



Evaluasi Kinerja Machine Learning pada Klasifikasi Penyakit Jantung Menggunakan Teknik Penyeimbangan Data

Eni Rohaini^{1*}, Gunardi Gunardi², Nurhayati Nurhayati³, Jasmir Jasmir⁴, Zahra Prisdian Tiararosa⁵

^{1,5} Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Dinamika Bangsa, Jambi, Indonesia

^{1,2,3} Program Studi Sistem Komputer, Fakultas Ilmu Komputer, Universitas Dinamika Bangsa, Jambi, Indonesia

Email: enirohaini0104@gmail.com^{1*}, gunardi@unama.ac.id², nurhayatihahim75@gmail.com³, jasmir@unama.ac.id⁴, rosemichhhh888@gmail.com⁵

Alamat: Universitas Dinamika Bangsa, Jl. Jendral Sudirman, Tehok Jambi Selatan, Jambi, Indonesia

*Penulis Korespondensi: enirohaini0104@gmail.com

Abstract. *Imbalanced data remains a significant issue in heart disease classification using machine learning, as it tends to cause models to overestimate the majority class while ignoring minority classes with high clinical value. This can lead to a decrease in accuracy and the model's ability to accurately detect disease cases. Therefore, this study aims to assess the effectiveness of oversampling techniques, namely Random Oversampling and Synthetic Minority Oversampling Technique (SMOTE), in improving the performance of the K-Nearest Neighbors (KNN), Naive Bayes (NB), and Random Forest (RF) algorithms. The dataset used comes from Kaggle and consists of 918 data sets with 12 attributes representing patient information related to heart disease prediction. The research stages include data preprocessing, baseline model testing, and re-evaluation using the two oversampling methods. Experimental results show that oversampling can improve the performance of all algorithms. KNN achieved the best results with SMOTE, with an accuracy of 72.98% and an F1-score of 75.39%. In the Naive Bayes algorithm, both oversampling techniques produced relatively stable performance, with the highest F1-score of 73.56% using SMOTE. Meanwhile, Random Forest showed the most optimal performance when combined with Random Oversampling, with an accuracy of 79.19% and an F1-score of 81.51%. These findings confirm that the success of data balancing techniques is strongly influenced by the characteristics of the classification algorithm used, and provide a practical contribution in determining strategies for handling imbalanced data in health research.*

Keywords: *Heart Disease; Machine Learning; Oversampling; Random Oversampling; SMOTE*

Abstrak. Permasalahan data tidak seimbang masih menjadi isu penting dalam klasifikasi penyakit jantung menggunakan *machine learning*, karena cenderung menyebabkan model lebih dominan memprediksi kelas mayoritas dan mengabaikan kelas minoritas yang memiliki nilai klinis tinggi. Kondisi tersebut dapat berdampak pada penurunan akurasi serta kemampuan model dalam mendeteksi kasus penyakit secara tepat. Oleh karena itu, penelitian ini bertujuan untuk mengkaji efektivitas teknik *oversampling*, yaitu *Random Oversampling* dan *Synthetic Minority Over-sampling Technique (SMOTE)*, dalam meningkatkan kinerja algoritma *K-Nearest Neighbors (KNN)*, *Naive Bayes (NB)*, dan *Random Forest (RF)*. Dataset yang digunakan berasal dari Kaggle dan terdiri atas 918 data dengan 12 atribut yang merepresentasikan informasi pasien terkait prediksi penyakit jantung. Tahapan penelitian meliputi prapemrosesan data, pengujian model *baseline*, serta evaluasi ulang dengan penerapan dua metode *oversampling*. Hasil eksperimen menunjukkan bahwa penerapan *oversampling* mampu meningkatkan performa seluruh algoritma. KNN mencapai hasil terbaik dengan *SMOTE*, dengan akurasi 72,98% dan *F1-score* 75,39%. Pada algoritma *Naive Bayes*, kedua teknik *oversampling* menghasilkan performa yang relatif stabil, dengan nilai *F1-score* tertinggi sebesar 73,56% menggunakan *SMOTE*. Sementara itu, *Random Forest* menunjukkan kinerja paling optimal ketika dikombinasikan dengan *Random Oversampling*, dengan akurasi 79,19% dan *F1-score* 81,51%. Temuan ini menegaskan bahwa keberhasilan teknik penyeimbangan data sangat dipengaruhi oleh karakteristik algoritma klasifikasi yang digunakan, serta memberikan kontribusi praktis dalam menentukan strategi penanganan data tidak seimbang pada penelitian di bidang kesehatan.

Kata kunci: *Machine Learning; Oversampling; Penyakit Jantung; Random Oversampling; SMOTE*

1. LATAR BELAKANG

Kemajuan teknologi yang terus meningkat telah memberikan dampak besar terhadap berbagai sektor, termasuk sektor kesehatan. Perkembangan ini mendorong lahirnya bidang *Biomedical Engineering*, yaitu cabang ilmu yang mengintegrasikan teknologi informasi, ilmu komputer, dan kedokteran guna meningkatkan efektivitas serta kualitas layanan kesehatan (Zhu et al., 2022). Melalui pendekatan tersebut, berbagai sistem berbasis kecerdasan buatan dapat dikembangkan untuk mendukung proses diagnosis, prediksi penyakit, dan pengambilan keputusan medis secara lebih cepat dan akurat. Salah satu permasalahan kesehatan yang mendapat perhatian serius adalah penyakit jantung. Jantung memiliki peran vital dalam menjaga kelangsungan hidup manusia dengan memompa darah ke seluruh bagian tubuh. Apabila fungsi jantung mengalami gangguan, maka suplai oksigen dan nutrisi ke organ-organ tubuh akan terganggu, sehingga berpotensi menimbulkan berbagai komplikasi serius (Sugiyarto et al., 2021). Namun, tingkat kesadaran masyarakat terhadap pentingnya pemeriksaan kesehatan jantung masih tergolong rendah, sehingga banyak kasus penyakit jantung tidak terdeteksi pada tahap awal. Padahal, penyakit ini dapat menyerang individu dari berbagai kelompok usia, termasuk mereka yang tidak memiliki riwayat penyakit sebelumnya. Oleh sebab itu, upaya deteksi dini penyakit jantung menjadi sangat penting untuk mencegah terjadinya komplikasi lebih lanjut serta menurunkan risiko kematian akibat penyakit jantung.

