PENANGANAN KETIDAKSEIMBANGAN DATA PADA KLASIFIKASI PENYAKIT CAMPAK MENGGUNAKAN KOMBINASI SMOTE DAN XGBOOST

Novita Ranti Muntiari[1], Kharis Hudaiby Hanif[2], Muliyadi[3], Mufida[4]

[1,3,4]Program Studi Promosi Kesehatan, Politeknik Kaltara, Tarakan, Kalimantan Utara, Indonesia
[2]Program Studi Teknik Komputer, Fakultas Teknik, Universitas Borneo Tarakan, Tarakan Kalimantan Utara, Indonesia

e-mail: [1]novitarantimuntiari@gmail.com, [2]hudaiby21@borneo.ac.id, [3]muliyadi@poltekkaltara.ac.id, [4]arianifidha@gmail.com

ABSTRAK

Ketidakseimbangan data menjadi salah satu tantangan utama dalam pengembangan model klasifikasi penyakit, karena dapat menyebabkan algoritma lebih dominan mengenali kelas mayoritas dan kurang optimal dalam mendeteksi kasus positif. Penelitian ini bertujuan untuk menganalisis penerapan kombinasi Synthetic Minority Over-sampling Technique (SMOTE) dan XGBoost dalam klasifikasi penyakit campak. Data yang digunakan berjumlah 1.000 data dengan fitur klinis meliputi usia, riwayat imunisasi, demam, batuk, pilek, konjungtivitis, ruam kulit, dan status campak. Data penelitian dipisahkan ke dalam dua subset, yaitu 80% untuk proses pelatihan model dan 20% untuk pengujian. Teknik SMOTE digunakan pada data pelatihan guna memperbaiki ketimpangan jumlah antar kelas, sementara algoritma XGBoost dimanfaatkan untuk membangun model klasifikasi. Kinerja model kemudian dinilai melalui confusion matrix serta metrik accuracy, precision, recall, dan F1-score. Hasil penelitian menunjukkan bahwa XGBoost tanpa SMOTE memperoleh accuracy 94,0%, precision 83,3%, recall 50,0%, dan F1-score 62,5%. Setelah diterapkan SMOTE, performa meningkat dengan accuracy 97,0%, precision 79,2%, recall 95,0%, dan F1-score 86,4%. Hasil ini menunjukkan bahwa kombinasi SMOTE dan XGBoost lebih efektif dalam meningkatkan kemampuan deteksi kasus positif campak pada data tidak seimbang.

Kata kunci: campak, klasifikasi, SMOTE, XGBoost, ketidakseimbangan data.

ABSTRACT

Data imbalance is one of the main challenges in developing disease classification models, as it can cause algorithms to recognize the majority class more dominantly and perform less optimally in detecting positive cases. This study aims to analyze the application of the combination of Synthetic Minority Over-sampling Technique (SMOTE) and XGBoost in measles disease classification. The data used consisted of 1,000 records with clinical features including age, immunization history, fever, cough, runny nose, conjunctivitis, skin rash, and measles status. The research data were divided into two subsets, namely 80% for the model training process and 20% for testing. The SMOTE technique was applied to the training data to address class distribution imbalance, while the XGBoost algorithm was used to build the classification model. Model performance was then evaluated using a confusion matrix and the metrics of accuracy, precision, recall, and F1-score. The results showed that XGBoost without SMOTE achieved an accuracy of 94.0%, precision of 83.3%, recall of 50.0%, and F1-score of 62.5%. After applying SMOTE, the

performance improved, with an accuracy of 97.0%, precision of 79.2%, recall of 95.0%, and F1-score of 86.4%. These results indicate that the combination of SMOTE and XGBoost is more effective in improving the detection capability of positive measles cases in imbalanced data..

Keywords*: measles, classification, SMOTE, XGBoost, imbalanced data*

1. PENDAHULUAN

Campak adalah penyakit infeksi akut yang disebabkan oleh virus dan memiliki tingkat penularan yang sangat tinggi serta tetap menjadi tantangan serius bagi kesehatan global, termasuk di Indonesia. Penyakit ini ditandai dengan gejala klinis spesifik seperti demam tinggi, batuk, pilek, konjungtivitis, serta munculnya bercak Koplik yang diikuti oleh ruam makulopapular. Meskipun upaya vaksinasi telah dilakukan secara meluas, fluktuasi kasus campak masih sering ditemukan akibat adanya kesenjangan cakupan imunisasi di beberapa wilayah. Klasifikasi kasus secara cepat sebagai prioritas utama dalam sistem surveilans kesehatan [1].

