JEECS (Journal of Electrical Engineering and Computer Sciences)
Vol. 10, No. 1, June 2025, pp. 69-75
e-ISSN: 2579-5392 p-ISSN: 2528-0260

Comparative Study of Obesity Levels Classification

Syahrazad Syaukat Al Malaky, Alisya Akbar Choirun Nisa, Siti Armiyanti, Rizky Syahputra Setyawan

Department of Informatics, Universitas Bhayangkara Surabaya, Jl. Ahmad Yani 114, Surabaya, East Java, Indonesia

Article Info

Article history:
Received: 6 June 2024
Revised: 11 June 2024
Accepted: 30 June 2024

Keyword:
Obesity
Classification
Random Forest Classifier
Decision Tree Classifier
Public Health

Abstract

Obesity is a growing global health problem, requiring accurate data analysis to understand and address contributing factors. The level of obesity can be identified based on eating habits and physical conditions, which consist of several parameters. However, the performance of widely used machine learning methods has not provided satisfactory results. Therefore, this study analyzes obesity data using pre-processing methods to improve data quality before classifying data. The dataset used is 2111 data and includes 17 variables/features. The classification methods are Random Forest Classifier, Light Gradient Boosting Machine (LGBM) Classifier, Decision Tree Classifier, and Extra Tree Classifier. The process of data pre-processing involves data integration, data labeling, data transformation, normalization, and data cleansing. After pre-processing the data, four algorithms were used to identify patterns in the obesity data. The Random Forest Classifier is used for its ability to handle unbalanced data and reduce the risk of overfitting. The LGBM Classifier is used for a probabilistic approach to classification. The Decision Tree Classifier is applied for straightforward interpretation and clear understanding of patterns, while the Extra Tree Classifier is applied to improve the variety and accuracy of classification. The experimental results showed that a good data pre-processing method significantly improved the performance of the classification. Among the four algorithms tested, the Random Forest Classifier and Extra Tree Classifier performed best in accuracy and generalizability. Combining appropriate data pre-processing with powerful classification algorithms can provide deep insights to address obesity problems and formulate effective public health interventions.

Corresponding author:
Syahrazad Syaukat Al Malaky, arzadalmalaki3@gmail.com

DOI: https://doi.org/10.54732/jeecs.v10i1.8

1. Introduction

Increasing extra fat, or obesity, can lead to health issues. The Body Mass Index (BMI) metric determines this condition. Puspitasari's explanation states that factors such as gender, calorie intake, marital status, hereditary history, physical activity, smoking status, and level of knowledge and education all impact obesity. People of all ages, including adults, teenagers [1], and even young children, can become obese. Insufficient Body Weight, Normal Body Weight, Level I Obesity, Level II Obesity, Type I Obesity, Type II Obesity, and Type III Obesity are the seven degrees of obesity, according to Fabio. Obesity's side effects can lead to several illnesses, such as Metabolic Syndrome, which in turn can lead to Diabetes and Cardiovascular disease. In children, obesity can also cause physical motor development abnormalities [2]. Researchers were drawn to develop algorithms that can predict obesity levels after realizing the extent of the detrimental effects associated with obesity [3]. As a result, while physical exercise has been recognized as a crucial component in combating obesity, the precise connection between physical activity and obesity is still unknown. Recent technological developments have made machine learning techniques an effective tool

for identifying intricate risk variables related to obesity. In contrast to traditional statistical methods, Machine Learning (ML) learns from data without following strict guidelines and does not just focus on the association between variables, a problem that traditional regression models frequently have [4].

Several studies have been carried out to support the handling and preventing obesity, such as detecting obesity in suburban regions by combining the multiclass AdaBoost algorithm with the extra tree classifier (ETC) [5]. This research achieved up to 88% accuracy using different sample sizes and stratified random sampling. Obesity detection with the AdaBoost and Light Gradient Boosting Machine (LGBM) Classifier method found that the classification accuracy was up to 91% [6]. Other research proves that LGBM achieves 95% accuracy [7]. These results indicate that the classification supports obesity prevention. Other research uses Fuzzy C-Means (FCM) to partition employee selection participant data [8]. Thus, obesity detection can be completed by classification. Other methods include Decision Tree [9], Random Forest (RF), Extra Tree (ET), Gradient Boosting (GB), Support Vector Machine (SVM), and Artificial Neural Network (ANN) [10]. Other problems that can be solved are the classification of potential customers [11], the classification of body weight types [12], and the classification of the relationship between physical activity and obesity [13].

