification model with an optimized K-Nearest Neighbor (KNN) algorithm using Information Gain feature selection and hyperparameter tuning with GridSearchCV. The main problems encountered are low accuracy in prediction and high computational complexity due to many irrelevant features. To overcome this, feature selection and hyperparameter optimization methods were used. The dataset used in this research comes from the UCI Machine Learning Repository, consisting of 100 data with 10 attributes. The KNN algorithm was chosen for its simplicity and ability to classify data with multiple classes and uneven distribution. However, its accuracy is highly dependent on the proper selection of features and parameters. The Information Gain method is used for selection of significant features against the target variable, reducing model complexity and computation time. Hyperparameter tuning is performed using GridSearchCV to find the best combination of parameters. The results showed that the application of Information Gain and GridSearchCV successfully improved the classification accuracy of KNN. The final model achieved 94% accuracy, better than the previous conventional method which only reached 84%. This increase in accuracy shows that KNN optimization with this approach is effective in improving male fertility classification performance. This research is expected to contribute to the development of male fertility diagnostic technology and the implementation of more accurate prediction models in clinical practice.
Purpose: The proposed model is a development based on previous research that focuses on developing the K-Nearest Neighbor algorithm with a model accuracy of 84%. this study uses feature selection techniques and hyperparameter tuning in the K-Nearest Neighbor (KNN) algorithm to improve the accuracy of the male fertility classification model.
Methods/Study design/approach: To improve the curation of the male fertility classification model and to optimize the model from previous research, this study uses the feature selection technique and hyperparameter tuning technique. For this technique, 2 types of optimization are carried out, namely feature selection using Information Gain and GridSearchCV hyperparameter tuning to get the best parameter combination for the proposed model. The fertility dataset has also been used in previous studies, used in this study.
Result/Findings: The proposed model obtained a high accuracy of 94%, which surpassed the model in the previous study which had an accuracy of 85% for the classification of fertility levels in men.
Novelty/Originality/Value: The novelty in this research is the addition of hyperparameter tuning techniques to optimize and obtain optimal parameters in the fertility classification model. This research also aims to improve and increase the accuracy of the previous model.

Keywords: K-Nearest Neighbor, Information Gain, GridSearchCV, machine learning, feature optimization
Received October 2024 / **Revised** March 2026 / **Accepted** March 2026

INTRODUCTION

Human fertility one of important role in the lives of individuals and societies, both in social and demographic contexts. Fertility, which refers to the ability of an individual or couple to have children, is highly significant in modern society. Rapid population growth in recent decades has given rise to a variety of challenges related to fertility in social, economic and health terms [1]. Understanding fertility dynamics is important to address these issues. Factors such as age, health conditions and lifestyle have a major influence on fertility [2].

Infertility, which is the inability of an individual or couple to conceive after 12 months of sexual intercourse without contraception, is estimated to affect a small number of couples of childbearing age

[1]*Corresponding author.
Email addresses: zaenalmuhammad211@students.unnes.ac.id (Muttaqin)
DOI: 10.15294/rji.v4i1.14868

worldwide. In men, infertility accounts for about 40-50% of all infertility cases. Research shows that common causes of male infertility are problems with hormones and the quality of sperm produced [3]. Male fertility screening includes a complete medical history and clinical evaluation as per the World Health Organization (WHO) guidelines. These steps include sperm analysis, hormone examination, microbiology, ultrasonography, and testicular biopsy to assess the reproductive organs [4].

As male infertility cases increase, various studies have been conducted to find more effective solutions. One approach that is starting to be widely used is the application of machine learning to classify fertility levels. Machine learning has the ability to identify patterns in complex data, which is often difficult for humans to recognize. The K-Nearest Neighbor (KNN) algorithm is one method that is popular due to its simplicity and ability to handle varied data. KNN is very suitable for data classification, especially when the number of classes is quite large [5]. Other research shows that KNN can produce high accuracy, such as in breast cancer detection with 97% accuracy [6].

However, KNN also has shortcomings, especially in terms of accuracy when the data contains irrelevant features or parameters that have not been optimized. Therefore, this study aims to improve fertility classification accuracy by using feature selection and hyperparameter optimization methods. Feature selection is performed using the Information Gain method to identify the most informative features related to the target variable. This can reduce model complexity and computation time by focusing only on relevant features [7]. On the other hand, hyperparameter optimization is performed using GridSearchCV, which automatically adjusts parameters to achieve the best performance of the KNN algorithm [8].