Permasalahan utama dalam diagnosis campak secara konvensional adalah adanya kemiripan gejala awal dengan infeksi saluran pernapasan lainnya, yang berpotensi menimbulkan kesalahan identifikasi. Di sisi lain, pemanfaatan teknologi informasi melalui implementasi *Machine Learning* telah menunjukkan potensi besar dalam membantu pengambilan keputusan medis [2], [3], [4]. Namun, dalam konteks data kesehatan, seringkali ditemui kendala berupa ketidakseimbangan kelas (*imbalanced data*). Dalam dataset campak, jumlah sampel pasien positif biasanya jauh lebih sedikit dibandingkan dengan jumlah pasien negatif. Kondisi ini menyebabkan algoritma klasifikasi cenderung mengalami bias terhadap kelas mayoritas, sehingga kemampuan model untuk mendeteksi pasien positif (kelas minoritas) menjadi sangat rendah, yang secara medis sangat berisiko.

Untuk mengatasi kendala tersebut, penggunaan algoritma yang tangguh seperti *eXtreme Gradient Boosting* (*XGBoost*) menjadi pilihan relevan karena kemampuannya dalam menangani data tabular dengan akurasi tinggi dan efisiensi waktu komputasi yang optimal [5]. Namun, *XGBoost* saja tidak cukup untuk menangani data yang sangat timpang. Oleh karena itu, diperlukan teknik prapemrosesan data seperti *Synthetic Minority Over-sampling Technique* (*SMOTE*) untuk menyeimbangkan distribusi kelas dengan membangkitkan sampel sintetis pada kelas minoritas, sehingga model dapat mempelajari karakteristik gejala campak secara lebih representatif [6], [7], [8].

Penelitian terdahulu telah mengeksplorasi klasifikasi penyakit menular dengan berbagai pendekatan. Pertama, penelitian oleh Prasetyo et al. (2022) dalam jurnal Jurnal Teknologi Informasi dan Ilmu Komputer yang berfokus pada penggunaan algoritma *Support Vector Machine* (*SVM*) untuk klasifikasi penyakit serupa, namun mencatat bahwa performa model menurun saat menghadapi data yang tidak seimbang [9]. Kedua, penelitian yang dilakukan oleh Sari & Wijaya (2023) dalam *Journal of Information Systems Engineering and Business Intelligence* menunjukkan bahwa penerapan teknik *Oversampling* dapat meningkatkan nilai *sensitivity* pada klasifikasi penyakit endemik, akan tetapi penelitian tersebut masih menggunakan algoritma dasar seperti *Decision Tree* yang rentan terhadap *overfitting* [10].

Kebaruan (*novelty*) dari penelitian ini terletak pada integrasi metode *SMOTE* dengan algoritma *XGBoost* yang diterapkan secara spesifik pada kasus penyakit campak. Berbeda dengan penelitian sebelumnya yang mayoritas menggunakan metode standar, penelitian ini menggabungkan teknik penanganan ketidakseimbangan data sintetis yang mutakhir dengan algoritma boosting yang teroptimasi. Melalui pendekatan ini, diharapkan nilai *Recall* dan *F1-Score* dapat meningkat secara signifikan tanpa mengorbankan akurasi keseluruhan. Dengan demikian, penelitian ini bertujuan untuk menghasilkan model klasifikasi yang lebih handal dengan judul "Penanganan Ketidakseimbangan Data pada Klasifikasi Penyakit Campak Menggunakan Kombinasi *SMOTE* dan *XGBoost*".

