Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

SEQUENTIAL DATA PREPROCESSING APPROACH FOR ENHANCED MATERNAL HEALTH RISK CLASSIFICATION PERFORMANCE

Ridha Adjie Eryadi[1], Wildan Hanif Hafidudin[2]

[12,]Department of Informatics, Digital Technology University, Kota Bandung, Indonesia

Jalan Cibogo Indah III, Bodogol, Kota Bandung
[1] ridhaadjie@digitechuniversity.ac.id, [2] wildan20123074@digitechuniversity.ac.id

Abstract

Maternal mortality is still a major health issue worldwide, and it, along with other reasons, has been leading to predictions that need risk-assessing systems to be improved. The current study performed the sequential outlier detection combining Interquartile Range followed by Local Outlier Factor methods on six machine learning algorithms using the UCI Maternal Health Risk dataset. The comprehensive preprocessing pipeline included the removal of duplicates, application of SMOTE for balancing, followed by Min-Max normalization and detection of outliers in a sequence. The performance of the model was evaluated through holdout validation and 10-fold cross-validation with statistical validation through Wilcoxon signed-rank tests and Cohen's d effect sizes. The Extra Trees Classifier resulted in a 98.34% accuracy rate, which is higher than that in previous studies. The distance-based methods showed the highest sensitivity, with KNN gaining 8.35% while tree-based ensembles were consistent with the accuracy gains. The statistical validation proved that there was a great extent of practical significance with a large effect size of more than 1.0 for the top performers, thereby establishing evidence-based guidelines for the application of sequential preprocessing in maternal health risk prediction systems.

Keywords : *Extra Trees Classifier, Interquartile Range, Local Outlier Factor, maternal health risk prediction, sequential outlier detection, SMOTE*

Abstrak

Angka kematian ibu hamil masih menjadi masalah kesehatan utama di seluruh dunia dan, bersama dengan alasan lain, telah menyebabkan prediksi bahwa sistem penilaian risiko perlu ditingkatkan. Studi ini melakukan deteksi outlier sekuensial yang menggabungkan metode Interquartile Range diikuti oleh Local Outlier Factor pada enam algoritma pembelajaran mesin menggunakan dataset UCI Maternal Health Risk. *Pipeline* pra-pemrosesan yang komprehensif mencakup penghapusan duplikat, penerapan SMOTE untuk penyeimbangan, diikuti oleh normalisasi Min-Max dan deteksi outlier secara berurutan. Kinerja model dievaluasi melalui validasi holdout dan validasi silang 10-fold dengan validasi statistik melalui uji peringkat bertanda Wilcoxon dan ukuran efek Cohen's d. Pengklasifikasi Extra Trees menghasilkan tingkat akurasi 98,34% yang lebih tinggi daripada studi sebelumnya. Metode berbasis jarak menunjukkan sensitivitas tertinggi dengan KNN memperoleh 8,35% sementara ensemble berbasis pohon konsisten dengan peningkatan akurasi. Validasi statistik membuktikan bahwa terdapat signifikansi praktis yang besar dengan ukuran efek yang besar, yaitu lebih dari 1.0, untuk para pelaku terbaik, sehingga menetapkan pedoman berbasis bukti untuk penerapan pra-pemrosesan sekuensial dalam sistem prediksi risiko kesehatan ibu hamil.

Kata kunci : *Extra Trees Classifier, Interquartile Range, Local Outlier Factor,* prediksi risiko kesehatan ibu hamil, deteksi outlier berurutan, *SMOTE*

1. INTRODUCTION

Maternal mortality continues to pose a considerable challenge to global health, with an estimated 287,000 maternal deaths occurring each year, leading to a maternal mortality ratio of 211 deaths per 100,000 live births globally [1]. Despite significant advancements in recent years, Indonesia still faces a substantial issue with maternal mortality. In 2020, the Ministry of Health reported a maternal mortality rate of 177 deaths per 100,000 live births, positioning Indonesia among the Southeast Asian nations with the highest rates of maternal mortality [2]. According to the Rencana Pembangunan Jangka Menengah Nasional (RPJMN) 2020-2024 [3], the Indonesian government aims to reduce the number of maternal deaths to 131 per 100,000 live births by 2024. Complications during pregnancy, such as high blood pressure, diabetes, and heart disease, significantly increase the risks associated with maternal morbidity and mortality. In Indonesia, over 75% of maternal deaths are attributed to hemorrhage, hypertension, and infections [4].