Meningkatnya kapasitas komputasi serta tersedianya data kesehatan dalam bentuk digital telah membuka peluang luas dalam penerapan teknologi informasi untuk bidang medis. Salah satu pendekatan yang banyak dimanfaatkan adalah *machine learning* (Rapacz et al., 2021), yang mampu membantu proses deteksi dan klasifikasi penyakit jantung secara lebih efektif. Melalui kemampuan belajar dari data, *machine learning* dapat mengidentifikasi pola-pola tersembunyi yang berguna dalam memprediksi risiko penyakit secara otomatis (Nur Riza Pahlevi & Badriyah, 2025). Berbagai *algoritma machine learning* telah diterapkan dalam penelitian kesehatan untuk meningkatkan ketepatan diagnosis serta efisiensi dalam pengolahan dan analisis data pasien.

Meskipun demikian, penerapan *machine learning* pada data medis masih menghadapi sejumlah tantangan, salah satunya adalah permasalahan ketidakseimbangan kelas (*imbalanced dataset*). Kondisi ini terjadi ketika jumlah data pasien tanpa penyakit jauh lebih dominan dibandingkan data pasien yang mengalami penyakit tertentu, termasuk penyakit jantung. Ketidakseimbangan distribusi kelas tersebut dapat menyebabkan model klasifikasi cenderung

berpihak pada kelas mayoritas, sehingga kemampuan dalam mengenali kelas minoritas menjadi menurun (Hasib et al., 2020). Dampaknya, model berpotensi menghasilkan prediksi yang kurang akurat, khususnya dalam mengidentifikasi pasien yang memiliki risiko penyakit jantung secara nyata.

Berbagai metode telah dikembangkan untuk mengatasi permasalahan ketidakseimbangan kelas pada data klasifikasi, yang secara umum dapat dikelompokkan ke dalam dua kategori, yaitu pendekatan berbasis data (*data-level approach*) dan pendekatan berbasis algoritma (*algorithm-level approach*) (Malek et al., 2023). Pendekatan berbasis data dilakukan dengan menyesuaikan distribusi data melalui proses *resampling*, baik dengan cara mengurangi dominasi kelas mayoritas (*undersampling*) maupun dengan menambah representasi kelas minoritas (*oversampling*). Dalam penelitian ini, penanganan ketidakseimbangan data difokuskan pada pendekatan tingkat data dengan menerapkan dua teknik *oversampling*, yaitu *Random Oversampling* dan *Synthetic Minority Over-sampling Technique (SMOTE)*, sebagai upaya untuk meningkatkan kinerja model klasifikasi.

Random Oversampling bekerja dengan menambah jumlah data kelas minoritas melalui duplikasi sampel yang sudah ada (Xin & Rashid, 2021). Pendekatan ini relatif mudah diterapkan dan terbukti mampu meningkatkan sensitivitas model terhadap kelas minoritas (Sampath et al., 2024). Di sisi lain, *Synthetic Minority Over-sampling Technique (SMOTE)* yang diperkenalkan oleh Nitesh V. Chawla (Liang et al., 2020) menghasilkan data sintesis baru dengan cara melakukan interpolasi antara sampel-sampel minoritas yang saling berdekatan. Teknik ini dianggap lebih unggul karena mampu mengurangi potensi terjadinya *overfitting* yang sering muncul akibat penggandaan data secara langsung (Sampath et al., 2024) (Soltanzadeh & Hashemzadeh, 2021).

Untuk mengevaluasi efektivitas kedua teknik *oversampling* tersebut, penelitian ini menerapkan tiga algoritma machine learning yang umum digunakan dalam klasifikasi data medis, yaitu *K-Nearest Neighbors (KNN)*, *Naive Bayes*, dan *Random Forest*. KNN merupakan algoritma berbasis jarak yang sederhana namun cukup efektif, khususnya pada dataset berukuran kecil hingga menengah, serta memiliki sensitivitas tinggi terhadap distribusi data (Halder et al., 2024). *Naive Bayes* adalah algoritma berbasis probabilistik dengan asumsi independensi antar fitur, yang dikenal efisien dalam menangani data berukuran besar dan relatif tahan terhadap noise (Ige et al., 2024). Sementara itu, *Random Forest* merupakan metode *ensemble learning* berbasis pohon keputusan yang mampu mengolah data dalam jumlah besar dan kompleks, serta efektif dalam mengurangi risiko *overfitting* (Mohammadagha, n.d.).

Melalui penggabungan teknik *Random Oversampling* dan *SMOTE* dengan algoritma machine learning tersebut, penelitian ini bertujuan untuk membangun model klasifikasi yang lebih seimbang dan memiliki tingkat akurasi yang lebih baik dalam mendeteksi risiko penyakit jantung. Diharapkan hasil penelitian ini dapat memberikan kontribusi dalam pengembangan sistem pendukung keputusan di bidang kesehatan, khususnya untuk membantu dokter dan tenaga medis dalam melakukan analisis awal terhadap potensi penyakit jantung berdasarkan data klinis pasien.

2. KAJIAN TEORITIS

Berbagai penelitian pun telah banyak menggunakan metode-metode *machine learning* yang sesuai dengan kapasitasnya. Samosir Amril, MS Hasibuan (Samosir et al., 2021) menggunakan beberapa metode *machine learning* seperti *Random Forest* (RF), *Naïve Bayes* (NB) dan *K-Nearest Neighbor* (KNN) untuk mendeteksi penyakit jantung dari dataset *Cleveland Clinic Foundation* sejumlah 304 *record* dengan nilai evaluasi akurasi masing masing metode sebesar RF =0,84; NB = 0,84, dan KNN =0,839. Sahar (Sahar, 2020) juga melakukan penelitian tentang penyakit jantung menggunakan KNN, dengan hasil nilai akurasi 67%, dataset yang digunakan diambil dari *UCI machine learning*. Alham dkk (Alham, 2021) menggunakan *machine learning* yang lain yaitu C4.5 untuk klasifikasi penyakit jantung dan menghasilkan akurasi yang sangat bagus yaitu 94,4% dari dataset penyakit jantung RSUD Dr. Soedarso Pontianak. Arni Sepharni (Sepharni et al., 2022), meneliti tentang klasifikasi penyakit jantung yang juga menggunakan algoritma C4.5, dan menghasilkan akurasi sebesar 79%, menggunakan *public dataset* dari *kaggle*.