2. METODE PENELITIAN

Gambar 1. Alur Penelitian

Berdasarkan Gambar 1 alur ini menggambarkan tahapan sistematis dari mulai pengumpulan data hingga evaluasi model. Penjelasan alur penelitian sebagai berikut [11][12]:

1. ***Data Acquisition*** : mengumpulkan data klinis campak (gejala dan label).
2. ***Preprocessing*** [13]: melakukan pembersihan data, penanganan *missing value*, dan *encoding* fitur kategorikal.
3. **Data *Splitting*** : memisahkan dataset ke dalam dua bagian, yaitu data latih yang digunakan untuk membangun model dan data uji yang digunakan untuk mengevaluasi kemampuan model terhadap data baru.[14].
4. ***Handling Imbalance* (*SMOTE*)** : menerapkan SMOTE hanya pada *Training Set* untuk menyeimbangkan jumlah kelas positif dan negatif [6].
5. ***Modeling* (*XGBoost*)** : melatih algoritma XGBoost pada data yang telah diseimbangkan [5].
6. ***Performance Evaluation*** : mengukur hasil prediksi menggunakan *Confusion Matrix.*

SMOTE

Metode penyeimbangan data yang dilakukan dengan membentuk sampel buatan pada kelas minoritas. [10], [15]. Algoritma ini mencari tetangga terdekat (K-*Nearest Neighbors*) dan menyisipkan titik data baru di antara mereka [16]. Rumus *SMOTE* ditunjukan pada persamaan (1).

$$x_{new} = x_i + gap \times (x_{i,knn} - x_i) \quad (1)$$

dimana x_{new} adalah samel sintetis, x_i sampel minoritas asli, dan gap adalah angka acak antara 0 dan 1.

XGBoost

Merupakan metode pengembangan dari algoritma *gradient boosting decision tree* yang dioptimalkan untuk menghasilkan proses komputasi yang cepat dan akurat. Algoritma ini memanfaatkan pemrosesan paralel serta mekanisme regularisasi untuk meningkatkan kinerja model sekaligus mengurangi risiko *overfitting*. [16]. Fungsi Objektif *XGBoost* ditunjukan pada persamaan (2).

$$Obj(\theta) = \sum_i L(y_i, y_i) + \sum_k (f_k)\, x_{new} \quad (2)$$

dimana L adalah *loss function.*

Pada penelitian ini, algoritma *XGBoost* digunakan untuk membangun model klasifikasi biner penyakit campak. Parameter model ditentukan untuk menjaga keseimbangan antara kemampuan model dalam mempelajari pola data dan risiko *overfitting*. Parameter yang digunakan pada model *XGBoost* tanpa *SMOTE* dan *XGBoost* dengan *SMOTE* dibuat sama agar perbandingan kinerja kedua model lebih objektif. Perbedaan utama kedua skenario hanya terletak pada kondisi data latih, yaitu data latih asli yang tidak seimbang dan data latih yang telah diseimbangkan menggunakan *SMOTE*. Parameter *XGBoost* ditunjukkan pada Tabel 1.

Tabel 1. Parameter *XGBoost*

Parameter	Nilai	Keterangan
objective	*binary*:*logistic*	Digunakan untuk klasifikasi biner status campak
eval_metric	*logloss*	Metrik evaluasi selama proses pelatihan model
n_estimators	100	Jumlah pohon keputusan yang dibangun dalam model
learning_rate	0,1	Mengatur kecepatan pembelajaran model
max_depth	3	Kedalaman maksimum setiap pohon
subsample	0,8	Proporsi data yang digunakan pada setiap iterasi
colsample_bytree	0,8	Proporsi fitur yang digunakan dalam pembentukan pohon
gamma	0	Nilai minimum loss reduction untuk melakukan pemisahan node
min_child_weight	1	Bobot minimum yang dibutuhkan pada child node
random_state	42	Menjaga konsistensi hasil eksperimen
scale_pos_weight	1	Digunakan karena penyeimbangan kelas dilakukan dengan *SMOTE*
n_jobs	-1	Menggunakan seluruh prosesor yang tersedia

Evaluasi Performa *(Confusion Matrix)*

Confusion Matrix adalah tabel yang digunakan untuk mendeskripsikan performa model klasifikasi pada sekumpulan data uji yang nilai sebenarnya sudah diketahui [17]. Tabel *Confusion Matrix* ditunjukkan pada Tabel 2.

Tabel 2. *Confusion Matrix*

	Terprediksi Negatif	**Terprediksi Positif**
Aktual Negatif	*True Negative (TN)*	*False Positive (FP)*
Aktual Positif	*False Negative (FN)*	*True Positive (TP)*

Akurasi

Menunjukkan tingkat ketepatan model dalam menghasilkan prediksi yang sesuai dibandingkan dengan keseluruhan data yang diuji. Rumus akurasi ditunjukan pada persamaan (3) [18].

