Meningkatkan Klasifikasi Penyakit Diabetes Menggunakan Metode Ensemble Softvoting Dengan SMOTE-ENN dan Optimasi Bayesian

Ilham Maulana[1], Siti Ernawati[2*]

[1]Ilmu Komputer, Universitas Nusa Mandiri Indonesia
[2]Sistem Informasi, Universitas Nusa Mandiri Indonesia

** Corresponding Author.E-mail:siti.ste@nusamandiri.ac.id*

Abstrak

Diabetes sebagai salah satu penyakit metabolik kronis, terus menjadi tantangan besar dalam kesehatan masyarakat global. Banyak kasus diabetes yang tidak terdiagnosis, sehingga deteksi dini menjadi langkah penting untuk mencegah atau menunda komplikasi serius. Dalam penelitian ini, mengusulkan pendekatan berbasis ensemble learning menggunakan Softvoting Classifier, dikombinasikan dengan teknik SMOTE-ENN untuk menangani ketidakseimbangan data serta optimasi Bayesian guna meningkatkan performa model. Penelitian ini menggunakan dataset PIMA Indian Diabetes sebanyak 768 data dan Diabetes Prediction Dataset (DPD) sebanyak 100.000 data. Proses klasifikasi melibatkan algoritma Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), k-Nearest Neighbors (K-NN), Extreme Gradient Boosting (XGB), Decision Tree (DT), dan AdaBoost (ADA), kemudian dikombinasikan dalam metode Softvoting untuk menghasilkan prediksi yang lebih akurat. Evaluasi model dilakukan dengan metode train-test split dan cross-validation (CV). Hasil eksperimen menunjukkan bahwa metode Softvoting dengan SMOTE-ENN dan Bayesian Optimation mampu meningkatkan kinerja model secara signifikan. Model terbaik pada data PIMA mencapai akurasi sebesar 0.987179 pada model RF, GB, K-NN, DT, ADA, dan Softvoting dengan menggunakan 10-fold cross-validation. Sedangkan model terbaik pada data DPD memperoleh akurasi 0.998059 pada eksperimen yang telah dilakukan SMOTE-ENN dan Bayesian Optimation pada model Gradiant Boosting. Pendekatan yang diusulkan menunjukkan efektivitas dalam meningkatkan akurasi deteksi diabetes, terutama dalam kondisi data yang tidak seimbang. Implementasi metode ensemble Softvoting dengan SMOTE-ENN dan Bayesian Optimation dapat menjadi solusi dalam diagnosis dini diabetes serta berkontribusi terhadap kemajuan machine learning dalam dunia medis.

Keywords: Ensemble Softvoting; Klasifikasi; Machine Learning; Optimasi Bayesian; SMOTE-ENN

Abstract

Diabetes, as one of the chronic metabolic diseases, continues to be a major challenge in global public health. Many cases of diabetes remain undiagnosed, making early detection a crucial step in preventing or delaying serious complications. This study proposes an ensemble learning-based approach using the Softvoting Classifier, combined with the SMOTE-ENN technique to handle data imbalance and Bayesian Optimization to enhance model performance. This study utilizes the PIMA Indian Diabetes dataset with 768 records and the Diabetes Prediction Dataset (DPD) with 100,000 records. The classification process involves Logistic

Regression (LR), Random Forest (RF), Gradient Boosting (GB), k-Nearest Neighbors (K-NN), Extreme Gradient Boosting (XGB), Decision Tree (DT), and AdaBoost (ADA), which are then combined using the Softvoting method to produce more accurate predictions. Model evaluation is performed using the train-test split method and cross-validation (CV). Experimental results show that the Softvoting method with SMOTE-ENN and Bayesian Optimization significantly improves model performance. The best model for the PIMA dataset achieved an accuracy of 0.987179 using RF, GB, K-NN, DT, ADA, and Softvoting with 10-fold cross-validation. Meanwhile, the best model for the DPD dataset obtained an accuracy of 0.998059 in experiments using SMOTE-ENN and Bayesian Optimization on the Gradient Boosting model. The proposed approach demonstrates effectiveness in improving diabetes detection accuracy, particularly in imbalanced data conditions. The implementation of the ensemble Softvoting method with SMOTE-ENN and Bayesian Optimization can serve as a solution for early diabetes diagnosis and contribute to advancements in machine learning in the medical field.