In this study, researchers create a classification system to predict obesity with predetermined variables. The dataset used in this study is an obesity level dataset taken from Kaggle which will then be pre-processed and trained on machine learning. The expected result of this study is the percentage of predictions that have been made by machine learning. We compare four classification methods: Random Forest, LGBM, Decision Tree, and Extra Tree Classifier. The results show that LGBM is superior to the others.

2. Research Methodology

2.1 Research Flowchart

The procedure in this study began by taking a dataset from Kaggle, where the data can be downloaded at https://www.kaggle.com/datasets/fatemehmehrparvar/obesity-levels. It is a dataset estimating obesity levels in individuals from Mexico, Peru and Columbia. The tools used in this study are colab.research.google.com. The dataset is then imported and continued with the Pre-Processing process of data. the pre-processed data then processed and trained using machine learning algorithms with classification types and testing accuracy with several methods, namely Random Forest Classifier, Light Gradient Boosting Machine (LGBM) Classifier, Decision Tree Classifier, and Extra Tree Classifier.

2.1.1 Pre-Processing Data

Data preprocessing is an initial data mining technique to convert raw data collected from various sources into clean information that can be used at a later stage [14]. This data mining techniques include data cleaning, data transformation, normalization, data integration, and data reduction.

2.1.2 Random Forest Classifier

Random Forest is a powerful decision tree ensemble method that can be used for a variety of pattern classification tasks by using a set of classification trees as a basic learning. This method relies on bootstrap aggregating and random sampling of subspace to build a committee. Therefore the final class label of each data instance is determined through a majority vote. Suppose {X, T} declares a set of training data where:

$X = x0,1, \ldots, xn-1$ and $T = t0,t1, \ldots, tn-1$.

Suppose h(x) presents a classification tree. For each individual tree h(x), the model selects a random sample by replacing the training data that has been collected and using that sample data to train h(x). This procedure aims to achieve better model performance because it has the ability to reduce the variance of the model without increasing the bias of the model. The feature bagging mechanism is also used by Random Forest in addition to sample bagging. That is, a subset of features is used to train h(x). This is a process to reduce the correlation of overall learning across committees. Usually for the case of pattern classification, the number of features selected by the individual h(x) is $\sqrt{D}$ where D is the total number of available features [15]. The use of the random forest method is suitable for analyzing the size of large datasets [16].

2.1.3 Light Gradient Boosting Machine (LGBM) *Classifier*

Light Gradient Boosting Machine is a gradient-boosting framework that has gained recognition for its ability to process large-scale datasets rapidly and effectively. The method employed in this system is a unique tree-based learning approach that emphasizes the growth of leaves in a hierarchical manner, resulting in reduced computational requirements and enhanced training speed. Light GBM expedites model construction and allows for real-time analysis in cardiovascular research, hence aiding prompt decision-making and patient care [17].

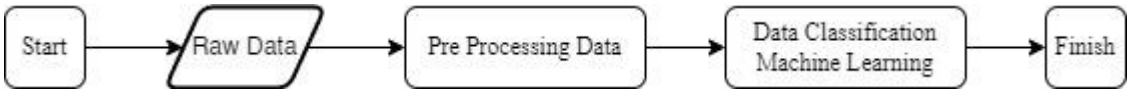

Figure 1. Research Flowchart

2.1.4 Decision Tree Classifier

Decision Tree is the most frequently used research method for classification problems. A decision tree is a structure that can be used to divide large data sets into smaller sets of records through a set of decision rules. The way Decision Tree works is to start by initializing training data with features (predictions) and labels (targets). Then data splitting is carried out (splitting) data on nodes that are divided based on the selected feature, each branch of the node represents one of the values or ranges of the selected feature [5], [18]. After that the process of selecting features and separating data is carried out in each subset of data, this process is repeated until it reaches the leaves of the tree, and the labels on the leaves can be predicted for new data.