Penelitian yang dilakukan oleh Khalid Amen mengenai diagnosis risiko penyakit jantung menggunakan *machine learning* dengan algoritma *Support Vector Machine* (SVM) dan *Random Forest* (LR) menunjukkan bahwa LR memiliki akurasi 82%, sedangkan SVM mencapai 80%. Data penelitian ini bersumber dari *UCI Heart Disease Dataset* yang terdiri dari 300 *record* dengan 14 atribut (Amen et al., 2020). Dalam penelitian yang dilakukan oleh Syamsul Bahri mengenai klasifikasi penyakit pada anak, algoritma C4.5 dibandingkan dengan *Naïve Bayes*. Hasilnya menunjukkan bahwa C4.5 memiliki akurasi tertinggi sebesar 90,00%, sementara *Naïve Bayes* mencapai 89,58% (Bahri et al., 2018).

Dari penelitian yang telah dilakukan oleh peneliti-peneliti terdahulu terkait klasifikasi terhadap penyakit gagal jantung telah banyak metode *data mining* yang diterapkan, diantaranya pada penelitian yang dilakukan oleh Yovi Pratama dkk. (Pratama et al., 2022) dan Dede Andri Muhammad Reza dkk. (Reza et al., 2022) kedua penelitian tersebut melakukan klasifikasi penyakit gagal jantung dengan menggunakan algoritma *K-NN*. Adapun penelitian dengan menerapkan algoritma *C4.5* sebagai metode klasifikasi yang dilakukan oleh Elin Nurlia dkk. (Elin Nurlia, 2021) yang menerapkan metode *C4.5* berbasis seleksi fitur *Forward Selection* didapatkan hasil sebesar 77,89% dengan nilai *AUC* 0,750 yang tergolong dalam kategori *fair classification*. Adapun penelitian yang melanjutkan penelitian tersebut dengan menerapkan keterbaruan menggunakan metode yang sama tetapi dengan basis yang berbeda yaitu *C4.5* dan *PSO* yang dilakukan oleh Agustiena Merdekawati (Merdekawati, 2022) didapat hasil sebesar 73,33% sedangkan menggunakan *PSO* sebesar 99,00% penerapan *PSO* mampu menyeleksi atribut pada *C4.5*.

Pada penelitian yang dilakukan Shakeela et. al (Shakeela et al., 2021) bertujuan untuk mengidentifikasi fitur-fitur esensial dalam dataset. Studi ini mengevaluasi metode yang dikembangkan untuk mendeteksi serangan jahat (*malicious attacks*) dalam jaringan komunikasi data. Keberhasilan desain penelitian ini bergantung pada ketersediaan fitur-fitur yang relevan dengan jenis serangan dalam skenario pemantauan waktu nyata.

Beberapa penelitian sebelumnya telah mengkaji efektivitas teknik oversampling dalam menangani data yang tidak seimbang. Nguyen et al. (Ngo et al., 2022) mengevaluasi penerapan *SMOTE* dan *Random Oversampling* pada algoritma K-Nearest Neighbors (KNN) untuk prediksi penyakit jantung, yang menghasilkan peningkatan signifikan pada *F1 Score* dan *Balanced Accuracy*. Zhao et al (Pan et al., 2020) mengimplementasikan kombinasi *SMOTE* dan Tomek Links pada model *Naive Bayes* dalam klasifikasi data biomedis, yang berhasil meningkatkan sensitivitas model terhadap kelas minoritas. Wang dan Li (X. Wang et al., 2021) mengusulkan integrasi *SMOTE* dengan *Random Forest* pada dataset kesehatan, yang secara signifikan meningkatkan *Balanced Accuracy* dan *F1 Score* dibandingkan model tanpa *resampling*

Selanjutnya Tingting Pan (Pan et al., 2020) dkk membahas metode pengambilan sampel yang efektif yang meningkatkan distribusi data. Satu metode yang diseimbangkan kembali, *Adaptive-SMOTE*, meningkatkan metode *SMOTE* dengan secara adaptif memilih kelompok data Inner dan Danger dari kelas minoritas sehingga kelas minoritas baru dikompilasi berdasarkan data yang dipilih, sehingga mencegah perluasan batas kategori dan memperkuat

karakteristik distribusi data asli. Dengan hasil penelitian menyatakan bahwa *Adaptive-SMOTE* mempunyai nilai F – measure dan *Accuracy* yang lebih tinggi dibandingkan metode sampling lain yang sudah ada dan ketahanan yang lebih tinggi terhadap classifier dan dataset dengan nilai Imb yang berbeda. *Gaussian Oversampling* lebih efisien ketika berhadapan dengan klasifikasi yang sangat tidak seimbang.

Kemudian Hai Ly NGO (Ngo et al., 2022) membahas tentang Komposisi citra deret waktu dan penggunaan teknik *SMOTE* untuk menyeimbangkan kumpulan data dalam pemetaan penggunaan penutupan lahan. Hasil penelitian menyediakan metode yang efektif untuk memantau perubahan tutupan lahan spasial dan temporal yang berperan sebagai kerangka kerja untuk penelitian relevan lainnya yang terkait dengan perubahan tutupan lahan, yang dapat mendukung perencanaan dan pengelolaan wilayah yang berkelanjutan.

3. METODE PENELITIAN

Untuk mencapai hasil penelitian yang maksimal, maka disusunlah serangkaian langkah penting untuk mengembangkan model ini menjadi lebih baik. Langkah-langkah yang dimaksud disajikan pada Gambar 1. Prosesnya dimulai dari pengumpulan dataset yang diambil dari situs kaggle.com, kemudian masuk ke tahap *preprocessing*. Setelah tahap *preprocessing*, kami melakukan 2 kali percobaan, percobaan pertama, proses masuk ke tahap machine learning dan tahap *classification performance evaluation*. Percobaan kedua, pada tahap machine learning dikolaborasikan dengan teknik *oversampling*, pada akhirnya adalah membandingkan hasil dari kedua percobaan tersebut.

A. Dataset

Pemilihan dataset ditentukan atas data yang akan diproses dan mencari data yang tersedia serta memperoleh data tambahan yang dibutuhkan, kemudian menintegrasikan data tersebut didalam dataset. Pada penelitian ini, dipilih dataset penyakit jantung yang dapat diakses dari public dataset dari situs Kaggle yaitu <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction> terdiri dari 918 data dan 12 atribut. (Kelas 0=410, Kelas 1= 508)

B. Preprocessing

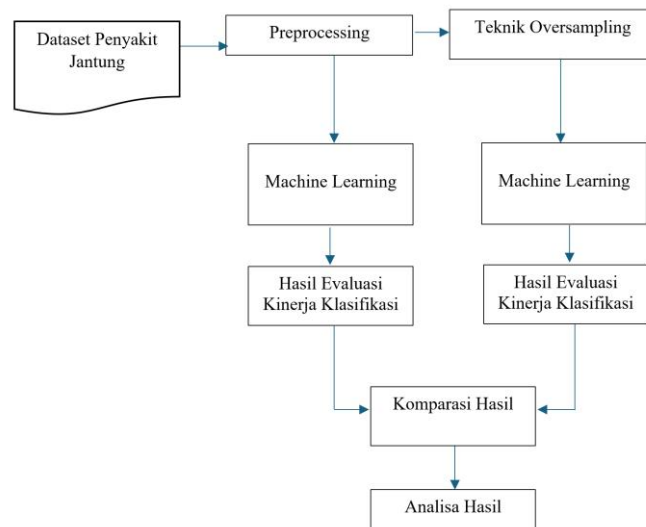
Selanjutnya, *fase pra-pemrosesan* dilakukan untuk membersihkan dan mempersiapkan data untuk klasifikasi. Ini mencakup identifikasi dan perbaikan data yang tidak konsisten dan salah.