Some studies have also shown that a combination of KNN and other optimization methods can improve results. For example, in a study that used a genetic algorithm to optimize the K parameter in KNN, the classification accuracy improved significantly on several medical datasets, with the highest result reaching 94.15% [8]. In the context of infertility, another study showed that the KNN algorithm combined with Random Forest (RF) was able to produce better classification than traditional methods such as Borderline-SMOTE and k-means-SMOTE, with accuracy reaching 87% [9].

In addition, other research on male fertility prediction using Classification and Regression Trees (CART) algorithm has also been conducted. This study shows that CART can be used to classify fertility with 81% accuracy, and the results can be applied to Android-based mobile applications for ease of use [7]. Further studies show that the combination of KNN with optimization techniques, such as XGBoost, can also improve prediction accuracy on various datasets, especially for cancer diseases [10].

Overall, this study concluded that applying optimization methods such as Information Gain and GridSearchCV to the KNN algorithm can improve the accuracy of male fertility prediction. In addition, this approach can also reduce model complexity, speed up the computational process, and provide more reliable results to be applied in a clinical context. This research is expected to make a significant contribution in fertility diagnostics and offer more accurate prediction tools to support future medical decisions.

METHODS

This research involves calculation and manipulation to develop a program to classify male fertility using K-Nearest Neighbor (KNN) with optimization techniques involving information gain for feature selection and GridSearchCV for hyperparameter tuning. This research follows certain stages, start from data collection, preprocessing using label encoder and SMOTE for overcoming unbalance data, feature selection, model training, hyperparameter optimization, and model evaluation. Each stage is designed to improve the accuracy and efficiency of the KNN model. To find out more the stage of this research will be displayed of a flowchart as in Figure 1.

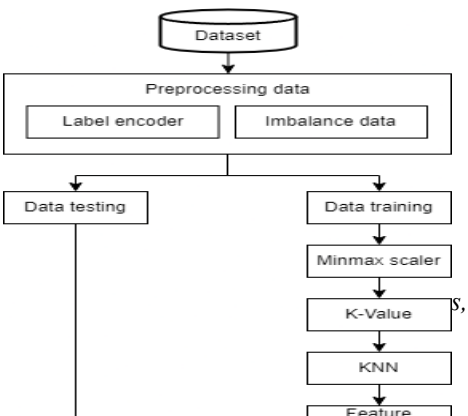

Figure 1. Research flowchart

Dataset

The data used in this research is collection from collected from volunteers provide a semen sample. Some of the features from dataset are in Table 1.

Table 1. Fertility dataset features

Features	Description
Season	Numeric value representing the season when the data was collected, with a range of values from -1 to 1.
Age	Age of the individual, ranging between 0 and 1.
childish_disease	Binary value indicating whether the individual had childhood diseases, 1 for "yes", 0 for "no".
Trauma	Binary value indicating whether the individual has experienced any serious trauma, 1 for "yes", 0 for "no".
surgical_intervention	Binary value representing whether the individual had undergone any surgical intervention, 1 for "yes", 0 for "no".
Fevers	Binary value indicating whether the individual had a high fever in the past year.
Alcoholic	Normalized value representing the frequency of alcohol consumption, ranging from 0 to 1.
Smoking	Categorical value representing the individual's smoking habit, 1 for regular smoker, 0 for non-smoker, -1 for occasional smoker.
Sitting	Normalized value indicating the number of hours the individual spends sitting per day.
Output	The target variable indicating fertility status, with "N" representing "Normal" and "O" representing "Altered" fertility.

The dataset used in this research was sourced from the UCI Machine Learning Repository, specifically the "fertility" dataset, which consists of 100 samples and 10 attributes. These attributes include factors related to lifestyle and health that are relevant to male fertility. The use of this publicly available dataset ensures reproducibility and comparability with other studies.

Preprocessing

Preprocessing are very important in this research and includes several processes, such as label encoder, and data balancing using SMOTE. This data balancing and label encoder aims to overcome bias. By using oversampling techniques on a dataset containing 100 data consisting of 88 normal class data and 12 altered data. This resulted in an indication of imbalance in the class, especially in the altered value which became the minority value. After using SMOTE, the resulting data from 100 data becomes 176 data, and by using a label encoder to convert the values into numeric formats 0 and 1. For example, the value 'N' is converted to 0 and the value 'O' is converted to 1. The use of this label encoder aims to ensure that the calculations in the KNN algorithm can run smoothly without being disturbed by non-numeric values. The result of this label encoding process has a value, where for the 'N' and 'O' values are successfully converted into 0 and 1. By making this change, the model becomes simpler and more efficient.