2.1.5 Extra Tree Classifier

According to Zhang, the Extra Tree (ET) or Extremely Randomized Tree is an algorithm that works like an RF algorithm. But in the *tagging process,* ET does not choose based on the previous tree but instead chooses randomly. Then choose which tree is the best, through optimization. ET can again replace a subset of the data set by recalling the entire subset or sample data so that the accuracy of the model can be improved [10].

2.2 Dataset

From the dataset there are 17 variables/features with 2111 records. The parameters used for obesity prediction to support this study are age, gender, height, weight, family history, eating habit (frequent consumption of high caloric food (FAVC), frequency of consumption of vegetables (FCVC), number of main meals (NCP), consumption of food between meals (CAEC), consumption of water daily (CH20), smoke monitoring, and consumption of alcohol (CALC)), physical condition (calories consumption monitoring (SCC), physical activity frequency (FAF)), time using technology devices (TUE), and transportation used (MTRANS).

2.3 Performance Metrics

2.3.1 Confusion Matrix

Confusion matrix is a method that aims to measure the extent to which a system is able to predict data accurately, by paying attention to how well the system is able to correctly predict predetermined classes. Through the confusion matrix, the performance of the system can be understood by looking at the comparison between the predicted value and the actual value in the multiclass classification. This approach provides a more comprehensive picture of the accuracy and reliability of the system in predicting data, which is a crucial step in evaluating the effectiveness of the developed model in addressing classification problems, or it can be seen as Table 1 [19].

True positive and true negative are the correct classifications on each label, while false negatives and false positives are the result of incorrect classifications. Evaluation measurements use accuracy, recall, and precision.

2.3.2 Accuracy

Accuracy is the comparison between True Positive and True Negative data with all True Positive counts, True Negative, False Positive, and False Negative. The accuracy equation uses the formula.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

2.3.3 Recall

Recall is the ratio between True Positive (TP) data and the sum of all actual data that is indeed positive. The equation for calculating the recall can use the following formula.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

2.3.4 Precision

Precision is a comparison between TP and a lot of data that is predicted to be positive. The precision equation uses the formula.

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

Table 1. Confusion matrix

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

3. Results and Discussions

Based on the data taken from Kaggle some parameters do not have a value in numeric, then in the initial process after the data is imported into the colab.research.google.com changes in the value of data are made in each attribute or parameter including age, gender, height, weight, eating habit attributes (frequent consumption of high caloric food (FAVC), frequency of consumption of vegetables (FCVC), number of main meals (NCP), consumption of food between meals (CAEC), consumption of water daily (CH20), and consumption of alcohol (CALC)), physical condition (calories consumption monitoring (SCC), physical activity frequency (FAF)), time using technology devices (TUE), and transportation used (MTRANS).

3.1 Results

After getting the raw data, the next step is to dance the unique values to generate a list of unique values in the list. Information of the column: consumption of alcohol (CALC), frequent consumption of high caloric food (FAVC), frequency of consumption of vegetables (FCVC). After searching for unique values, the next step is to find the normalization result to change the values in the dataset so that they are on the same scale. This technique is especially important when we use machine learning algorithms that are sensitive to the scale of features. The data presented in Table 2 and Table 3.

Table 2. Results of the unique value calculation table.