Keywords: Ensemble Softvoting; Klasifikasi; Machine Learning; Optimasi Bayesian; SMOTE-ENN

1. Pendahuluan

Diabetes mellitus memiliki peran penting dalam meningkatkan risiko kematian, terutama melalui komplikasi kardiovaskular dan penyakit terkait lainnya. Penelitian menunjukkan bahwa diabetes mellitus dapat melipatgandakan risiko kematian akibat penyakit jantung koroner (Anharudin & Mila Tejamaya, 2022)(Anharudin & Mila Tejamaya, 2022). Selain itu, diabetes juga berkontribusi terhadap peningkatan insiden penyakit ginjal kronis, yang merupakan salah satu penyebab utama kematian di kalangan penderita diabetes.

Deteksi dini dapat mencegah komplikasi kronis seperti neuropati, nefropati, dan retinopati, yang sering kali terjadi pada pasien diabetes (Lukitaningtyas, Kurniasih, & Pariyem, 2022). Salah satu metode yang efektif untuk deteksi dini adalah melalui pemeriksaan gula darah secara rutin, yang memungkinkan identifikasi awal terhadap kondisi diabetes (Hatmanti et al., 2023). Selain itu, edukasi kepada masyarakat dan kader kesehatan tentang pentingnya deteksi dini juga berperan besar dalam meningkatkan kesadaran dan pengetahuan mengenai diabetes (Nistiandani, Siswoyo, Sakinah, Rahmah, & Lisar, 2024). Deteksi dini diabetes mellitus sangat penting untuk mencegah komplikasi serius yang dapat mengancam jiwa.

Klasifikasi diabetes menggunakan data medis menghadapi berbagai tantangan yang signifikan. Salah satu tantangan utama adalah kompleksitas dan variabilitas data medis itu sendiri, yang sering kali mencakup berbagai fitur yang tidak terstruktur dan beragam, seperti hasil laboratorium, riwayat kesehatan, dan faktor demografis (Prasetyo & Nabiilah, 2023). Hal ini dapat menyulitkan algoritma

klasifikasi dalam mengidentifikasi pola yang konsisten. Selain itu, pemilihan algoritma yang tepat juga menjadi tantangan. Berbagai metode, seperti K-NN, Support Vector Machine (SVM), dan Naïve Bayes, memiliki kelebihan dan kekurangan masing-masing dalam hal akurasi dan kecepatan (Dewi Nasien et al., 2024) (Anisa & Jumanto, 2022). Misalnya, SVM sering kali memberikan hasil yang lebih baik pada data yang tidak seimbang, tetapi memerlukan waktu pelatihan yang lebih lama.

Masalah *overfitting* juga perlu diperhatikan, di mana model yang terlalu kompleks dapat mempelajari *noise* dalam data pelatihan dan tidak generalisasi dengan baik pada data baru (Ferdina, Satyahadewi, & Kusnandar, 2023). Oleh karena itu, teknik seperti pengurangan dimensi dan pemilihan fitur yang tepat sangat penting untuk meningkatkan performa model klasifikasi (Yusa, Coastera, & Yandika, 2022) (Alpiansah & Ramdhani, 2023). Penggunaan metode ensemble seperti *Softvoting*, bersama dengan teknik resampling seperti SMOTE-ENN, dan optimasi Bayesian, memiliki banyak manfaat dalam meningkatkan akurasi dan keandalan model klasifikasi diabetes. Metode ensemble, khususnya *Softvoting*, menggabungkan prediksi dari beberapa model untuk menghasilkan keputusan akhir, dan membantu mengurangi variabilitas dan meningkatkan akurasi prediksi.