K-value

Finding the optimal K value is a crucial step to ensure the KNN model do able to produce accurate predictions. This process usually involves trying different values of K and evaluating the model's performance for each value using appropriate evaluation metrics, such as accuracy, precision, recall, or

F1-score. Thus, choosing the right K helps achieve balance, so that the model can generalize well to new data [11]. If the value of K too small, the model can become too sensitive to noise in the data, which can lead to overfitting. Conversely, if the K value is too large, the model may become too generalized and fail to capture important patterns in the data, which can lead to underfitting.

K-Nearest Neighbor

K-Nearest Neighbor or (KNN) identify as one of the methods used to classify objects based on the nearest training sample in the problem space [12]. The training data is projected into a dimensional space where each dimension describes a property of the data. KNN uses a control-based algorithm. The goal of the KNN algorithm is to classify new objects based on attributes and training data. The new test sample results are classified in the most KNN categories. In the classification process, this algorithm does not use a model that must be adapted and only relies on memory, in the classification process the K value is required to ensure that there are no draws in the nearest neighbor search. The KNN algorithm uses the nearest neighbor classification as the predicted value of the new test sample [13]. The Euclidean distance is used to determine the distance between 2 points, the formula that can be used to find the euclidean distance can use the Formula 1.

$$Euclidean\ distance_{(a,b)} = \sqrt{\sum_{g=1}^{p} \left(x_{ag} - x_{bg}\right)^2} \quad (1)$$

Feature selection with Information Gain

Information gain are most commonly used feature selection technique. It is a filter-based feature selection technique [14]. Information gain uses simple feature classification and noise removal, then identifies the features that have the most important information about a particular class. The best feature is determined by calculating the entropy of the feature [15], to find entropy can be calculated using Formula 2.

$$Entropy\ (S) = \sum_{i=1}^{c} - P_i P_i) \quad (2)$$

After getting the entropy value, the information gain value can be calculated using Formula 3 [16].

$$Gain\ (S, A) = Entropy\ (s) - \sum_{i=1}^{n} \frac{|S_i|}{S} Entropy\left(S_i\right) \quad (3)$$

Hyperparameter tuning GridSearchCV

Hyperparameter one of the process of finding and selecting the optimal combination of hyperparameter values for a machine learning model or algorithm [17]. Hyperparameter values should be determined before starting training and set manually or through auto-tuning. If a model has to consider multiple parameters, it can be difficult to select good parameters. Therefore, hyperparameter tuning techniques are used to optimize model performance by finding the best combination of parameters for the data [18]. To get the number of combinations in GridSearchCV can use Formula 4.

$$Number\ of\ combinations = \prod_{i=1}^{n} Number\ of\ values\ in\ -\ i \quad (4)$$

Confusion matrix

Confusion matrix are method of calculating accuracy in the performance of a classification method. This matrix contains information that compares the classification results of the method with the actual classification. This confusion matrix contains four main terms, namely: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) [19]. This confusion matrix provides a clear picture to display results related to predictions made in text or data mining. Therefore, using this technique can help to provide information about the model used whether it is accurate or not. For more details about the main terms in the confusion matrix can be seen from Table 2 below.

Table 2. Confusion Matrix Table

Actual	Prediction	
	True	False
True	True Positive (TP)	True Negative (TN)
False	False Positive (FP)	False Negative (FN)

In the process of calculating model performance evaluation, it can be done by using a classification report that produces results such as accuracy, precision, recall, and f1-score. The classification report results can be calculated using the Formula 5 until 8.

RESULT AND DISCUSSION

This research uses the K-Nearest Neighbor algorithm which is used to classify male fertility data, where K-Nearest Neighbor was chosen because it has several advantages where KNN is able to handle multi-class data, and is adaptable to take into account the data. Although it is simple and uses an instant-based approach, it is very sensitive to noise in the dataset. After the process, Information Gain is used to perform feature selection to evaluate the most informative features in relation to the class label. Feature selection using Information Gain can also increase the accuracy of the KNN algorithm, reduce the likelihood of overfitting, and overcome feature immaturity and improve model performance by removing less informative features. After Information Gain is added to increase the accuracy and address the features and improve the model performance, then to increase the accuracy in the model, GridSearch CV is used because the most informative features with Information Gain, GridSearchCV can be used to find the optimal KNN parameter combination, and reduce the risk of overfitting that may occur when using non-optimal parameters.