No	Age	Gender	Height	Weight	CALC	FAVC	FCVC
1	14	Male	1,71	72,00	no	yes	3,00
2	15	Female	1,65	86,00	no	yes	3,00
3	16	Female	1,55	45,00	no	yes	2,00
4	16	Female	1,55	54,93	Sometimes	yes	1,75
5	16	Female	1,57	49,00	Sometimes	yes	2,00
6	16	Female	1,60	57,00	no	yes	3,00
7	16	Female	1,60	65,00	no	yes	2,96
8	16	Female	1,60	65,00	Sometimes	yes	2,54
9	16	Female	1,61	65,00	no	yes	1,00
10	16	Female	1,61	65,00	no	yes	1,32
11	16	Female	1,61	66,74	no	yes	2,21
12	16	Female	1,62	65,06	no	yes	2,39
13	16	Female	1,62	67,18	no	yes	1,62
14	16	Female	1,62	67,91	no	yes	2,85
15	16	Female	1,63	85,80	no	yes	2,06
16	16	Female	1,64	67,44	no	yes	1,31
17	16	Female	1,65	85,58	no	yes	2,95
18	16	Female	1,66	58,00	no	no	2,00
19	16	Female	1,71	45,25	Sometimes	yes	2,91
20	16	Female	1,74	50,00	Sometimes	yes	2,19
21	16	Female	1,75	49,93	Sometimes	yes	2,49
22	16	Female	1,78	44,76	Sometimes	yes	2,91
23	16	Female	1,82	47,12	Sometimes	yes	3,00
24	16	Female	1,83	43,53	Sometimes	yes	2,95
25	16	Male	1,67	50,00	no	yes	2,00

26	16	Male	1,69	50,00	Sometimes	yes	2,00
27	16	Male	1,69	52,63	Sometimes	yes	2,00
28	16	Male	1,75	50,00	Sometimes	yes	2,31
29	16	Male	1,82	71,00	Sometimes	yes	2,00

Table 3. Results of normalization

No	Age	Gender	Height	Weight	CALC	FAVC	FCVC
1	0,00	Male	0,49	0,25	no	yes	1,00
2	0,02	Female	0,38	0,35	no	yes	1,00
3	0,04	Female	0,19	0,04	no	yes	0,50
4	0,04	Female	0,19	0,12	Sometimes	yes	0,38
5	0,04	Female	0,23	0,07	Sometimes	yes	0,50
6	0,04	Female	0,28	0,13	no	yes	1,00
7	0,04	Female	0,28	0,19	no	yes	0,98
8	0,04	Female	0,28	0,19	Sometimes	yes	0,77
9	0,04	Female	0,30	0,19	no	yes	0,00
10	0,04	Female	0,30	0,19	no	yes	0,16
11	0,04	Female	0,30	0,21	no	yes	0,61
12	0,04	Female	0,32	0,19	no	yes	0,70
13	0,04	Female	0,32	0,21	no	yes	0,31
14	0,04	Female	0,32	0,22	no	yes	0,93
15	0,04	Female	0,34	0,35	no	yes	0,53
16	0,04	Female	0,36	0,21	no	yes	0,16
17	0,04	Female	0,38	0,35	no	yes	0,98
18	0,04	Female	0,40	0,14	no	no	0,50
19	0,04	Female	0,49	0,05	Sometimes	yes	0,96
20	0,04	Female	0,55	0,08	Sometimes	yes	0,60
21	0,04	Female	0,57	0,08	Sometimes	yes	0,75
22	0,04	Female	0,62	0,04	Sometimes	yes	0,96
23	0,04	Female	0,70	0,06	Sometimes	yes	1,00
24	0,04	Female	0,72	0,03	Sometimes	yes	0,98
25	0,04	Male	0,42	0,08	no	yes	0,50
26	0,04	Male	0,45	0,08	Sometimes	yes	0,50
27	0,04	Male	0,45	0,10	Sometimes	yes	0,50
28	0,04	Male	0,57	0,08	Sometimes	yes	0,66
29	0,04	Male	0,70	0,24	Sometimes	yes	0,50

After finding the normalization result, the last step is to calculate the change numeric category: Convert the category to an integer number. Each unique category is assigned a sequential number. Information of the column : frequency of consumption of vegetables (FCVC), gender female (GF), gender male GM), consumption of alcohol no (CALCN), consumption of alcohol sometimes (CALCS), consumption of alcohol always (CALCA), consumption of alcohol frequently(CALCF), frequent consumption of high caloric food yes (FAVCY), frequent consumption of high caloric food no (FAVCN). The data presented in Table 4.