Machine learning technologies have significantly impacted healthcare prediction systems, improving maternal health outcomes via data-driven risk assessment. The UCI Maternal Health Risk dataset has become a vital resource for academics seeking to create and evaluate predictive models for categorizing maternal health risk [5, 6]. Healthcare data analytics has significantly progressed, with machine learning algorithms demonstrating accuracies ranging from 70% to 98%, contingent upon the techniques utilized [7, 8]. These breakthroughs enable healthcare workers to identify high-risk pregnancies, allowing for prompt intervention. This could significantly reduce the mortality rate of women during pregnancy. The quality of the data and the methodologies employed for its preparation are crucial for the effective functioning of predictive systems. This underscores the necessity of rigorous approaches for data purification and outlier detection in maternal health forecasting [9].

Most current research uses single outlier detection methods on their own, with the Interquartile Range (IQR) being the most common one [10, 11, 12]. Density-based methods like Local Outlier Factor (LOF) are still not well studied in maternal health forecasting. SMOTE has become the leading oversampling method in several studies [5, 11, 13], successfully mitigating class imbalance in healthcare datasets. Most studies concentrate on discrete preprocessing stages or singular algorithm optimization, neglecting a thorough examination of the sequential impacts of outlier removal across diverse classifier types (tree-based, distance-based, linear, and ensemble methods) [14].

This study seeks to fill the identified research gaps by examining the cumulative effects of sequential outlier detection utilizing IQR, followed by LOF, on maternal health risk prediction models. This study aims to examine the effects of sequential outlier removal on dataset attributes and model efficacy, to evaluate the sensitivity of various machine learning algorithms to sequential preprocessing, and to confirm performance enhancements through statistical methods, including the Wilcoxon signed-rank test and Cohen’s d effect size.

2. LITERATURE REVIEW

Recent analyses of the UCI Maternal Health Risk dataset demonstrate that model efficacy is significantly affected by preprocessing methodologies and algorithmic choices. Traditional methods that use only one classifier have been able to reach accuracy levels between 70.21% and 98%, depending on the methods used [7, 8]. Pawar et al. [7] utilized Random Forest with feature ranking and IQR preprocessing, whereas Raza et al. [8] achieved better accuracy by using SVM with DT-BiLTCN feature extraction. Other studies using LightGBM, Decision Tree, and Random Forest also showed how important it is to choose the right features and get rid of outliers [6, 10, 12]. The studies demonstrate that preprocessing decisions impact performance results, yet researchers typically analyze these decisions as separate components.

For improving predictions substantially more accurately, more advanced hybrid and ensemble-based methods have been suggested. Raihen and Akter [11] integrated SVM with RFE and SMOTE, whereas Togunwa et al. [13] presented a hybrid ANN–Random Forest model. Jamel et al. [15] attained elevated accuracy through PCA-based feature engineering and a voting classifier. These methods certainly render things precise, but they mostly focus on how complicated the model is

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

instead of how to perform preprocessing strategies.

IQR is the most commonly used method for finding outliers in healthcare analytics because it is both strong and easy to use [16, 17]. Density-based methods like LOF can find local multivariate anomalies, which makes them useful in addition to other methods [18–20]. Nevertheless, in maternal health prediction studies, LOF is infrequently integrated with statistical methods and is commonly utilized in isolation.

Class imbalance remains to be considered a major problem in maternal health datasets, and SMOTE has been widely used to fix this problem [21]. Prior research indicates that SMOTE enhances classification performance, although its efficacy may be compromised by the existence of outliers in minority classes [22]. Even with this dependency, not many studies look at how to handle outliers and oversample in a single preprocessing framework.

There aren't many statistical validation methods that can be applied in research on predicting maternal health risks. Many studies report accuracy, but fewer use non-parametric significance tests or effect size measures to check how reliable and useful performance improvements are [23]. Consequently, the accuracy of reported gains frequently remains ambiguous.

After peering at the literature, we understand that there are some gaps. Most studies use a single approach to find outliers, focus on one algorithm at a time, and don't use strict statistical validation [6, 10–12, 15]. This study, on the contrary hand, looks at sequential outlier detection, compares different algorithm families, and uses statistical importance and effect size analysis to give a more complete picture.