Rumus :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

F1-Score

Merupakan ukuran evaluasi yang menggabungkan nilai precision dan recall dalam satu indikator melalui pendekatan rata-rata harmonik. Rumus *F1-Score* ditunjukan pada persamaan (4), *Precision* ditunjukan pada persamaan (5) dan *Recall* ditunjukan pada persamaan (5) [19], [20].

Rumus :

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

Sumber Dataset

Dataset yang digunakan dalam penelitian ini bersumber dari repositori data terbuka Kaggle dengan nama dataset Measles Vaccination and Infection. Dataset awal kemudian disesuaikan dengan kebutuhan penelitian klasifikasi penyakit campak berbasis gejala klinis. Jumlah data yang digunakan sebanyak 1.000 record, terdiri atas data pasien dengan status negatif dan positif campak. Status campak digunakan sebagai label klasifikasi biner, yaitu 0 untuk negatif campak dan 1 untuk positif campak. Berdasarkan distribusi kelas, dataset memiliki karakteristik tidak seimbang, dengan jumlah kelas negatif lebih dominan dibandingkan kelas positif. Dari total 1.000 data, sebanyak 900 data termasuk kelas negatif dan 100 data termasuk kelas positif. Kondisi ini menunjukkan rasio ketidakseimbangan sekitar 90% kelas negatif dan 10% kelas positif. Detail sumber dataset sebagai berikut :

1. Jumlah data: 1.000 *record*.
2. Variabel dan keterangan data dapat dilihat di Tabel 3.

Tabel 3. *Dataset* awal

No	**Variabel/Fitur**	**Keterangan**
1	Usia	Usia pasien/responden
2	Riwayat imunisasi	Status pernah/tidak pernah imunisasi campak
3	Demam	Gejala demam pada pasien
4	Batuk	Gejala batuk

No	Variabel/Fitur	Keterangan
5	Pilek	Gejala pilek
6	Konjungtivitis	Gejala radang mata/kemerahan pada mata
7	Ruam kulit	Munculnya rash/ruam makulopapular
8	Status campak	Target klasifikasi: 0 = negatif, 1 = positif

3. Label target data dapat dilihat di Tabel 4.

Tabel 4. Label Target

Label	Kategori	Makna
0	Negatif	Pasien tidak diklasifikasikan sebagai campak
1	Positif	Pasien diklasifikasikan sebagai campak

Proses *Preprocessing* dan Modifikasi Data

Tahap *preprocessing* dilakukan sebelum proses pelatihan model untuk memastikan data siap digunakan dalam algoritma *machine learning*. Proses *preprocessing* meliputi pemeriksaan data kosong, pemeriksaan data duplikat, penyesuaian format variabel, serta pengkodean data kategorikal ke dalam bentuk numerik. Variabel gejala seperti demam, batuk, pilek, konjungtivitis, dan ruam kulit dikodekan dalam bentuk biner, yaitu 0 untuk tidak ada gejala dan 1 untuk ada gejala. Riwayat imunisasi juga disesuaikan menjadi bentuk numerik agar dapat diproses oleh model. Modifikasi data dilakukan dengan memilih variabel yang relevan dengan gejala klinis campak dan menyesuaikan struktur data agar sesuai dengan kebutuhan klasifikasi. Dataset setelah proses *preprosessing* ditunjukkan pada Tabel 5.

Tabel 5. Dataset Setelah Proses *Preprosessing*

No	Usia	Imunisasi	Demam	Batuk	Pilek	Konjungtivitis	Ruam Kulit	Status Campak
1	5	0	1	1	1	1	1	1
2	8	1	1	0	1	0	0	0
3	3	0	1	1	1	1	1	1
4	12	1	0	1	0	0	0	0
5	6	1	1	1	1	0	0	0
6	4	0	1	1	1	1	1	1
7	10	1	0	0	1	0	0	0
8	7	1	1	1	0	0	0	0
9	2	0	1	1	1	1	1	1
10	9	1	0	1	1	0	0	0
dst								

3. HASIL DAN PEMBAHASAN

Dataset kemudian dibagi menjadi data latih dan data uji menggunakan rasio 80:20. Data latih terdiri dari 800 *record*, sedangkan data uji terdiri dari 200 *record*. Pembagian dilakukan secara stratified agar proporsi kelas negatif dan positif tetap terjaga pada data latih dan data uji. Setelah pembagian data, teknik *SMOTE* hanya diterapkan pada data latih untuk menyeimbangkan jumlah kelas minoritas tanpa memengaruhi distribusi data uji. Dengan asumsi pembagian dilakukan secara *stratified*, distribusi data sebelum *SMOTE* dapat dirinci pada Tabel 6.