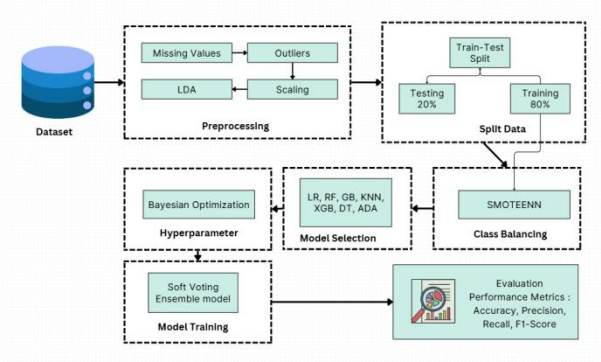
Teknik SMOTE-ENN, yang menggabungkan *oversampling* dan *undersampling*, efektif dalam menangani masalah ketidakseimbangan data. Dengan menciptakan contoh sintetis untuk kelas minoritas dan menghapus contoh yang tidak relevan dari kelas mayoritas, SMOTE-ENN dapat meningkatkan kinerja model dalam klasifikasi diabetes (Izraiq et al., 2024) (Xu et al., 2022). Penelitian menunjukkan bahwa penggunaan SMOTE-ENN dapat secara signifikan meningkatkan *recall* dan akurasi model dibandingkan dengan metode resampling tunggal (Abdulsadig & Rodriguez-Villegas, 2024). Optimasi Bayesian juga berperan penting dalam menemukan *hyperparameter* terbaik untuk model, yang dapat meningkatkan kinerja model secara keseluruhan. Dengan menggunakan pendekatan ini, peneliti dapat secara efisien bereksperimen terhadap parameter dan menemukan kombinasi yang optimal untuk meningkatkan akurasi prediksi diabetes. Kombinasi dari ketiga metode ini memberikan pendekatan yang kuat dan efektif dalam klasifikasi diabetes.

Tujuan penelitian yang dilakukan yaitu untuk mengoptimalkan performa model *ensemble Softvoting* untuk klasifikasi penyakit diabetes dengan menerapkan *Bayesian Optimization* untuk menemukan

kombinasi *hyperparameter* terbaik, sehingga meningkatkan akurasi deteksi diabetes secara signifikan pada dataset medis, dan mengevaluasi efektivitas metode *Ensemble Softvoting* yang telah dioptimalkan melalui Bayesian Optimization dalam mendeteksi diabetes menggunakan metrik evaluasi seperti *accuracy, precision, recall,* dan *F1-score* untuk memastikan model memiliki akurasi dan sensitivitas yang optimal dalam klasifikasi diabetes.

2. Bahan dan Metode

Penelitian yang dibahas mengenai peningkatkan klasifikasi penyakit diabetes menggunakan metode *ensemble Softvoting* dengan SMOTE-ENN dan optimasi bayesian membutuhkan pemahaman mendalam tentang konsep-konsep dasar dalam machine learning. *Machine learning* berperan penting dalam pengembangan sistem yang dapat secara otomatis mempelajari pola dari data tanpa memerlukan pemrograman eksplisit. Dalam penelitian ini, machine learning digunakan untuk membangun model klasifikasi data medis, seperti dataset PIMA dan DPD, untuk mendeteksi diabetes. Sementara itu, *Bayesian Optimization* diterapkan untuk menemukan kombinasi *hyperparameter* terbaik pada algoritma yang digunakan dalam *ensemble Softvoting*, dengan tujuan meningkatkan akurasi prediksi dan menghasilkan model yang lebih efisien dalam mengklasifikasikan kondisi medis. gambar 1 menunjukan alur dari penelitian yang dilakukan.



Gambar 2. Alur Penelitian

2.1 Dataset

Penelitian ini menggunakan dua dataset, yaitu PIMA Indian Diabetes sebanyak 768 data dan Diabetes Prediction Dataset (DPD) sebanyak 100.000 data, untuk mengembangkan dan mengevaluasi model klasifikasi yang diusulkan. Dengan memanfaatkan PIMA dan DPD, penelitian ini mendapatkan manfaat dari kondisi terkendali pada dataset yang lebih kecil dan representasi yang lebih luas dari dataset yang lebih besar, memastikan evaluasi yang komprehensif terhadap model *ensemble* yang diusulkan.