3. RESEARCH METHOD

The research evaluates the implications of sequential outlier detection on maternal health risk prediction by employing various machine learning algorithms within a systematic preprocessing framework. The overall workflow includes cleaning the data, finding outliers in order, balancing the classes, normalizing the features, training the model, and testing it. The Interquartile Range (IQR) method serves first to find sequential outliers, and then the Local Outlier Factor (LOF) method is applied to identify both statistical and density-based anomalies in the dataset.

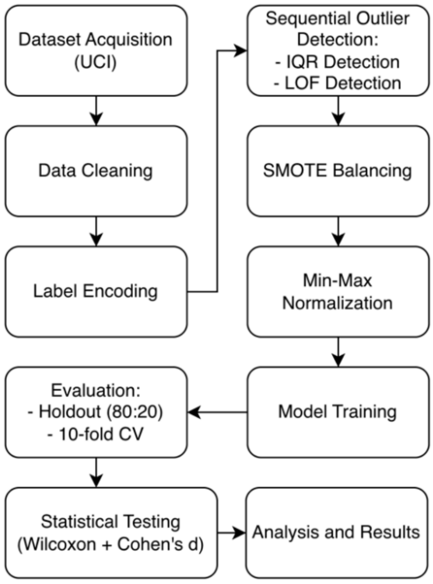

Figure 1. Research Methodology

3.1. Dataset Description

The UCI Maternal Health Risk dataset is obtained from the UCI Machine Learning Repository. It has 1,014 maternal health records that were collected using an IoT-based monitoring system [24]. The dataset has six numerical features: Age, Systolic Blood Pressure, Diastolic Blood Pressure, Blood Sugar, Body Temperature, and Heart Rate. It also has one categorical target variable that shows the level of risk for pregnant women. The dataset has no missing values, but it has class imbalance and outliers, which are common in medical data from the real world [5].

3.2. Data Preprocessing Pipeline

3.2.1. Data Cleaning and Encoding

To avoid bias and overfitting, duplicate records were removed, leaving a smaller dataset of 799 instances [9]. Label Encoding was used to change the categorical risk level variable into a number, which is what is usually done in ordinal encoding [25].

3.2.2. Sequential Outlier Detection

The outlier detection process has two steps, and it employs a combination of statistical and

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

density-based methods. First, the Interquartile Range (IQR) method was used, and the calculation was as follows:

$$IQR = Q_3 - Q_1 \quad (1)$$

Outliers are indicated by the observations x that are outside the range:

$$x < Q_1 - 1.5 \times IQR \text{ or } x > Q_3 + 1.5 \times IQR \quad (2)$$

This method has a robust statistical detection of outliers, which is not influenced or affected by extreme values [17]. Then, the LOF technique was employed along with the best hyperparameters (n_neighbors=20, contamination=0.05), which have been determined through grid search. LOF determines the local density deviation of a data point concerning its neighbors using the following equation:

$$LOF_k(p) = \frac{\sum_{o \in N_k(p)} \frac{lrd_k(o)}{|N_k(p)|}}{lrd_k(p)} \quad (3)$$

where $lrd_k(p)$ stands for local reachability density and $N_k(p)$ for k-nearest neighbors [19]. The technique based on density is capable of recognizing multivariate outliers while traditional statistical methods might overlook them [18].

3.3. Feature Scaling and Data Balancing

By using Min-Max normalization on their data, this study was able to get all of their features to have the same range. Distance-based classifiers need this method because their algorithms need all data points to have the same value distribution. The researchers used SMOTE to make fake samples, which helped them fix the problem of class imbalance by making fake samples for the minority classes. The document will show how SMOTE works through demonstrating how it uses standard interpolation-based oversampling methods to do its job, as shown in [21, 22].

3.4. Machine Learning Algorithms

We used six machine learning algorithms to see how sequential preprocessing affected different model families. These algorithms have been extensively utilized in healthcare prediction tasks, encompassing Extra Trees and Random Forest for ensemble learning [26, 27], XGBoost for gradient boosting [28, 29], K-Nearest Neighbors for distance-based classification [30, 31], Logistic Regression for baseline linear modeling [32], and Decision Tree for rule-based classification [33].