C. *Random Oversampling*

Random Oversampling bekerja pada kelas minoritas, menyeimbangkan data dengan teknik acak, serta tidak meniadakan atau menghapus pengamatan. Akan tetapi, dengan adanya replikasi pengamatan pada data asli dapat menyebabkan *overfitting*.

D. *SMOTE*

SMOTE merupakan metode yang populer diterapkan dalam rangka menangani ketidakseimbangan kelas. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan cara sampling ulang sampel kelas minoritas. *SMOTE* bekerja dengan memilih sampel yang dekat pada *feature space*, menggambar garis di antara sampel di *feature space* dan menggambar sampel baru pada titik di sepanjang garis itu. Lebih jelasnya, contoh acak dari kelas minoritas pertama kali dipilih. Kemudian k dari tetangga terdekat untuk sampel itu ditemukan (biasanya $k = 5$). Tetangga yang dipilih secara acak dipilih dan contoh sintetis dibuat pada titik yang dipilih secara acak antara dua contoh dalam *feature space* tersebut.



Gambar 1. Kerangka Kerja Penelitian

E. **K Nearest Neighbor.**

KNN mengklasifikasikan objek berdasarkan kedekatan titik data pelatihan (Chivukula et al., 2021) (Islam et al., 2022) (Zhang et al., 2024). KNN tidak melibatkan fase pelatihan luring (Nguyen et al., 2023). Sebaliknya, KNN menyimpan semua dokumen pelatihan dan menghitung jarak selama fase prediksi (Nadeem et al., 2023). KNN menetapkan kelas berdasarkan tetangga terdekat dan kategorinya (Yang & Liu, n.d.).

F. Naïve Bayes

Naïve Bayes adalah pengklasifikasi probabilistik sederhana yang mengasumsikan independensi fitur (Berrar, 2018) (Alwan et al., 2022). NB menyeimbangkan kinerja dengan efisiensi komputasi dan berkinerja baik dengan ukuran sampel kecil karena regularisasi yang melekat (S. Wang et al., 2023) (Anderson et al., 2024). Namun, KNN mengalami kesulitan dengan interaksi antar fitur (Badar & Fisichella, 2024).

G. Random Forest

Random Forest adalah metode pembelajaran ensemble yang membangun beberapa pohon keputusan independen (Khaleel et al., 2023) (Assegie et al., 2022). Setiap pohon memberikan suara pada kelas contoh uji, dan suara mayoritas menentukan prediksi akhir (Sekulić et al., 2020). *Random Forest* mengatasi *overfitting* dengan menggabungkan beragam pohon yang dibuat melalui pemilihan fitur dan data acak (Syahputra & Wibowo, 2023).

4. HASIL DAN PEMBAHASAN

Berdasarkan hasil evaluasi kinerja model klasifikasi menggunakan algoritma *K-Nearest Neighbors* (KNN), KNN dengan teknik *SMOTE*, dan KNN dengan *Random Oversampling*, diperoleh nilai metrik klasifikasi yang bervariasi.

Tabel 1. Confusion Matrix KNN dengan dan tanpa *OverSampling*

Metrix	KNN	KNN+ <i>SMOTE</i>	KNN+Random OS
TP	313	380	350
FP	238	105	152
FN	146	143	143
TN	221	290	273

Pada Tabel 1, terlihat bahwa metode KNN baseline tanpa oversampling menghasilkan *True Positive (TP)* sebanyak 313 dan *True Negative (TN)* sebesar 221. Namun, nilai *False Positive (FP)* masih cukup tinggi (238), sementara *False Negative (FN)* sebesar 146. Hal ini menunjukkan bahwa KNN *baseline* masih kesulitan dalam membedakan kelas secara seimbang, khususnya dalam mengurangi kesalahan prediksi pada kelas negatif.

Ketika dilakukan teknik *SMOTE* (*Synthetic Minority Over-sampling Technique*), terjadi peningkatan signifikan pada jumlah *TP* (380) dan *TN* (290), sekaligus penurunan drastis pada *FP* (105). Sementara itu, *FN* hanya sedikit berkurang (143). Kondisi ini menunjukkan bahwa *SMOTE* mampu memperbaiki distribusi data latih sehingga model KNN lebih optimal dalam mengklasifikasikan kedua kelas.

Sementara itu, teknik *Random Oversampling* (*Random OS*) juga meningkatkan jumlah *TP* (350) dan *TN* (273) dibandingkan baseline, meskipun tidak sebaik *SMOTE*. Nilai *FP* pada *Random OS* (152) lebih besar dibanding *SMOTE* (105), yang menunjukkan bahwa metode ini masih menghasilkan lebih banyak kesalahan dalam memprediksi kelas negatif.

Hal tersebut semakin diperkuat dengan hasil evaluasi pada Tabel 2. Nilai akurasi baseline KNN hanya mencapai 58,17%, yang menandakan performa klasifikasi masih rendah. Setelah diterapkan *SMOTE*, akurasi meningkat signifikan hingga 72,98%, sedangkan *Random OS* juga memberikan peningkatan menjadi 67,86%.

Dari sisi *precision*, baseline KNN hanya 56,81%, artinya banyak kesalahan pada prediksi positif. Setelah *oversampling*, *SMOTE* memberikan peningkatan *precision* yang sangat besar hingga 78,35%, sedangkan *Random OS* berada di posisi tengah dengan 69,72%.

Pada *metrik recall*, baseline KNN sudah relatif cukup baik (68,19%), tetapi masih meningkat dengan *Random OS* (70,99%) dan *SMOTE* (72,65%). Ini menunjukkan bahwa kedua teknik *oversampling* membantu model lebih baik dalam mengenali kelas positif.