2.2. Data Preprocessing

Data *preprocessing* adalah tahap penting untuk memastikan kualitas data sebelum analisis atau pemodelan *machine learning.* Proses ini mencakup penanganan *missing values* dengan imputasi atau penghapusan,

identifikasi dan penanganan outlier menggunakan *boxplot, z-score*, atau IQR, serta normalisasi atau standarisasi data numerik. Data kategorikal dikonversi ke format numerik menggunakan encoding seperti *one-hot* atau *label encoding.* Jika terdapat ketidakseimbangan kelas, teknik seperti SMOTE atau *undersampling* digunakan untuk menyeimbangkan distribusi data. Dengan preprocessing yang baik, dataset menjadi lebih bersih, terstruktur, dan siap untuk pelatihan model, yang berkontribusi pada peningkatan performa dan validitas hasil penelitian.

2.3. Split Data

Dalam eksperimen ini, data dibagi menjadi dua bagian dengan proporsi 80% untuk data latih dan 20% untuk data uji. Pembagian ini bertujuan untuk memastikan bahwa model dapat belajar dari sebagian besar data yang tersedia, sementara sisanya digunakan untuk menguji kinerja model terhadap data yang belum pernah dilihat sebelumnya. Dengan 80% data latih, model memiliki cukup informasi untuk mengenali pola dan melakukan generalisasi, sedangkan 20% data uji membantu mengevaluasi sejauh mana model dapat mengklasifikasikan data baru dengan akurasi yang baik. Pendekatan ini umum digunakan dalam machine learning untuk menghindari *overfitting*, di mana model terlalu menyesuaikan diri dengan data latih dan gagal berkinerja baik pada data baru. Selain itu, pemisahan data dilakukan secara acak *(random split)* untuk memastikan distribusi kelas tetap seimbang, sehingga hasil evaluasi lebih representatif terhadap performa model secara keseluruhan.

2.4. Class Balancing (SMOTE-ENN)

Metode SMOTE-ENN *(Synthetic Minority Over-sampling Technique-Edited Nearest Neighbors)* merupakan pendekatan yang efektif untuk menangani masalah ketidakseimbangan kelas dalam dataset. SMOTE berfungsi untuk menghasilkan data sintetis dari kelas minoritas dengan cara interpolasi antara titik data minoritas dan tetangga terdekatnya, sedangkan ENN bertugas untuk menghapus contoh yang salah klasifikasi dari kedua kelas (Prabha & Sasikala, 2024) (Ullah et al., 2022). Kombinasi kedua metode ini tidak hanya meningkatkan jumlah data minoritas tetapi juga membersihkan data dari noise yang dapat mengganggu proses pembelajaran.

Penelitian menunjukkan bahwa penerapan SMOTE-ENN dapat meningkatkan akurasi model *machnie learning*, seperti *Random Forest* dan *Support Vector Machine*, dalam berbagai aplikasi, termasuk diagnosis medis dan prediksi penyakit. Misalnya, dalam analisis data diabetes dan keguguran, penggunaan SMOTE-ENN terbukti lebih efektif dibandingkan dengan metode *oversampling*

atau *undersampling* yang berdiri sendiri (Yang et al., 2022) (Nuanmeesri & Poomhiran, 2023). Dengan demikian, SMOTE-ENN menjadi pilihan yang dalam pengembangan model yang memerlukan keseimbangan antara kelas minoritas dan mayoritas.