Tabel 6. Pembagian Dataset

Jenis Data	Negatif	Positif	Total
Data latih	720	80	800
Data Uji	180	20	200
Total	900	100	1.000

SMOTE hanya diterapkan pada data latih, bukan pada data uji. Hal ini penting untuk mencegah kebocoran data atau data leakage. Setelah *SMOTE*, data latih menjadi lebih seimbang karena jumlah kelas positif ditingkatkan secara sintetis hingga mendekati jumlah kelas negatif. Setelah menggunakan *SMOTE* dapat dilihat pada Tabel 7.

Tabel 7. Proses Dataset

Kondisi Data Latih	Negatif	Positif	Total
Sebelum SMOTE	720	80	800
Setelah SMOTE	720	720	1.440

Dengan demikian, *SMOTE* tidak menambah data uji, tetapi hanya memperbaiki distribusi kelas pada data latih agar model *XGBoost* dapat mempelajari pola kelas positif secara lebih optimal.

Confusion Matrix XGBoost Tanpa *SMOTE*

Model XGBoost dilatih menggunakan data yang masih tidak seimbang. Dari 200 data uji, terdapat 180 data negatif dan 20 data positif. Hasil pengujian model ditunjukkan pada *confusion matrix* yang ditunjukkan pada Tabel 8.

Tabel 8. *Confusion Matrix XGBoost* Tanpa *SMOTE*

Aktual \ Prediksi	Prediksi Negatif	Prediksi Positif	Total
Aktual Negatif	178	2	180
Aktual Positif	10	10	20
Total Prediksi	188	12	200

Berdasarkan Tabel 6 dimana TN adalah 178 artinya pasien negatif yang benar diprediksi negatif, FP adalah 2 artinya pasien negatif yang salah diprediksi positif, FN adalah 10 artinya pasien positif yang salah diprediksi negatif, dan TP dalah 10 artinya pasien positif yang benar diprediksi positif.

Perhitungan manual *XGBoost* Tanpa *SMOTE*

$$Accuracy = \frac{10 + 178}{10 + 178 + 2 + 10} = \frac{188}{200} = 0{,}94 = 94\%$$

$$F1 - Score = 2 \times \frac{0{,}8333 \times 0{,}50}{0{,}8333 + 0{,}50} = 0{,}625 = 62{,}5\%$$

$$Precision = \frac{10}{10 + 2} = 0{,}833 = 83\%$$

$$Recall = \frac{19}{20} = 0{,}95 = 95\%$$

Berdasarkan hasil tersebut, kombinasi *SMOTE* dan *XGBoost* menghasilkan akurasi sebesar 97%, *precision* sebesar 79,2%, *recall* sebesar 95%, dan *F1-score* sebesar 86,4%.

Perbandingan Kinerja Model

Tabel 9. Perbandingan Kinerja Model

Metode	*Accuracy*	*Precision*	*Recall*	*F1-score*
XGBoost tanpa SMOTE	94,0%	83,3%	50,0%	62,5%
XGBoost + SMOTE	97,0%	79,2%	95,0%	86,4%

Berdasarkan Tabel 9 perbandingan tersebut menunjukkan bahwa penggunaan SMOTE memberikan peningkatan signifikan pada kemampuan model dalam mendeteksi kasus positif campak. Nilai *recall* meningkat dari 50,0% menjadi 95,0%, sedangkan *F1-score* meningkat dari 62,5% menjadi 86,4%. Peningkatan ini menunjukkan bahwa model tidak hanya memiliki akurasi yang lebih baik, tetapi juga menjadi lebih seimbang dalam mengenali kelas minoritas.