Label encoder

The data contained in the information provided has different information than desired. The existence of different data so that the KNN algorithm cannot run properly. In the output column there are values with N and O values as shown in Figure 2.

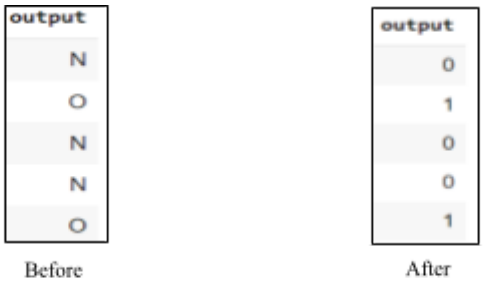

Figure 2. Before and after encoder

In Figure 2, it can be seen that the values in the data are not as desired, so it is necessary to use a label encoder to convert these values into numeric formats 0 and 1. For example, the value of 'N' is converted to 0 and the value of 'O' is converted to 1. The use of this label encoder aims to ensure that the calculations in the KNN algorithm can run smoothly without being disturbed by non-numeric values. The result of this label encoding process has a value, where for the 'N' and 'O' values are successfully changed to 0 and 1. By making this change, the model becomes simpler and more efficient, making it easier to implement and produce more optimal performance.

SMOTE

Prior to processing, the dataset used had the disadvantage of a significant amount of difference between the classes in the dataset. This imbalance must be corrected in order for the classification model to function optimally and produce more precise predictions [20]. Some approaches that can be used to address this issue include oversampling underrepresented classes, undersampling more dominant classes, or a combination of both [21]. In addition, techniques such as SMOTE or adjusting the weights in the classification model can also be applied to overcome the imbalance, the different classes in the dataset can be seen in the Figure 3.

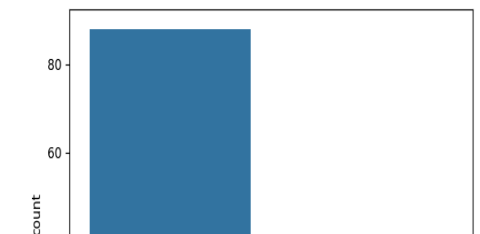

Figure 3. Class size comparison

With the existence of class imbalance by using oversampling where the dataset containing 100 data consists of 88 normal class data and 12 altered data. This resulted in an indication of imbalance in the class, especially in the altered value which became the minority value. After using SMOTE, the resulting data from 100 data becomes 176 data or 50% for normal class and 50% for altered class as shown in Figure 4.

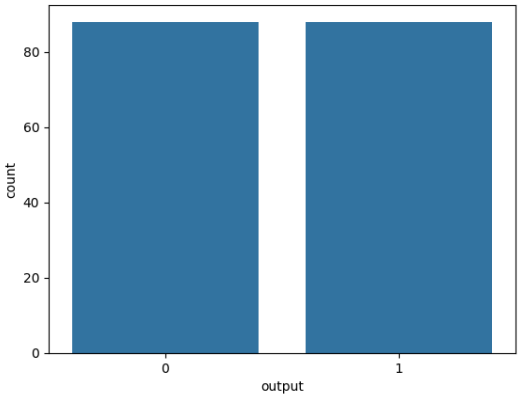

Figure 4. SMOTE result

Splitting data

In data processing, data splitting is the process of dividing data into two parts, namely training data or training data, and testing data or test data. A ratio of 80:20 is the most accurate. In this research, the method of data division, as shown in Table 3.

Table 3. Data splitting result

Category	Total data
Data Training	141
Data Testing	35

Based on Table 3, from a total of 176 research data, 141 data are used for training data, while the rest of the training data, namely 35 data, will be included in the testing data.

Feature scaling

Implementation using MinMax Scaler feature scaling is used to ensure that each feature has a uniform scale so that no single feature dominates the calculation of the distance between KNN algorithm data. In this feature scaling stage, all feature values will be changed in such a way that the same range is between 0 and 1. This is done so that the distance between data can be calculated more consistently, thus increasing the efficiency of the model, especially for algorithms such as KNN that are sensitive to feature scaling. The results of feature scaling can be seen in Table 4.