The Extra Trees Classifier makes a group of very random decision trees, and the final prediction is shown as:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} T_m(x) \quad (4)$$

where T_m denotes individual trees, and M is the total number of trees.

In the same way, Random Forest makes predictions by having multiple decision trees trained on bootstrap samples that vote on the best one:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \quad (5)$$

To find a balance between model complexity and accuracy, the XGBoost model uses gradient boosting with regularization. It has a regularization term that looks like this:

$$\Omega(f_k) = \gamma T + \frac{\lambda}{2} \sum_{j=1}^{T} w_j^2 \quad (6)$$

Here, T is the number of leaves and w_j is the weight of each leaf.

The K-Nearest Neighbors (KNN) classifier uses Euclidean distance to give class labels based on how close points are in feature space:

$$d(x, x') = \sqrt{\sum_{i=1}^{p} (x_i - x_i')^2} \quad (7)$$

This property makes KNN very sensitive to scaling features and having outliers.

The sigmoid function is used by the Logistic Regression model to figure out the chances of each class:

$$P(y = 1|x) = \frac{1}{1+e^{-z}} \quad (8)$$

where z is a linear combination of the input features.

Lastly, the Decision Tree model picks the best splits based on Information Gain, which is calculated as:

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} H(S_v) \quad (9)$$

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

This criterion allows for recursive partitioning of the dataset into subsets that are more and more similar to each other.

3.5. Model Optimization and Evaluation

The study employed GridSearchCV in conjunction with stratified 10-fold cross-validation to enhance hyperparameter tuning outcomes, thereby validating the model configuration and delivering an equitable assessment of model performance [11]. The model evaluation utilized an 80:20 train-test split alongside conventional multi-class metrics, including accuracy, precision, recall, and F1-score [34, 35].

- **Accuracy** = (TP + TN) / (TP + TN + FP + FN)
- **Precision** = TP / (TP + FP)
- **Recall** = TP / (TP + FN)
- **F1-Score** = 2 × (Precision × Recall) / (Precision + Recall)

3.6. Statistical Validation

We utilized the Wilcoxon signed-rank test to evaluate performance disparities between two groups and employed Cohen's d effect size to ascertain the practical significance of these findings. The study offers fundamental statistical data, featuring standard statistical techniques and recognized criteria for the interpretation of research results [36–38].

4. RESULT AND DISCUSSION

4.1. Experimental Setup

The experiment was performed in a Google Colaboratory (Colab) setting, which allowed the running of Python scripts, the use of cloud computing facilities, and the availability of various ML tools. The UCI Maternal Health Risk dataset was subjected to comprehensive preprocessing procedures, encompassing data cleansing, outlier identification, and class balancing.

4.2. Data Preprocessing Results

Table 1 shows how preprocessing changes the characteristics of a dataset. After removing duplicates, the dataset size went from 1,014 to 799 instances, which shows that there were extra records. Using IQR to find outliers in order, followed by LOF, cut the dataset down to 463 instances. This demonstrated that there were actually a lot of unexpected observations, specifically in the Mid Risk and High Risk classes. These categories are of clinical significance as they refer to pregnancies necessitating enhanced examination.

Table 1. Dataset Characteristics Across Processing Stages

Processing Stage	Total	High Risk	Mid Risk	Low Risk
Raw Data	1,014	272	336	406
After Cleaning	799	233	221	345
After IQR+LOF	463	43	119	301
After SMOTE	903	301	301	301

From a practical point of view, this result shows how important it is to have strong preprocessing in real-world maternal health systems, where sensor-based measurements and manual data collection can be noisy. Removal of abnormal observations makes the data more probable and helps automated risk assessments be more accurate. The next step proved to use SMOTE, that's successfully balanced the class distributions and made sure that all risk categories learned equally well.

4.3. Model Performance Evaluation

Table 3 summarizes all of the evaluation results, which look at the way the model applies in two distinct methods: through the use of internal performance metrics and by comparing it to other studies. The use of sequential preprocessing with IQR and LOF made most of the algorithm families tested in this accuracy research improvements.