Analisis Hasil

Hasil pengujian menunjukkan bahwa *XGBoost* tanpa *SMOTE* memperoleh akurasi tinggi sebesar 94%, namun nilai tersebut belum sepenuhnya menggambarkan kemampuan model dalam mendeteksi kasus positif campak. Hal ini disebabkan oleh dominasi kelas negatif pada data uji, sehingga model cenderung lebih mudah mengenali kelas mayoritas. Kelemahan utama terlihat dari nilai *false negative* sebanyak 10 kasus, yang berarti separuh dari 20 pasien positif campak salah diklasifikasikan sebagai negatif. Setelah diterapkan *SMOTE*, jumlah *false negative* menurun dari 10 kasus menjadi 1 kasus, sehingga *recall* meningkat dari 50% menjadi 95%. Hal ini menunjukkan bahwa kombinasi *SMOTE* dan *XGBoost* lebih sensitif dalam mendeteksi kasus positif campak. Meskipun *precision* sedikit menurun dari 83,3% menjadi 79,2% akibat peningkatan *false positive*, kondisi ini masih lebih dapat diterima dalam konteks penyakit menular karena pasien yang dicurigai positif dapat diarahkan untuk pemeriksaan lanjutan.

4. KESIMPULAN

Berdasarkan hasil pengujian, dapat disimpulkan bahwa data penyakit campak yang digunakan dalam penelitian ini memiliki karakteristik tidak seimbang, dengan dominasi kelas negatif dibandingkan kelas positif. Kondisi tersebut menyebabkan model XGBoost tanpa SMOTE memiliki akurasi tinggi, tetapi recall rendah, sehingga kurang optimal dalam mendeteksi kasus positif campak.

Penerapan SMOTE pada data latih terbukti mampu meningkatkan performa model XGBoost, terutama pada metrik *recall* dan *F1-score*. *Recall* meningkat dari 50,0% menjadi 95,0%, sedangkan F1-score meningkat dari 62,5% menjadi 86,4%. Meskipun precision sedikit menurun, peningkatan recall lebih bernilai dalam konteks deteksi penyakit menular karena dapat mengurangi risiko pasien positif tidak teridentifikasi.

Dengan demikian, kombinasi SMOTE dan XGBoost merupakan pendekatan yang lebih efektif dibandingkan XGBoost tanpa SMOTE dalam menangani ketidakseimbangan data pada klasifikasi penyakit campak. Pendekatan ini dapat menjadi dasar pengembangan model prediksi berbasis machine learning untuk mendukung deteksi dini dan penguatan sistem surveilans penyakit menular. Penelitian selanjutnya disarankan menggunakan data klinis campak yang lebih besar, valid, dan berasal dari sumber resmi fasilitas kesehatan agar model lebih representatif. Selain itu, perlu dilakukan perbandingan dengan algoritma lain serta validasi lapangan untuk memastikan model dapat digunakan sebagai pendukung deteksi dini penyakit campak.

UCAPAN TERIMA KASIH

Terima kasih atas dukungan dana penelitian yang di berikan oleh Politeknik Kaltara melalui LPPM.

DAFTAR PUSTAKA

[1] I. W. Adhi, S. Gemilang, and I. W. Supriana, “Case-Based Reasoning untuk Diagnosis Penyakit Campak Menggunakan Metode Bayesian Model,” vol. 2, pp. 801–806, 2024.

[2] L. Chaves and G. Marques, “applied sciences Data Mining Techniques for Early Diagnosis of Diabetes,” *Appl. Sci.*, vol. 11, no. 2218, pp. 1–12, 2021.

[3] K. H. Hanif, N. R. Muntiari, D. Harto, and D. S. Wiranata, “Perbandingan Analisis Sentimen Komentar Mahasiswa Prodi Teknik Komputer Menggunakan Algoritma Decision Tree , Support Vector Machine (SVM), dan Random Forest,” vol. 12, no. 01, 2026.

[4] W. Apriliah, I. Kurniawan, M. Baydhowi, and T. Haryati, “Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest,” *Sistemasi*, vol. 10, no. 1, p. 163, 2021, doi: 10.32520/stmsi.v10i1.1129.

[5] W. Chang, Y. Liu, Y. Xiao, X. Yuan, X. Xu, and S. Zhang, “Phương Pháp Xư Lý Dư Liệu Luận Án.Pdf,” 2019.