Dampak nyata terlihat pada *F1-score*, yang menggabungkan *precision* dan *recall*. Baseline KNN hanya mencapai 61,98%, sedangkan dengan *SMOTE* meningkat tajam menjadi 75,39%. *Random OS* juga menunjukkan perbaikan menjadi 70,35%, namun tetap lebih rendah dibanding *SMOTE*.

Secara keseluruhan, hasil ini menunjukkan bahwa *SMOTE* adalah teknik *oversampling* yang paling efektif dalam meningkatkan performa KNN pada penelitian ini. *SMOTE* tidak hanya meningkatkan akurasi, tetapi juga memberikan keseimbangan yang lebih baik antara *precision* dan *recall*, sehingga menghasilkan *F1-score* tertinggi. Sementara itu, *Random Oversampling* juga memberikan perbaikan dibanding baseline, tetapi belum mampu menandingi efektivitas *SMOTE*.

Tabel 2. Evaluasi kinerja KNN dengan dan tanpa Oversampling

Evaluasi	KNN	KNN + <i>SMOTE</i>	KNN + Random OS
Accuracy	58,17%	72,98%	67,86%
Precision	56,81%	78,35%	69,72%
Recall	68,19%	72,65%	70,99%
F1-score	61,98%	75,39%	70,35%

Berdasarkan Tabel 3, algoritma *Naive Bayes* (NB) baseline menghasilkan *True Positive* (*TP*) sebesar 339 dan *True Negative* (*TN*) sebanyak 302, dengan *False Positive* (*FP*) sebesar 143 dan *False Negative* (*FN*) sebesar 134. Distribusi ini menunjukkan bahwa NB sudah cukup

seimbang dalam mengklasifikasikan data, dengan jumlah kesalahan prediksi yang relatif lebih rendah dibandingkan baseline KNN pada tabel sebelumnya.

Tabel 3. *Confusion Matrix Naïve Bayes dengan dan tanpa Oversampling*

Metrix	NB	NB + SMOTE	NB+Random OS
TP	339	366	372
FP	143	138	144
FN	134	125	124
TN	302	289	278

Setelah dilakukan teknik *SMOTE*, jumlah *TP* meningkat menjadi 366 dan *FN* menurun menjadi 125, yang berarti model lebih mampu mengenali kelas positif. Namun, *TN* justru sedikit menurun dari 302 menjadi 289, dengan *FP* relatif stabil di angka 138. Hal ini menandakan bahwa *SMOTE* lebih mengarahkan perbaikan pada kemampuan deteksi kelas positif meskipun sedikit mengorbankan prediksi kelas negatif.

Sedangkan pada *Random Oversampling*, *TP* meningkat lebih tinggi lagi menjadi 372 dan *FN* menurun ke 124, yang menunjukkan kemampuan model dalam mendeteksi kelas positif lebih baik daripada *SMOTE*. Namun, hal ini diikuti dengan penurunan *TN* menjadi 278 dan peningkatan *FP* menjadi 144. Kondisi ini memperlihatkan bahwa *Random OS* meningkatkan *recall*, tetapi dengan konsekuensi *precision* sedikit menurun dibandingkan *SMOTE*.

Hal tersebut sejalan dengan hasil evaluasi pada Tabel 4. Akurasi *baseline* NB adalah 69,82%, kemudian meningkat menjadi 71,35% dengan *SMOTE*, dan 70,80% dengan *Random OS*. Nilai *precision* NB *baseline* berada di 70,33%, meningkat menjadi 72,61% dengan *SMOTE*, dan 72,09% dengan *Random OS*.

Dari sisi *recall*, NB *baseline* mencapai 71,67%, kemudian meningkat menjadi 74,54% dengan *SMOTE*, dan sedikit lebih tinggi lagi (75,00%) dengan *Random OS*. Peningkatan ini menunjukkan bahwa kedua metode *oversampling* efektif dalam memperkuat kemampuan model mendeteksi kelas positif.

Sementara itu, *F1-score* *baseline* NB adalah 70,99%, naik menjadi 73,56% dengan *SMOTE*, dan 73,51% dengan *Random OS*. Peningkatan nilai F1 ini menegaskan bahwa kedua teknik *oversampling* berhasil memperbaiki kinerja NB, meskipun perbedaan antara *SMOTE* dan *Random OS* relatif kecil.

Secara keseluruhan, NB dengan *SMOTE* menghasilkan performa yang paling stabil dengan keseimbangan antara *precision* dan *recall*, sementara NB dengan *Random OS* lebih unggul pada *recall* tetapi sedikit menurunkan *precision*. Kedua metode *oversampling* sama-

sama meningkatkan performa dibanding baseline, tetapi SMOTE cenderung lebih konsisten dalam menjaga keseimbangan metrik evaluasi.

Tabel 4. Evaluasi kinerja Naïve Bayes dengan dan tanpa Oversampling

Evaluasi	NB	NB + SMOTE	NB + Random OS
Accuracy	69,82%	71,35%	70,80%
Precision	70,33%	72,61%	72,09%
Recall	71,67%	74,54%	75,00%
F1-score	70,99%	73,56%	73,51%

Pada Tabel 5, model *Random Forest* (RF) *baseline* menghasilkan *True Positive* (TP) sebesar 361 dan *True Negative* (TN) sebanyak 207. Namun, jumlah *False Positive* (FP) masih cukup tinggi yaitu 202, disertai *False Negative* (FN) sebesar 148. Hal ini menunjukkan bahwa *baseline* RF masih kesulitan dalam mengurangi kesalahan prediksi, terutama pada kelas negatif.

Tabel 5. *Confusion Matrix* *Random Forest* dengan dan tanpa *Oversampling*

Metrix	RF	RF + SMOTE	RF+Random OS
TP	361	412	421
FP	202	103	94
FN	148	115	97
TN	207	288	306

Dengan penerapan *SMOTE*, jumlah *TP* meningkat menjadi 412 dan *FN* menurun menjadi 115, yang berarti model semakin mampu mengenali kelas positif. Selain itu, *TN* juga meningkat signifikan menjadi 288, serta *FP* menurun menjadi 103. Perbaikan ini menunjukkan bahwa *SMOTE* membantu RF dalam menyeimbangkan distribusi kelas, sehingga prediksi menjadi lebih akurat pada kedua kelas.

Lebih lanjut, dengan *Random Oversampling* (Random OS), *TP* meningkat lebih jauh menjadi 421 dan *FN* turun menjadi 97. Selain itu, *TN* juga meningkat hingga 306, sementara *FP* turun drastis menjadi hanya 94. Kondisi ini memperlihatkan bahwa *Random OS* mampu menghasilkan keseimbangan yang lebih baik dibanding *SMOTE*, karena peningkatan terjadi baik pada deteksi kelas positif maupun pengurangan kesalahan pada kelas negatif.