Table 4. Datasets that have been minmax scaled

No	*Season*	*Age*	*Childish_disease*	…..	*Smoking*	*Sitting*
1	0.000000	0.220000	1.0	…..	0.0	0.340426
2	0.129639	0.425136	0.0	…..	0.5	0.286944
3	0.000000	0.220000	1.0	…..	0.5	0.468085
4	1.000.000	0.520126	0.0	…..	0.5	0.363563
5	0.000000	0.160000	1.0	…..	1.0	0.468085
…..	…..	…..	…..	…..	…..	…..
96	0.000000	0.560000	1.0	…..	0.0	0.202128
97	0.335000	0.720000	1.0	…..	0.0	0.202128
98	0.972165	0.362603	0.0	…..	0.5	0.459625
99	0.335000	0.380000	1.0	…..	0.0	0.734043
100	1.000.000	0.378815	1.0	…..	0.0	0.206224

K-Value search result

In the K-value search that has been carried out using KNN, the most optimal k-value is produced at the K=2 value with the minimum error rate of 0.0556. In the K-value search process, various k-values from 1 until 20 are also performed, and the error rate is calculated based on the average wrong prediction on the test data [22]. The following Table 5 shows the value of each experiment on the value of K.

Table 5. K-Value test result

K	Error Rate
1	0.1111
2	0.0556
3	0.1111
4	0.0833
5	0.0833
6	0.0833
7	0.1111
8	0.1111
9	0.1111
10	0.1111
11	0.1111
12	0.1111
K	Error Rate
13	0.1111
14	0.1111
15	0.1111
16	0.1111
17	0.1111
18	0.1111
19	0.1111

From Table 5, it can be seen that K = 2 produces the lowest error rate, so it was chosen as the optimal K-value.

K-Nearest Neighbor algorithm classification

The data is carried out at the data preprocessing stage. In addition, the parameters used at this stage use the parameters that have been provided. K-Nearest Neighbor parameters can be seen in Table 6.

Table 6. Default parameter k-nearest neighbor

Parameters	Default value	Description
n_neighbors	5	integer
weights	'uniform'	string
algorithm	'auto'	string
leaf_size	30	integer
p	2	integer
metric	'minkowski'	string
n_jobs	1	integer

The results of evaluating the performance of the model with the confusion matrix obtained from the K-Nearest Neighbor classification using the parameters in Table 4, can be seen in Table 7.

Table 7. confusion matrix k-nearest neighbor

Actual	Predicted	
	Positive	Negative
Positive	19	4
Negative	0	13

Classification of fertility levels in adult men using the K-Nearest Neighbor algorithm gets an accuracy value of 89%. With these results it can be proven that K-Nearest Neighbor has good performance in

classifying fertility levels in adult men but at this level the value of K-Nearest Neighbor can be increased again by adding selection features that can increase the performance of a more efficient model.

Feature selection using Information Gain

The use of the best parameters selected using feature selection Information Gain. Information Gain itself is a method used to get the most informative features or those that contribute most to predicting fertility rates in adult men, and the resulting value of the most informative features can be seen in Table 8.

Table 8. Most informative feature results from Information Gain

Features	Information_Gain Value
Age	0.317777
Sitting	0.288187
Alcoholic	0.247332
Season	0.119793
Trauma	0.107581
Fevers	0.052042
Surgical_intervention	0.050532
Childish_disease	0.021808
Smoking	0.006661

Information Gain measures how much information a feature provides to the prediction of the class or target variable in the dataset. The test results of the K-Nearest Neighbor method using the best parameters of the best Informtaion Gain feature selection can be seen in Table 9.

Table 9. Confusion matrix with the best parameter information gain

Actual	Predicted	
	Positive	Negative
Positive	22	1
Negative	2	11

Classification using the KNN algorithm by finding the best parameter using Information Gain produces an accuracy value of 92%. With increasing results, it can be proven that the K-Nearest Neighbor algorithm with the best parameter choice using information gain provides excellent results for classification.