[6] M. Rezapour, “Sentiment classification of skewed shoppers’ reviews using machine learning techniques, examining the textual features,” *Eng. Reports*, vol. 3, no. 1, pp. 1–13, 2021, doi: 10.1002/eng2.12280.

[7] B. Algama *et al.*, "Analisa Perbandingan Metode Arithmetic Mean Filtering Dan Metode Konvolusi Pada Citra," vol. 5, no. 02, 2023.
[8] A. W. Putera *et al.*, "Klasifikasi Sms Spam Menggunakan Algoritma K-Nearest Neighbor," vol. 5, no. 01, 2023.
[9] T. M. Prasetyo, A. Amrullah, S. Syahrir, and B. N. Sari, "Implementasi Algoritma Svm (Support Vector Machine) Dalam Klasifikasi Penyakit Paru-Paru Berdasarkan," vol. 6, no. 1, 2022.
[10] D. Andriyani, A. Faqih, and S. E. Permana, "The Effect of SMOTE Application on Support Vector Machine Performance in Sentiment Classification on Imbalanced Datasets," vol. 4, no. 2, 2025.
[11] N. R. Muntiari, K. Nisa, A. S. Sandi A, I. A. Ashari, K. H. Hanif, and R. W. Dwinanto, "Comparison of random forest algorithm, support vector machine, and k-nearest neighbor for diabetes disease classification," *AIP Conf. Proc.*, vol. 2706, pp. 1–8, 2023, doi: 10.1063/5.0120218.
[12] K. H. Hanif, A. Fadllullah, N. R. Muntiari, and I. A. Fahrezi, "A Comparative Sentiment Analysis of Computer Engineering Student Feedback Using Decision Trees and SVM," vol. 10, no. 1, pp. 71–82, 2025, doi: 10.31572/inotera.Vol10.Iss1.2025.ID436.
[13] N. R. Muntiari and K. H. Hanif, "Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning," *J. Ilmu Komput. dan Teknol.*, vol. 3, no. 1, pp. 1–6, 2022, doi: 10.35960/ikomti.v3i1.766.
[14] N. R. Muntiari, K. Nisa, A. S. Sandi, I. A. Ashari, A. Kharis Hudaiby Hanif, and R. W. Dwinanto, "Comparison of random forest algorithm, support vector machine, and k-nearest neighbor for diabetes disease classification," no. May, 2023.
[15] R. Muntiar, Novita Ranti, Kharis Hudaiby Hanif, Syamsiah, "Klasifikasi Penyakit Preekslamsia Pada Ibu Hamil Menggunakan Perbandingan Algoritma Machine Learning," vol. 13, no. 2, pp. 96–102, 2025.
[16] A. Zaelani, M. Fikriansyah, M. Syahdan, and I. A. Hakim, "Peran Keamanan Basis Data Relasional dalam Menjamin Kualitas Data untuk Proses Data Mining : Studi Kasus Klasifikasi Aktivitas Akses Berisiko," vol. 4, no. 2, pp. 5286–5292, 2025.
[17] N. R. Muntiari, K. H. Hanif, and W. Rahmaniar, "Application of the Certainty Factor Method for Diagnosing Osteoarthritis Using the Python Programming Language," *J. Adv. Heal. Informatics Res.*, vol. 1, no. 1, pp. 21–27, 2023.
[18] N. R. Muntiari, I. C. Nisa, A. Sriekaningih, A. Y. Adyatma, and M. Yusril, "Penerapan Algoritma YOLOv8 Dalam Indentifikasi Wajah secara Real- Time menggunakan CCTV untuk Presensi Siswa," vol. 4, no. 3, pp. 1155–1165, 2024.
[19] T. Azhima, Y. Siswa, T. Informatika, F. Sains, U. Muhammadiyah, and K. Timur, "Komparasi Optimasi Chi-Square , CFS , Information Gain Dan ANOVA Dalam Evaluasi Peningkatan Akurasi Algoritma Klasifikasi Data Performa Akademik Mahasiswa," vol. 18, no. 1, 2023.
[20] R. M. 2 Muntiari, Novita Ranti, "A Bibliometric Analysis of Knowledge Distillation in Medical Image Segmentation," vol. 2, no. 3, pp. 115–126, 2024, doi: 10.59247/jahir.v2i3.297.