Table 4. Results of change numeric category

No	Age	GM	GF	Height	Weight	CALCN	CALCS	CALCA	CALCF	FAVCY	FAVCN	FCVC
1	0,00	1,00	0,00	0,49	0,25	1,00	0,00	0,00	0,00	1,00	0,00	1,00
2	0,02	0,00	1,00	0,38	0,35	1,00	0,00	0,00	0,00	1,00	0,00	1,00
3	0,04	0,00	1,00	0,19	0,04	1,00	0,00	0,00	0,00	1,00	0,00	0,50
4	0,04	0,00	1,00	0,19	0,12	0,00	1,00	0,00	0,00	1,00	0,00	0,38
5	0,04	0,00	1,00	0,23	0,07	0,00	1,00	0,00	0,00	1,00	0,00	0,50
6	0,04	0,00	1,00	0,28	0,13	1,00	0,00	0,00	0,00	1,00	0,00	1,00
7	0,04	0,00	1,00	0,28	0,19	1,00	0,00	0,00	0,00	1,00	0,00	0,98
8	0,04	0,00	1,00	0,28	0,19	0,00	1,00	0,00	0,00	1,00	0,00	0,77
9	0,04	0,00	1,00	0,30	0,19	1,00	0,00	0,00	0,00	1,00	0,00	0,00
10	0,04	0,00	1,00	0,30	0,19	1,00	0,00	0,00	0,00	1,00	0,00	0,16
11	0,04	0,00	1,00	0,30	0,21	1,00	0,00	0,00	0,00	1,00	0,00	0,61
12	0,04	0,00	1,00	0,32	0,19	1,00	0,00	0,00	0,00	1,00	0,00	0,70
13	0,04	0,00	1,00	0,32	0,21	1,00	0,00	0,00	0,00	1,00	0,00	0,31
14	0,04	0,00	1,00	0,32	0,22	1,00	0,00	0,00	0,00	1,00	0,00	0,93
15	0,04	0,00	1,00	0,34	0,35	1,00	0,00	0,00	0,00	1,00	0,00	0,53
16	0,04	0,00	1,00	0,36	0,21	1,00	0,00	0,00	0,00	1,00	0,00	0,16
17	0,04	0,00	1,00	0,38	0,35	1,00	0,00	0,00	0,00	1,00	0,00	0,98

18	0,04	0,00	1,00	0,40	0,14	1,00	0,00	0,00	0,00	0,00	1,00	0,50
19	0,04	0,00	1,00	0,49	0,05	0,00	1,00	0,00	0,00	1,00	0,00	0,96
20	0,04	0,00	1,00	0,55	0,08	0,00	1,00	0,00	0,00	1,00	0,00	0,60
21	0,04	0,00	1,00	0,57	0,08	0,00	1,00	0,00	0,00	1,00	0,00	0,75
22	0,04	0,00	1,00	0,62	0,04	0,00	1,00	0,00	0,00	1,00	0,00	0,96
23	0,04	0,00	1,00	0,70	0,06	0,00	1,00	0,00	0,00	1,00	0,00	1,00
24	0,04	0,00	1,00	0,72	0,03	0,00	1,00	0,00	0,00	1,00	0,00	0,98
25	0,04	1,00	0,00	0,42	0,08	1,00	0,00	0,00	0,00	1,00	0,00	0,50
26	0,04	1,00	0,00	0,45	0,08	0,00	1,00	0,00	0,00	1,00	0,00	0,50
27	0,04	1,00	0,00	0,45	0,10	0,00	1,00	0,00	0,00	1,00	0,00	0,50
28	0,04	1,00	0,00	0,57	0,08	0,00	1,00	0,00	0,00	1,00	0,00	0,66
29	0,04	1,00	0,00	0,70	0,24	0,00	1,00	0,00	0,00	1,00	0,00	0,50

Table 5. Accuracy model evaluation table

Classification Method	Accuracy Score	% Accuracy	Recall Score	% Recall	Precision Score	% Precision
Random Forest	0,93	93%	0,93	93%	0,94	94%
LGBM	0,97	97%	0,97	97%	0,97	97%
Decision Tree	0,92	92%	0,92	92%	0,92	92%
Extra Tree	0,90	90%	0,93	93%	0,94	94%

After doing the data pre-processing stage from looking for unique values to change numeric category, the next step is the classification method.