Temuan ini diperkuat dengan hasil evaluasi pada Tabel 6. Nilai akurasi *baseline* RF hanya 61,87%, meningkat menjadi 76,25% dengan *SMOTE*, dan lebih tinggi lagi 79,19% dengan *Random OS*. Pada metrik *precision*, *baseline* RF berada di 64,12%, meningkat menjadi 80,00% dengan *SMOTE*, dan lebih baik lagi dengan *Random OS* (81,74%).

Dari sisi *recall*, *baseline* RF menghasilkan 70,92%, meningkat menjadi 78,17% setelah *SMOTE*, dan tertinggi dengan *Random OS* (81,27%). Peningkatan ini menunjukkan bahwa *Random OS* lebih efektif dalam membantu RF mengenali kelas positif dengan lebih konsisten.

Terakhir, pada *F1-score*, *baseline* RF hanya 67,35%, lalu meningkat menjadi 79,07% dengan *SMOTE*, dan mencapai nilai tertinggi 81,51% dengan *Random OS*. Nilai ini menegaskan bahwa kombinasi RF + *Random Oversampling* menghasilkan performa terbaik, dengan keseimbangan optimal antara *precision* dan *recall*.

Secara keseluruhan, *Random Oversampling* terbukti paling efektif dalam meningkatkan performa RF dibanding *baseline* dan *SMOTE*. Jika *SMOTE* memberikan peningkatan signifikan terutama pada akurasi dan *recall*, *Random OS* justru lebih unggul karena memberikan perbaikan menyeluruh pada seluruh metrik evaluasi.

Tabel 6. Evaluasi kinerja *Random Forest* dengan dan tanpa *Oversampling*

Evaluasi	RF	RF + <i>SMOTE</i>	RF + RANDOM OS
Accuracy	61,87%	76,25%	79,19%
Precision	64,12%	80,00%	81,74%
Recall	70,92%	78,17%	81,27%
F1-score	67,35%	79,07%	81,51%

Hasil penelitian ini menunjukkan bahwa penerapan teknik *oversampling* tidak hanya memberikan peningkatan kinerja pada algoritma klasifikasi, tetapi juga mengungkapkan pola yang berbeda terkait sensitivitas tiap algoritma terhadap metode *balancing* data. Temuan ini menjadi poin penting yang membedakan penelitian ini dengan studi sebelumnya, karena tidak hanya berfokus pada evaluasi performa tunggal, tetapi juga menganalisis hubungan antara jenis algoritma dengan teknik *oversampling* yang digunakan.

Pada algoritma KNN, penggunaan *SMOTE* terbukti lebih efektif dibandingkan *Random Oversampling*. Hal ini dapat dijelaskan karena KNN sangat bergantung pada distribusi data dalam ruang *vektor*, sehingga data sintetis hasil *SMOTE* yang lebih representatif mampu memperkecil kesalahan klasifikasi dibandingkan duplikasi sederhana pada *Random OS*. Dengan demikian, penelitian ini memperlihatkan bukti empiris bahwa pemilihan teknik *balancing* perlu mempertimbangkan sifat dasar algoritma yang digunakan.

Berbeda dengan KNN, algoritma *Naive Bayes* menunjukkan performa yang relatif stabil baik dengan *SMOTE* maupun *Random Oversampling*. Hal ini dapat dipahami karena asumsi independensi fitur pada NB membuat algoritma ini tidak terlalu terpengaruh pada distribusi ulang data minoritas. Penemuan ini penting karena menegaskan bahwa untuk model berbasis probabilistik, efek *oversampling* bersifat moderat dan tidak menghasilkan perubahan drastis seperti pada algoritma berbasis jarak atau *ensemble*.

Sementara itu, algoritma *Random Forest* justru paling diuntungkan dengan penerapan oversampling, khususnya *Random Oversampling*. Hasil penelitian menunjukkan adanya peningkatan akurasi dan *F1-score* yang lebih tinggi dibandingkan SMOTE. Hal ini menarik karena berlawanan dengan pola pada KNN, dan dapat dijelaskan oleh sifat *ensemble* pada RF yang lebih tangguh terhadap *noise*, sehingga duplikasi data minoritas dari *Random OS* justru memperkuat kemampuan model dalam membentuk pohon keputusan yang lebih seimbang.

Dari temuan ini, dapat disimpulkan bahwa *novelty* penelitian terletak pada analisis mendalam mengenai interaksi antara algoritma klasifikasi dengan metode *oversampling* yang digunakan. Alih-alih hanya melaporkan hasil akurasi atau *F1-score*, penelitian ini menunjukkan bahwa efektivitas teknik *balancing* bergantung pada karakteristik dasar algoritma. Dengan demikian, hasil penelitian ini dapat menjadi rujukan praktis dalam pemilihan strategi *oversampling* yang lebih tepat sasaran, sesuai dengan algoritma yang digunakan dalam klasifikasi data tidak seimbang.

5. KESIMPULAN DAN SARAN

Kesimpulan Penelitian ini membuktikan bahwa penerapan teknik *oversampling* mampu meningkatkan performa model klasifikasi pada dataset penyakit jantung yang tidak seimbang. Hasil eksperimen menunjukkan bahwa: (1) KNN paling optimal dengan SMOTE karena sifat algoritma ini sangat bergantung pada distribusi data, sehingga sampel sintesis yang representatif mampu mengurangi kesalahan klasifikasi; (2) *Naive Bayes* relatif stabil terhadap kedua metode *oversampling*, dengan SMOTE memberikan keseimbangan terbaik antara *precision* dan *recall*; (3) *Random Forest* memperoleh peningkatan tertinggi dengan *Random Oversampling*, berkat sifat *ensemble* yang tangguh terhadap *noise* sehingga duplikasi data minoritas justru memperkuat performa model. Dengan demikian, hasil penelitian ini dapat dijadikan rujukan praktis bagi pengembangan sistem prediksi penyakit jantung maupun studi klasifikasi medis lainnya, khususnya dalam menentukan strategi penanganan data tidak seimbang. Untuk penelitian selanjutnya, beberapa hal dapat menjadi perhatian, yaitu eksplorasi teknik *balancing* lanjutan seperti *Borderline-SMOTE*, *Adaptive-SMOTE*, atau kombinasi SMOTE dengan Tomek Links dapat dilakukan untuk menghasilkan distribusi data yang lebih representatif. Kemudian perlu dilakukan perbandingan dengan algoritma *ensemble* lain seperti *Gradient Boosting* atau *XGBoost* untuk melihat potensi peningkatan akurasi.