Hyperparameter tuning GridSearchCV

After determining the best features using Information Gain, the next step is to perform testing using the most optimized parameters. To achieve this, GridSearchCV is applied as a hyperparameter tuning method in the K-Nearest Neighbor algorithm. For the process of finding the best combination for Hyperparameter tuning GridSearchCV can automatically find the best and optimal combination. The process to find the best combination can be seen in Table 10.

Table 10. The process of finding the best parameters

No	n_neighbors	Weights	P	Accuracy
1	3	'uniform'	1	0.85
2	3	'uniform'	2	0.80
3	3	'distance'	1	0.86
4	3	'distance'	2	0.86
5	5	'uniform'	1	0.81
6	5	'uniform'	2	0.79
7	5	'distance'	1	0.86
8	5	'distance'	2	0.84
9	7	'uniform'	1	0.83
10	7	'uniform'	2	0.81
11	7	'distance'	1	0.86
12	7	'distance'	2	0.83

After finding the best parameters using Hyperparameter tuning GridSearchCV which has gone through 12 experiments using cross validation 5 produces the best parameters for the K-Nearest Neighbor model. The best parameter values generated by Hyperparameter tuning GridSearchCV can be seen in Table 11.

Table 11. GridSearchCV result

Parameter	Value
n_neighbors	7
weights	'distance'
p	1

After the best parameter results are generated by Hyperparameter tuning GridSearchCV, then the results of the best model will be tested using the confusion matrix, the results of testing using the best parameters can be seen in Table 12.

Table 12. Confusion matrix GridSearchCV

Actual	Predicted	
	Positive	Negative
Positive	23	0
Negative	2	11

With the results of these calculations, the KNN algorithm with the best parameter Information Gain and the best parameter from GridSearchCV produces an accuracy value of 94%. With these results, it can be proven that the optimization performed by GridSearchCV hyperparameter tuning is very effective in improving the performance of the KNN model. This result shows that this method is able to find the optimal combination of parameters, which is significant in improving the accuracy of the model.

CONCLUSION

This study successfully enhanced the K-Nearest Neighbor (KNN) algorithm for male fertility classification by utilizing Information Gain for feature selection and GridSearchCV for hyperparameter tuning, resulting in a significant accuracy increase to 94% from the previous 84%. The findings confirm that these optimization techniques effectively address issues related to computational complexity and irrelevant features, leading to a more accurate and efficient model. The optimized KNN model demonstrates strong potential for clinical applications, providing a reliable tool for predicting male fertility. Future research could expand on this work by applying the model to larger, more diverse datasets or by comparing its performance with other advanced machine learning algorithms, laying a solid foundation for further advancements in fertility diagnostics and related healthcare applications.

REFERENCES

[1] N. E. Skakkebæk *et al.*, "Environmental factors in declining human fertility," 2022. doi: 10.1038/s41574-021-00598-8.

[2] E. Spolaore and R. Wacziarg, "Fertility and Modernity," *Econ. J.*, vol. 132, no. 642, 2022, doi: 10.1093/ej/ueab066.

[3] M. Ramalingam, S. Kini, and T. Mahmood, "Male fertility and infertility," *Obstet. Gynaecol. Reprod. Med.*, vol. 24, no. 11, pp. 326–332, Nov. 2014, doi: 10.1016/J.OGRM.2014.08.006.

[4] N. Panth, A. Gavarkovs, M. Tamez, and J. Mattei, "The Influence of Diet on Fertility and the Implications for Public Health Nutrition in the United States," 2018. doi: 10.3389/fpubh.2018.00211.

[5] S. Zulaikhah Hariyanti Rukmana, A. Aziz, and W. Harianto, "OPTIMASI ALGORITMA K-NEAREST NEIGHBOR (KNN) DENGAN NORMALISASI DAN SELEKSI FITUR UNTUK KLASIFIKASI PENYAKIT LIVER," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, 2022, doi: 10.36040/jati.v6i2.4722.

[6] S. W. Binabar and Ivandari, "Optimasi Parameter K pada Algoritma KNN untuk Deteksi Penyakit Kanker Payudara," *IC-Tech*, vol. XII, no. 2, pp. 11–18, 2017.

[7] Arif Rahman Hakim, Dewi Marini Umi Atmaja, Amat Basri, and Andri Ariyanto, "Performance Analysis of Classification and Regression Tree (CART) Algorithm in Classifying Male Fertility Levels with Mobile-Based," *Tech-E*, vol. 7, no. 1, pp. 10–20, 2023, doi: 10.31253/te.v7i1.2110.