3.2 Discussions

The classification results are presented in Table 5. Based on the test results using random forest classifier, it was found that the score accuracy was 0.93 or 93%. Based on the results of data processing using the decision tree classifier technique, it was found that the score accuracy was 0.92 or 92%. Based on the results of data processing using the LGBM Classifier technique, it was found that the score accuracy was 0.97 or 97%. Based on the results of data processing using the Extra Tree Classifier technique, it was found that the accuracy of the score was 0.90 or 90%. Based on the 4 (four) Classification techniques that have been used, then a comparison of accuracy scores can be seen on Based on the existing results, it can be seen that the LGBM Classifier technique has the best accuracy.

Based on the test results using random forest classifier, it was found that the recall score was 0.93 or 93%. Based on the results of data processing using the decision tree classifier technique, it was found that the score accuracy was 0.92 or 92%. Based on the results of data processing using the LGBM Classifier technique, it was found that the score accuracy was 0.97 or 97%. Based on the results of data processing using the Extra Tree Classifier technique, it was found that the score accuracy was 0.93 or 93%. Based on the 4 (four) Classification techniques that have been used, then a comparison of accuracy scores can be seen on Based on the existing results, it can be seen that the LGBM Classifier technique has the best recall.

Based on the test results using random forest classifier, it was found that the precision score was 0.94 or 94%. Based on the results of data processing using the decision tree classifier technique, it was found that the score accuracy was 0.92 or 92%. Based on the results of data processing using the LGBM Classifier technique, it was found that the score accuracy was 0.97 or 97%. Based on the results of data processing using the Extra Tree Classifier technique, it was found that the score accuracy was 0.94 or 94%. Based on the 4 (four) Classification techniques that have been used, then a comparison of accuracy scores can be seen on Based on the existing results, it can be seen that the LGBM Classifier technique has the best precision.

4. Conclusion

Our research concluded that machine learning algorithms can be used to make predictions of obesity levels. In each technique of the machine learning algorithm, it was found to have a different accuracy, recall, and precision score. The Light Gradient Boosting Machine (LGBM) Classifier has the highest accuracy, precision, and recall scores, so the model can be applied to predict obesity levels. This research can open insights for application in other fields that require prediction analysis, classification, and estimation. The researcher's suggestions for this study include adding other classification methods to compare what methods are worth using for this dataset.

References

[1] A. Mutia, J. Jumiyati, and K. Kusdalinah, "Pola Makan Dan Aktivitas Fisik Terhadap Kejadian Obesitas Remaja Pada Masa Pandemi Covid-19," *Journal of Nutrition College*, vol. 11, no. 1, pp. 26–34, 2022, doi: 10.14710/jnc.v11i1.32070.

[2] L. N. W. Fajzrina and R. R. Diana, "Analisis Dampak Obesitas Terhadap Perkembangan Fisik Motorik Anak Usia 5 Tahun," *Ceria: Jurnal Program Studi Pendidikan Anak Usia Dini*, vol. 11, no. 1, pp. 62–74, 2022, doi: 10.31000/CERIA.V11I1.6640.

[3] L. Setiyani, A. N. Indahsari, and R. Roestam, "Analisis Prediksi Level Obesitas Menggunakan Perbandingan Algoritma Machine Learning dan Deep Learning," *JTERA (Jurnal Teknologi Rekayasa)*, vol. 8, no. 1, pp. 139–146, 2023, doi: 10.31544/jtera.v8.i1.2022.139-146.

[4] D. S. Saputra, Jajat, I. Damayanti, K. Sultoni, Y. Ruhayati, and N. I. Rahayu, "Prediksi BMI Berdasarkan Level Aktivitas Fisik dengan Metode Analisis Machine Learning," *Jurnal Pendidikan Kesehatan Rekreasi*, vol. 10, no. 1, pp. 165–175, 2024, doi: 10.59672/jpkr.v10i1.3499.

[5] A. M. Patel and A. Suthar, "AdaBoosted Extra Trees Classifier for Object-Based Multispectral Image Classification of Urban Fringe Area," *International Journal of Image and Graphics*, vol. 22, no. 3, Dec. 2020, doi: 10.1142/S0219467821400064.