UCAPAN TERIMA KASIH

Kami ingin menyampaikan rasa terima kasih kami kepada Yayasan Dinamika Bangsa Jambi melalui Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) Universitas Dinamika Bangsa atas dukungan finansial dan fasilitas serta atas keberhasilan penyelesaian penelitian ini melalui program Hibah Penelitian Dosen Universitas Dinamika Bangsa dengan nomor 012/MOU/LPPM/UNAMA/IV/2025.

DAFTAR REFERENSI

- Alham, S. R. J. I. (2021). Sistem Diagnosis Penyakit Jantung Koroner Dengan Menggunakan Algoritma C4.5 Berbasis Website (Studi Kasus: RSUD Dr. Soedarso Pontianak). *Petir*, 14(2), 214–222. <https://doi.org/10.33322/petir.v14i2.1338>
- Alwan, J. K., Jaafar, D. S., & Ali, I. R. (2022). Diabetes diagnosis system using modified Naive Bayes classifier. *Indonesian Journal of Electrical Engineering and Computer Science*, 28(3), 1766–1774. <https://doi.org/https://doi.org/10.11591/ijeecs.v28.i3.pp1766-1774>
- Amen, K., Zohdy, M., & Mahmoud, M. (2020). *Machine Learning for Multiple Stage Heart Disease Prediction*. 205–223. <https://doi.org/10.5121/csit.2020.101118>
- Anderson, C. J., Cadeddu, R., Anderson, D. N., Huxford, J. A., VanLuik, E. R., Odeh, K., Pittenger, C., Pulst, S. M., & Bortolato, M. (2024). A novel naïve Bayes approach to identifying grooming behaviors in the force-plate actometric platform. *Journal of Neuroscience Methods*, 403(July 2023), 110026. <https://doi.org/10.1016/j.jneumeth.2023.110026>
- Assegie, T. A., Subhashni, R., Kumar, N. K., Manivannan, J. P., Duraisamy, P., & Engidaye, M. F. (2022). Random forest and support vector machine-based hybrid liver disease detection. *Bulletin of Electrical Engineering and Informatics*, 11(3), 1650–1656. <https://doi.org/https://doi.org/10.11591/eei.v11i3.3787>
- Badar, M., & Fisichella, M. (2024). Fair-CMNB: Advancing Fairness-Aware Stream Learning with Naïve Bayes and Multi-Objective Optimization. *Big Data and Cognitive Computing*, 8(2). <https://doi.org/https://doi.org/10.3390/bdcc8020016>
- Bahri, S., Marisa Midyanti, D., Hidayati, R., Sistem Komputer, J., & Mipa, F. (2018). Perbandingan Algoritma Naive Bayes dan C4.5 Untuk Klasifikasi Penyakit Anak. *Seminar Nasional Aplikasi Teknologi Informasi (SNATi)*, 11–2018.
- Berrar, D. (2018). Bayes' theorem and naive bayes classifier. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1–3(2018), 403–412. <https://doi.org/10.1016/B978-0-12-809633-8.20473-1>
- Chivukula, R., Jaya Lakshmi, T., Uday, S. S., & Pavani, S. T. (2021). Classifying clinically actionable genetic mutations using KNN and SVM. *Indonesian Journal of Electrical Engineering and Computer Science*, 24(3), 1672–1679. <https://doi.org/https://doi.org/10.11591/ijeecs.v24.i3.pp1672-1679>
- Elin Nurlia, U. E. (2021). *PENERAPAN FITUR SELEKSI FORWARD SELECTION UNTUK MENENTUKAN KEMATIAN AKIBAT GAGAL JANTUNG MENGGUNAKAN*. 6(1), 42–50.

- Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, *11*(1). <https://doi.org/10.1186/s40537-024-00973-y>
- Hasib, K. M., Iqbal, M. S., Shah, F. M., Mahmud, J. Al, Popel, M. H., Showrov, M. I. H., Ahmed, S., & Rahman, O. (2020). A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem. *Journal of Computer Science*, *16*(11), 1546–1557. <https://doi.org/10.3844/JCSSP.2020.1546.1557>
- Ige, T., Kiekintveld, C., Piplai, A., Wagglar, A., Kolade, O., & Matti, B. H. (2024). *An investigation into the performances of the Current state-of-the-art Naive Bayes, Non-Bayesian and Deep Learning Based Classifier for Phishing Detection: A Survey*. <http://arxiv.org/abs/2411.16751>
- Islam, M. S., Hasan, M. M., Rahim, M. A., Hasan, A. M., Mynuddin, M., Khandokar, I., & Islam, M. J. (2022). Machine Learning-Based Music Genre Classification with Pre-Processed Feature Analysis. *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika*, *7*(3), 491. <https://doi.org/https://doi.org/10.26555/jiteki.v7i3.22327>
- Khaleel, A. A., Al-Azzawi, A. A. M., & Alkhazraji, A. M. (2023). Random forest for lung cancer analysis using Apache Mahout and Hadoop based on software defined networking. *Indonesian Journal of Electrical Engineering and Computer Science*, *32*(2), 1086–1093. <https://doi.org/https://doi.org/10.11591/ijeecs.v32.i2.pp1086-1093>
- Liang, X. W., Jiang, A. P., Li, T., Xue, Y. Y., & Wang, G. T. (2020). LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM. *Knowledge-Based Systems*, *196*. <https://doi.org/10.1016/j.knosys.2020.105845>
- Malek, N. H. A., Yaacob, W. F. W., Wah, Y. B., Md Nasir, S. A., Shaadan, N., & Indratno, S. W. (2023). Comparison of ensemble hybrid sampling with bagging and boosting machine learning approach for imbalanced data. *Indonesian Journal of Electrical Engineering and Computer Science*, *29*(1), 598–608. <https://doi.org/10.11591/ijeecs.v29.i1.pp598-608>
- Merdekawati, A. (2022). *Komparasi Algoritma Data Mining dan Perancangan Aplikasi Prediksi Harapan Hidup Pasien Gagal Jantung*. *14*(3), 188–202.
- Mohammadagha, M. (n.d.). *Hyperparameter Optimization Strategies for Tree-Based Machine Learning Models Prediction : A Comparative Study of AdaBoost , Decision Trees , and Random Forest*.
- Nadeem, M., Arshad, A., Riaz, S., Zahra, S. W., Rashid, M., Band, S. S., & Mosavi, A. (2023). Preventing Cloud Network from Spamming Attacks Using Cloudflare and KNN. *Computers, Materials and Continua*, *74*(2), 2641–2659. <https://doi.org/https://doi.org/10.32604/cmc.2023.028796>
- Ngo, H. L., Nguyen, H. D., Loubiere, P., Tran, T. Van, Şerban, G., Zelenakova, M., Breţcan, P., & Laffly, D. (2022). The composition of time-series images and using the technique SMOTE ENN for balancing datasets in land use/cover mapping. *Acta Montanistica Slovaca*, *27*(2), 342–359. <https://doi.org/10.46544/AMS.v27i2.05>
- Nguyen, L. V., Vo, Q. T., & Nguyen, T. H. (2023). Adaptive KNN-Based Extended Collaborative Filtering Recommendation Services. *Big Data and Cognitive Computing*, *7*(2). <https://doi.org/https://doi.org/10.3390/bdcc7020106>
- Nur Riza Pahlevi, M., & Badriyah, T. (2025). Implementasi dan Optimasi Hyperparameter pada Model Machine learning untuk Prediksi Diabetes dengan Integrasi Aplikasi Telemedicine. *JEPIN (Jurnal Edukasi Dan Penelitian Informatika)*, *11*(2), 287–296.