Table 2. Model Performance Comparison

Model	Raw Data	After IQR+LOF	Improvement
Extra Trees	92.27%	98.34%	+6.07%
KNN	88.89%	97.24%	+8.35%
Random Forest	90.82%	96.69%	+5.87%
XGBoost	90.82%	96.13%	+5.31%
Decision Tree	91.30%	94.48%	+3.18%

ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

Model	Raw Data	After IQR+LOF	Improvement
Logistic Regression	67.63%	66.30%	-1.33%

Logistic Regression, on the other hand, did not perform as well after removing outliers. This finding suggests that linear models may show less receptivity to anomalous observations within this dataset, emphasizing the necessity for preprocessing strategies to be adapted to machine learning characteristics rather than applied uniformly.

4.4. Comparative and Practical Implications

The proposed approach exceeds or matches state-of-the-art methods when compared to previous studies, as shown in Table 3, utilizing the same dataset, all while leveraging simpler model architectures. Previous studies that demonstrate similar accuracy typically depend on sophisticated feature engineering or hybrid deep learning frameworks. This study, on the reverse side, reveals that systematic data preprocessing alone might enhance the performance of standard machine learning algorithms.

Table 3. Comparison with Previous Studies

Study	Method	Accuracy
This Study	**Extra Trees + IQR+LOF**	**98.34%**
Jamel et al. [15]	Voting (ETC+MLP) + PCA	98.25%
Raza et al. [8]	SVM + DT-BiLTCN	98.00%
Togunwa et al. [13]	MaternalNET-RF + IQR	95.00%
Mutlu et al. [12]	Decision Tree + IQR	89.16%
Khadidos et al. [5]	Ensemble Stacking	86.00%
Raihen & Akter [11]	SVM + RFE	86.13%
Noviandy et al. [6]	LightGBM + IQR	84.73%
Alamsyah et al. [10]	Random Forest + PCA	82.18%
Pawar et al. [7]	Random Forest + IQR	70.21%

This finding is essential from the point standpoint of placing the result into implementation. In medical environments with a shortage of assets, it is often simpler to enhance preprocessing pipelines than to use models that require a lot of computing capacity. The proposed sequential outlier detection framework improves accuracy, stability, and fairness, making it an appropriate decision for use in hospitals, community health centers, and IoT-based systems for monitoring maternal health.

In general, the results show that sequential outlier detection is a useful and effective way of enhancing predictions of maternal health risks. This helps to identify high-risk pregnancies early and makes clinical decision support more reliable.

5. CONCLUSION AND SUGGESTIONS

The current study validates that sequential outlier detection leveraging the Interquartile Range (IQR) performed effectively by the Local Outlier Factor (LOF) considerably improves the effectiveness and resilience of machine learning models for predicting maternal health risks. The proposed preprocessing framework lifts the accuracy of multiple different kinds of machine learning algorithms by a lot. The Extra Trees Classifier had the highest accuracy of 98.34% on the UCI Maternal Health Risk dataset. The Wilcoxon signed-rank test and Cohen's d effect size demonstrate just how these improvements are both statistically significant and practically significant. It demonstrates the necessity that the matter remains to systematic preprocess data in reliable maternal health risk prediction systems.

Future studies would need to assess the practical use of this framework using maternal health datasets collected across different clinical and geographic settings. Further enhancements can be made by applying adaptive preprocessing methods, using explainable AI methods in order to render clinical data easier for users to comprehend, and extending the framework to handle real-time or streaming maternal health data for assistance using continuous monitoring of risks and early intervention.

REFERENCES:

[1] World Health Organization, "Maternal mortality: Key facts," WHO, 2020.