[8] R. T. Prasetio, "SELEKSI FITUR DAN OPTIMASI PARAMETER k-NN BERBASIS ALGORITMA GENETIKA PADA DATASET MEDIS," *J. Responsif Ris. Sains dan Inform.*, vol. 2, no. 2, pp. 213–221, 2020, doi: 10.51977/jti.v2i2.319.

[9] B. Karlik, A. Mohammed, Y. Bahir, and B. Koçer, "Comprising Feature Selection and Classifier Methods with SMOTE for Prediction of Male Infertility Machine Learning Applications in Medicine View project Machine Learning and Deep Learning Applications in Plasma and Thermodynamics View project Comprising F," *Artic. Int. J. Fuzzy Syst.*, no. February, 2016, [Online]. Available: https://www.researchgate.net/publication/337307643

[10] Muflih Ihza Rifatama, Mohammad Reza Faisal, Rudy Herteno, Irwan Budiman, and Muhammad Itqan Mazdadi, "Optimasi Algoritma K-Nearest Neighbor Dengan Seleksi Fitur Menggunakan Xgboost," *J. Inform. dan Rekayasa Elektron.*, vol. 6, no. 1, pp. 64–72, 2023, doi: 10.36595/jire.v6i1.723.

[11] K. Joshi, S. Jain, S. Kumar, and N. Roy, "Optimization of K-Nearest Neighbors for Classification BT - Futuristic Trends in Network and Communication Technologies," P. K. Singh, G. Veselov, V. Vyatkin, A. Pljonkin, J. M. Dodero, and Y. Kumar, Eds., Singapore: Springer Singapore, 2021, pp. 205–214.

[12] Advernesia, "Pengertian dan Cara Kerja Algoritma," https://www.advernesia.com/.

[13] A. Contreras-Valdes, J. P. Amezquita-Sanchez, D. Granados-Lieberman, and M. Valtierra-Rodriguez, "Predictive data mining techniques for fault diagnosis of electric equipment: A review," *Appl. Sci.*, vol. 10, no. 3, pp. 1–24, 2020, doi: 10.3390/app10030950.

[14] Z. Karimi, M. Mansour Riahi Kashani, and A. Harounabadi, “Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods,” *Int. J. Comput. Appl.*, vol. 78, no. 4, pp. 21–27, 2013, doi: 10.5120/13478-1164.
[15] P. Bereziński, B. Jasiul, and M. Szpyrka, “An Entropy-Based Network Anomaly Detection Method,” *Entropy*, vol. 17, no. 4, pp. 2367–2408, Apr. 2015, doi: 10.3390/e17042367.
[16] R. Haqmanullah Pambudi, B. Darma Setiawan, and Indriati, “Penerapan Algoritma C4.5 Untuk Memprediksi Nilai Kelulusan Siswa Sekolah Menengah Berdasarkan Faktor Eksternal,” *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 7, pp. 2637–2643, 2018, [Online]. Available: http://j-ptiik.ub.ac.id
[17] S. C. Gupta and N. Goel, “Predictive Modeling and Analytics for Diabetes using Hyperparameter tuned Machine Learning Techniques,” in *Procedia Computer Science*, 2022. doi: 10.1016/j.procs.2023.01.104.
[18] T. Agrawal, *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*. 2020. doi: 10.1007/978-1-4842-6579-6.
[19] Karsito and S. Susanti, “Klasifikasi Kelayakan Peserta Pengajuan Kredit Rumah Dengan Algoritma Naïve Bayes Di Perumahan Azzura Residencia,” *J. Teknol. Pelita Bangsa*, vol. 9, pp. 43–48, 2019.
[20] D. Dablain, B. Krawczyk, and N. V. Chawla, “DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 34, no. 9, 2023, doi: 10.1109/TNNLS.2021.3136503.
[21] N. Nnamoko and I. Korkontzelos, “Efficient treatment of outliers and class imbalance for diabetes prediction,” *Artif. Intell. Med.*, vol. 104, 2020, doi: 10.1016/j.artmed.2020.101815.
[22] S. Lonang, A. Yudhana, and M. K. Biddinika, “Performance Analysis for Classification of Malnourished Toddlers Using K-Nearest Neighbor,” *Sci. J. Informatics*, vol. 10, no. 3, pp. 313–322, 2023, doi: 10.15294/sji.v10i3.45196.