- Pan, T., Zhao, J., Wu, W., & Yang, J. (2020). Learning imbalanced datasets based on SMOTE and Gaussian distribution. *Information Sciences*, 512, 1214–1233. <https://doi.org/10.1016/j.ins.2019.10.048>
- Pratama, Y., Prayitno, A., Nazrian, D., Aini, N., R, Y. R., & Rasywir, E. (2022). *BULLETIN OF COMPUTER SCIENCE RESEARCH* Klasifikasi Penyakit Gagal Jantung Menggunakan Algoritma K-Nearest Neighbor. 3(1), 52–56. <https://doi.org/10.47065/bulletincsr.v3i1.203>
- Rapacz, S., Cholda, P., & Natkaniec, M. (2021). A method for fast selection of machine-learning classifiers for spam filtering. *Electronics (Switzerland)*, 10(17). <https://doi.org/https://doi.org/10.3390/electronics10172083>
- Reza, D. A. M., Siregar, A. M., & Rahmat. (2022). Penerapan Algoritma K-Nearest Neighbord Untuk Prediksi Kematian Akibat Penyakit Gagal Jantung. *Scientific Student Journal for Information, Technology and Science*, III(1), 105–112.
- Sahar, S. (2020). Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Clasiffier Pada Dataset Penyakit Jantung. *Indonesian Journal of Data and Science*, 1(3), 79–86. <https://doi.org/10.33096/ijodas.v1i3.20>
- Samosir, A., Hasibuan, M. S., Justino, W. E., & Hariyono, T. (2021). Komparasi Algoritma Random Forest, Naïve Bayes dan K- Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung. *Prosiding Seminar Nasional Darmajaya*, 1(0), 214–222. <https://jurnal.darmajaya.ac.id/index.php/PSND/article/view/2955>
- Sampath, P., Elangovan, G., Ravichandran, K., Shanmuganathan, V., Pasupathi, S., Chakrabarti, T., Chakrabarti, P., & Margala, M. (2024). Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique. *Scientific Reports*, 14(1), 1–15. <https://doi.org/10.1038/s41598-024-78519-8>
- Sekulić, A., Kilibarda, M., Heuvelink, G. B. M., Nikolić, M., & Bajat, B. (2020). Random forest spatial interpolation. *Remote Sensing*, 12(10), 1–29. <https://doi.org/https://doi.org/10.3390/rs12101687>
- Sepharni, A., Hendrawan, I. E., & Rozikin, C. (2022). Klasifikasi Penyakit Jantung dengan Menggunakan Algoritma C4.5. *STRING (Satuan Tulisan Riset Dan Inovasi Teknologi)*, 7(2), 117. <https://doi.org/10.30998/string.v7i2.12012>
- Shakeela, S., Shankar, N. S., Reddy, P. M., Tulasi, T. K., & Koneru, M. M. (2021). Optimal ensemble learning based on distinctive feature selection by univariate ANOVA-F statistics for IDS. *International Journal of Electronics and Telecommunications*, 67(2), 267–275. <https://doi.org/10.24425/ijet.2021.135975>
- Soltanzadeh, P., & Hashemzadeh, M. (2021). RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem. *Information Sciences*, 542, 92–111. <https://doi.org/10.1016/j.ins.2020.07.014>
- Sugiyarto, A. W., Abadi, A. M., & Sumarna. (2021). Classification of heart disease based on PCG signal using CNN. *Telkomnika (Telecommunication Computing Electronics and Control)*, 19(5), 1697–1706. <https://doi.org/10.12928/TELKOMNIKA.v19i5.20486>
- Syahputra, H., & Wibowo, A. (2023). Comparison of Support Vector Machine (SVM) and Random Forest Algorithm for Detection of Negative Content on Websites. *Jurnal Ilmiah Teknik Elektro Komputer Dan Informatika (JITEKI)*, 9(1), 165–173. <https://doi.org/https://doi.org/10.26555/jiteki.v9i1.25861>

- Wang, S., Ren, J., & Bai, R. (2023). A semi-supervised adaptive discriminative discretization method improving discrimination power of regularized naive Bayes. *Expert Systems with Applications*, 225(April), 120094. <https://doi.org/10.1016/j.eswa.2023.120094>
- Wang, X., Zhai, M., Ren, Z., Ren, H., Li, M., Quan, D., Chen, L., & Qiu, L. (2021). Exploratory study on classification of diabetes mellitus through a combined Random Forest Classifier. *BMC Medical Informatics and Decision Making*, 21(1), 1–14. <https://doi.org/10.1186/s12911-021-01471-4>
- Xin, L. K., & Rashid, N. binti A. (2021). Prediction of depression among women using random oversampling and random forest. *2021 International Conference of Women in Data Science at Taif University, WiDSTaif 2021*. <https://doi.org/10.1109/WIDSTAI52235.2021.9430215>
- Yang, Y., & Liu, X. (n.d.). *A re-examination of text categorization methods*.
- Zhang, J., Li, Y., Shen, F., He, Y., Tan, H., & He, Y. (2024). Hierarchical text classification with multi-label contrastive learning and KNN. *Neurocomputing*, 577(January), 127323. <https://doi.org/https://doi.org/10.1016/j.neucom.2024.127323>
- Zhu, Y., Kong, B., Liu, R., & Zhao, Y. (2022). Developing biomedical engineering technologies for reproductive medicine. *Smart Medicine*, 1(1). <https://doi.org/10.1002/smmd.20220006>