[6] R. Ahsana, R. Rohmat Saedudin, and V. P. Widartha, "Perbandingan Akurasi Algoritma Adaboost Dan Algoritma Lightgbm Untuk Klasifikasi Penyakit Diabetes," *e-Proceeding of Engineering*, vol. 8, no. 5, pp. 9738–9748, 2021.

[7] B. Shamreen Ahamed and M. Sumeet Arya, "Prediction of Type-2 Diabetes using the LGBM Classifier Methods and Techniques," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 12, pp. 223–231, 2021.

[8] A. Dharma *et al.*, "Deteksi Pola Pasien Kanker Serviks dengan Algoritma Extra Trees dan K-Nearest Neighbor," *JIKOMSI Jurnal Ilmu Komputer dan Sistem Informasi*, vol. 3, no. 2, pp. 32–36, 2020.

[9] G. Stein, B. Chen, A. S. Wu, and K. A. Hua, "Decision tree classifier for network intrusion detection with GA-based feature selection," *Proceedings of the Annual Southeast Conference*, vol. 2, pp. 2136–2141, 2005, doi: 10.1145/1167253.1167288.

[10] A. Satria, R. M. Badri, and I. Safitri, "Prediksi Hasil Panen Tanaman Pangan Sumatera dengan Metode Machine Learning," *Digital Transformation Technology*, vol. 3, no. 2, pp. 389–398, 2023, doi: 10.47709/digitech.v3i2.2852.

[11] L. Sari, A. Romadloni, R. Lityaningrum, and H. D. Hastuti, "Implementation of LightGBM and Random Forest in Potential Customer Classification," *TIERS Information Technology Journal*, vol. 4, no. 1, pp. 43–55, 2023, doi: 10.38043/tiers.v4i1.4355.

[12] T. Hidayatulloh *et al.*, "Klasifikasi Tipe Berat Tubuh Menggunakan Metode Support Vector Machine," vol. 18, no. 1, pp. 71–77, 2023.

[13] N. Nisrina, F. Fahdhienie, and Rahmadhaniah, "Hubungan Aktivitas Fisik, Umur dan Jenis Kelamin Terhadap Obesitas Pekerja Kantor Bupati Aceh Besar," *Jurnal Promotif Preventif*, vol. 6, no. 5, pp. 746–752, 2023, doi: https://doi.org/10.47650/jpp.v6i5.973.

[14] A. Riadi and R. Sulaehani, "Analisis Implementasi Preprocessing Dengan Otsu-Gaussian Pada Pengenalan Wajah," *ILKOM Jurnal Ilmiah*, vol. 11, no. 3, pp. 200–205, 2019, doi: 10.33096/ilkom.v11i3.457.200-205.

[15] H. Nhat-Duc and T. Van-Duc, "Comparison of histogram-based gradient boosting classification machine, random Forest, and deep convolutional neural network for pavement raveling severity classification," *Automation in Construction*, vol. 148, p. 104767, Apr. 2023, doi: 10.1016/J.AUTCON.2023.104767.

[16] F. Hamami and I. A. Dahlan, "Klasifikasi Cuaca Provinsi Dki Jakarta Menggunakan Algoritma Random Forest Dengan Teknik Oversampling," *Jurnal Teknoinfo*, vol. 16, no. 1, pp. 87–92, 2022, doi: 10.33365/jti.v16i1.1533.

[17] A. Hussain and A. Aslam, "Cardiovascular Disease Prediction Using Risk Factors: A Comparative Performance Analysis of Machine Learning Models," *Journal on Artificial Intelligence*, vol. 6, no. 1, pp. 129–152, 2024, doi: 10.32604/jai.2024.050277.

[18] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis : Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 13, no. 2, pp. 67–75, 2020, doi: 10.51903/e-bisnis.v13i2.247.

[19] M. A. P. Sri, A. Wahyono, and M. A. Aziz, "Deteksi Pola Kejadian Bencana Menggunakan Algoritma Naïve Bayes di Kabupaten Boyolali," *Jitu*, vol. 8, no. 1, pp. 97–106, 2024, doi: https://doi.org/10.36596/jitu.v8i1.1119.