2.5. Model Selection

Dalam proses *model selection*, beberapa algoritma *machine learning* dibandingkan untuk menentukan model dengan performa terbaik dalam menyelesaikan tugas klasifikasi. Model yang diuji meliputi *Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), k-Nearest Neighbors (K-NN), XGBoost (XGB), Decision Tree (DT), dan AdaBoost (ADA). Logistic Regression (LR)* digunakan sebagai baseline karena sifatnya yang sederhana dan interpretatif. RF dan DT bekerja dengan membangun pohon keputusan, di mana RF lebih unggul dalam mengurangi *overfitting* melalui *ensemble learning. Gradient Boosting (GB), XGBoost (XGB),* dan *AdaBoost (ADA)* merupakan algoritma *boosting* yang meningkatkan akurasi dengan menggabungkan beberapa model lemah menjadi model kuat. Sementara itu, K-NN mengklasifikasikan data berdasarkan kedekatan dengan tetangga terdekatnya.

2.6. Hyperparameter

Pada tahap ini, dilakukan pencarian kombinasi *hyperparameter* yang optimal agar model dapat bekerja dengan baik, sehingga model dapat memberikan hasil yang lebih akurat (Ernawati & Wati, 2024). Hyperparameter yang digunakan dalam penelitian yaitu *Bayesian Optimization (BO),* yang merupakan metode yang sangat efektif untuk melakukan tuning *hyperparameter* dalam model *machine learning.* Metode ini menggunakan model probabilistik, seperti *Gaussian Processes,* untuk memprediksi fungsi yang tidak diketahui dan mengeksplorasi ruang parameter dengan cara yang efisien (Oei et al., 2023) (Amou, Xia, Kamhi, & Mouhafid, 2022). Dalam konteks ini, BO memungkinkan pemilihan *hyperparameter* yang optimal dengan meminimalkan jumlah evaluasi yang diperlukan, sehingga menghemat waktu dan sumber daya.

Dalam penelitian yang dilakukan oleh Oei et al., BO digunakan untuk mengoptimalkan hyperparameter dalam model prediksi risiko pasca-stroke, menunjukkan peningkatan kinerja model yang signifikan (Oei et al., 2023). Selain itu, Amou et al. menerapkan BO untuk mengoptimalkan hyperparameter dalam klasifikasi tumor otak menggunakan CNN, dengan fungsi akuisisi yang membantu menyeimbangkan eksplorasi dan eksploitasi ruang parameter (Amou et al., 2022). Penelitian lain oleh Muzayanah membandingkan efektivitas BO dengan metode lain seperti *GridSearchCV* dalam

konteks prediksi diabetes, menunjukkan bahwa BO lebih efisien dalam menemukan kombinasi hyperparameter yang optimal. Berikut merupakan Teorema Bayes dapat dinyatakan sebagai:

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} = \frac{P(D|H)P(H)}{\int_{H'} P(D|H')P(H')dH'} = \frac{P(D,H)}{\int_{H'} P(D,H')dH'} \quad (1)$$

Dimana:

$P(H|D)$ adalah probabilitas posterior, yaitu probabilitas hipotesis H setelah melihat data D.

$P(D|H)$ adalah likelihood, yaitu probabilitas mendapatkan data D jika hipotesis H benar.

$P(H)$ adalah prior probability, yaitu probabilitas awal dari hipotesis H sebelum melihat data.

$P(D)$ adalah evidence, yaitu probabilitas total dari data D di semua kemungkinan hipotesis.

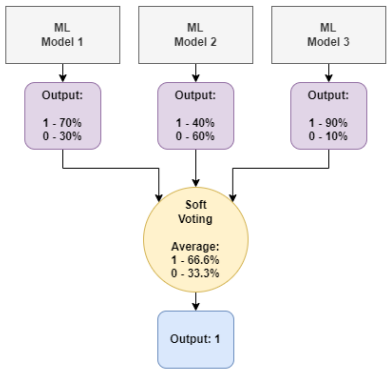
$\int_H P(D|H')P(H')dH'$ adalah menunjukkan bahwa P(D) dihitung dengan menjumlahkan (mengintegralkan) semua kemungkinan kombinasi H' (hipotesis alternatif) yang dapat menjelaskan D.