[2] Kementerian Kesehatan Republik Indonesia, "Profil kesehatan Indonesia tahun 2020," Kemenkes RI, 2021.
[3] Badan Perencanaan Pembangunan Nasional, "Rencana Pembangunan Jangka Menengah Nasional (RPJMN) 2020-2024," Bappenas, 2020.
[4] Kementerian Kesehatan Republik Indonesia, "Laporan kinerja kementerian kesehatan tahun 2021," Kemenkes RI, 2022.
[5] A. O. Khadidos, F. Saleem, S. Selvarajan, Z. Ullah, and A. O. Khadidos, "Ensemble machine learning framework for predicting maternal health risk during pregnancy," *Scientific Reports*, vol. 14, no. 1, p. 21483, Sep. 2024, doi: 10.1038/s41598-024-71934-x.
[6] T. R. Noviandy, S. I. Nainggolan, R. Raihan, I. Firmansyah, and R. Idroes, "Maternal health risk detection using Light Gradient Boosting Machine approach," *Infolitika Journal of Data Science*, vol. 1, no. 2, pp. 48-55, Dec. 2023, doi: 10.60084/ijds.v1i2.123.
[7] L. Pawar, J. Malhotra, A. Sharma, D. Arora, and D. Vaidya, "A robust machine learning predictive model for maternal health risk," in *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Aug. 2022, pp. 882-888, doi: 10.1109/icesc54411.2022.9885515.
[8] A. Raza, H. U. R. Siddiqui, K. Munir, M. Almutairi, F. Rustam, and I. Ashraf, "Ensemble learning-based feature engineering to analyze maternal health during pregnancy and health risk prediction," *PLoS ONE*, vol. 17, no. 11, p. e0276525, Nov. 2022, doi: 10.1371/journal.pone.0276525.
[9] A. Jarmakovica, "Machine learning-based strategies for improving healthcare data quality: an evaluation of accuracy, completeness, and reusability," *Frontiers in Artificial Intelligence*, vol. 8, p. 1621514, Jul. 2025, doi: 10.3389/frai.2025.1621514.
[10] D. P. Alamsyah, Y. Ramdhani, and L. Susanti, "Maternal Health Risk Classification: Random Forest and Evolutionary Algorithms," in *2023 International Conference on Information Security and Systems (ICISS)*, Sep. 2023, pp. 1-6, doi: 10.1109/iciss59129.2023.10292044.
[11] M. N. Raihen and S. Akter, "Comparative assessment of several effective machine learning classification methods for maternal health risk," *Computational Journal of Mathematical and Statistical Sciences*, vol. 3, no. 1, pp. 161-176, Feb. 2024, doi: 10.21608/cjmss.2024.259490.1036.
[12] H. B. Mutlu, F. Durmaz, N. Yücel, E. Cengil, and M. Yildirim, "Prediction of Maternal Health Risk with Traditional Machine Learning Methods," *NATURENGS MTU Journal of Engineering and Natural Sciences*, Jun. 2023, doi: 10.46572/naturengs.1293185.
[13] T. O. Togunwa, A. O. Babatunde, and K.-U.-R. Abdullah, "Deep hybrid model for maternal health risk classification in pregnancy: synergy of ANN and random forest," *Frontiers in Artificial Intelligence*, vol. 6, p. 1213436, Jul. 2023, doi: 10.3389/frai.2023.1213436.
[14] M. Rahman, R. Noor, S. Mallik, N. Santa, S. Deb, and A. Pathak, "Classification of health risk levels for pregnant women using support vector machine (SVM) algorithm," *IOSR Journal of Computer Engineering*, vol. 26, no. 3, pp. 7-17, 2024, doi: 10.9790/0661-2603010717.
[15] L. Jamel *et al.*, "Improving prediction of maternal health risks using PCA features and TreeNet model," *PeerJ Computer Science*, vol. 10, p. e1982, Apr. 2024, doi: 10.7717/peerj-cs.1982.
[16] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset," in *Advances in Intelligent Systems and Computing*,

2018, pp. 511-518, doi: 10.1007/978-981-10-7563-6_53.

[17] Ch. S. K. Dash, A. K. Behera, S. Dehuri, and A. Ghosh, "An outliers detection and elimination framework in classification task of data mining," *Decision Analytics Journal*, vol. 6, p. 100164, Jan. 2023, doi: 10.1016/j.dajour.2023.100164.

[18] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data and Cognitive Computing*, vol. 5, no. 1, p. 1, Dec. 2020, doi: 10.3390/bdcc5010001.

[19] E. F. Agyemang, "Anomaly detection using unsupervised machine learning algorithms: A simulation study," *Scientific African*, vol. 26, p. e02386, Sep. 2024, doi: 10.1016/j.sciaf.2024.e02386.

[20] A. H. Abuzaid, "Identifying density-based local outliers in medical multivariate circular data," *Statistics in Medicine*, vol. 39, no. 21, pp. 2793-2798, May 2020, doi: 10.1002/sim.8576.

[21] D. Elreedy and A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Information Sciences*, vol. 505, pp. 32-64, Jul. 2019, doi: 10.1016/j.ins.2019.07.070.

[22] S. Matharaarachchi, M. Domaratzki, and S. Muthukumarana, "Enhancing SMOTE for imbalanced data with abnormal minority instances," *Machine Learning With Applications*, vol. 18, p. 100597, Oct. 2024, doi: 10.1016/j.mlwa.2024.100597.

[23] J. Hansen, S. Ahern, and A. Earnest, "Evaluations of statistical methods for outlier detection when benchmarking in clinical registries: a systematic review," *BMJ Open*, vol. 13, no. 7, p. e069130, Jul. 2023, doi: 10.1136/bmjopen-2022-069130.

[24] M. Ahmed, "Maternal health risk," *UC Irvine Machine Learning Repository*, Jan. 2020, doi: 10.24432/c5dp5d.

[25] D. Thakur, T. Gera, V. Bhardwaj, A. A. AlZubi, F. Ali, and J. Singh, "An enhanced diabetes prediction amidst COVID-19 using ensemble models," *Frontiers in Public Health*, vol. 11, p. 1331517, Dec. 2023, doi: 10.3389/fpubh.2023.1331517.

[26] U. J. Nzenwata, E. Edwin, E. A. Chukwu, D. Osilaja, J. O. Hinmikaiye, and C. Enyinnah, "Extra trees model for heart disease prediction," *Journal of Data Analysis and Information Processing*, vol. 13, no. 2, pp. 125-139, Jan. 2025, doi: 10.4236/jdaip.2025.132008.

[27] J. Lötsch and B. Mayer, "A biomedical case study showing that tuning random forests can fundamentally change the interpretation of supervised data structure exploration aimed at knowledge discovery," *BioMedInformatics*, vol. 2, no. 4, pp. 544-552, Oct. 2022, doi: 10.3390/biomedinformatics2040034.

[28] M. Feng, X. Wang, Z. Zhao, C. Jiang, J. Xiong, and N. Zhang, "Enhanced heart attack prediction using eXtreme Gradient boosting," *Journal of Theory and Practice of Engineering Science*, vol. 4, no. 4, pp. 9-16, Apr. 2024, doi: 10.53469/jtpes.2024.04(04).02.

[29] Rizka Dahlia, Lady Agustin Fitriana, and Syarah Seimahuira, "ANALISIS ALGORITMA GRADIENT BOOSTING DALAM PENGARUH MASYARAKAT MEMILIH RUMAH SEWA," misi, vol. 8, no. 1, pp. 35–44, Jan. 2025, doi: 10.36595/misi.v8i1.1356.

[30] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of heart disease using K-nearest neighbor and genetic algorithm," *arXiv preprint arXiv:1508.02061*, May 2015, doi: 10.48550/arxiv.1508.02061.

[31] G. P. A. Brahmantha, E. Utami, and A. Yaqin, "KLASIFIKASI GENRE ANIME BERDASARKAN SINOPSIS MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBORS", *misi*, vol. 7, no. 1, pp. 15–24, Feb. 2024.

[32] M. Iwagami *et al.*, "Comparison of machine-learning and logistic regression models for prediction of 30-day unplanned readmission in electronic health records: A development and validation study,"

Jurnal Manajemen Informatika & Sistem Informasi (MISI)
Volume 9, Nomor 1, Januari 2026.
ISSN 2614-1701 (media cetak)
ISSN 2614-3739 (media online)
DOI : 10.36595/misi.v5i2
http://e-journal.stmiklombok.ac.id/index.php/misi

PLOS Digital Health, vol. 3, no. 8, p. e0000578, Aug. 2024, doi: 10.1371/journal.pdig.0000578.
[33] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, "A comparative study for predicting heart diseases using data mining classification methods," *arXiv preprint arXiv:1704.02799*, Apr. 2017, doi: 10.48550/arxiv.1704.02799.
[34] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, May 2009, doi: 10.1016/j.ipm.2009.03.002.
[35] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," *arXiv preprint arXiv:2008.05756*, Aug. 2020, doi: 10.48550/arxiv.2008.05756.
[36] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1-30, 2006.
[37] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1988.
[38] G. M. Sullivan and R. Feinn, "Using effect size—or why the P value is not enough," *Journal of Graduate Medical Education*, vol. 4, no. 3, pp. 279-282, Sep. 2012, doi: 10.4300/jgme-d-12-00156.1.