2.7. Model Training

Ensemble learning adalah teknik dalam *machine learning* yang menggabungkan beberapa model untuk meningkatkan akurasi prediksi dan mengurangi risiko *overfitting*. Metode ini berfungsi dengan memanfaatkan kekuatan dari berbagai algoritma pembelajaran untuk menghasilkan hasil yang lebih baik dibandingkan dengan model tunggal Helmut (Helmut & Murdiansyah, 2023) (Ye, Li, Zhang, & Xu, 2024). Terdapat beberapa pendekatan dalam *ensemble learning*, termasuk *bagging*, *boosting*, dan *stacking*, yang masing-masing memiliki cara unik dalam menggabungkan prediksi dari model-model dasar.

Salah satu metode *Ensemble learning* yaitu *Voting*. *Voting* adalah metode *ensemble learning* yang menggabungkan prediksi dari beberapa model untuk menghasilkan keputusan akhir berdasarkan suara *(vote)* dari masing-masing model. Dalam *voting*, setiap model memberikan prediksi, dan kelas dengan jumlah suara terbanyak *(majority voting)* atau dengan bobot tertentu *(weighted voting)* dipilih sebagai hasil akhir (Desiani et al., 2023) (Desiani et al., 2024). Metode ini sangat populer karena kesederhanaannya dan kemampuannya untuk meningkatkan akurasi prediksi dengan mengurangi varians yang mungkin terjadi pada model tunggal. *Voting* memiliki dua jenis yaitu *hardvoting* dan *softvoting*. Dalam penelitian ini menggunakan *softvoting*. Metode ini merupakan teknik yang efektif untuk meningkatkan akurasi klasifikasi dengan menggabungkan prediksi dari beberapa model. Dalam *softvoting* menghitung rata-rata probabilitas prediksi dari setiap model dan memilih kelas dengan probabilitas

tertinggi sebagai hasil akhir (Das, Panda, & Rautray, 2023). Alur perhitungan rata-rata probabilitas prediksi dari setiap model disajikan dalam gambar 3.



Gambar 4. Alur Proses Softvoting

2.8. Evaluation

Dalam proses evaluasi model, digunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk menilai kinerja masing-masing algoritma. *Accuracy* mengukur persentase prediksi yang benar dari keseluruhan data, namun metrik ini bisa kurang efektif jika terjadi ketidakseimbangan kelas. Oleh karena itu, digunakan precision, yang menunjukkan seberapa banyak prediksi positif yang benar dibandingkan total prediksi positif, serta *recall*, yang mengukur sejauh mana model dapat menemukan semua sampel positif dalam dataset. Untuk mendapatkan keseimbangan antara *precision* dan *recall*, digunakan *F1-score*, yang merupakan rata-rata harmonis dari keduanya. Model dengan *F1-score* tinggi menunjukkan performa yang baik dalam menangani *trade-off* antara *precision* dan *recall*, terutama dalam kasus klasifikasi dengan distribusi kelas yang tidak seimbang. Evaluasi dilakukan dengan membandingkan hasil metrik dari beberapa algoritma, sehingga dapat dipilih model terbaik yang memiliki keseimbangan optimal antara keempat metrik tersebut.

3. Hasil dan Diskusi

Penelitian ini menggunakan dua dataset, yaitu PIMA Indian Diabetes sebanyak 768 data dan *Diabetes Prediction Dataset (DPD)* sebanyak 100.000 data, untuk mengembangkan dan mengevaluasi model klasifikasi yang diusulkan. Menggunakan pendekatan *Softvoting* untuk meningkatkan akurasi klasifikasi diabetes. Metode ini menggabungkan prediksi dari beberapa model dasar, seperti *Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), K-Nearest Neighbors (K-NN), XGBoost (XGB), Decision Tree (DT),* dan *AdaBoost (ADA).* Kombinasi model ini dipilih karena masing-masing memiliki kekuatan unik yang dapat saling melengkapi.

3.1. Data Preprocessing

Langkah pertama dalam preprocessing adalah mengidentifikasi dan menangani missing values. Hal ini dilakukan karena missing values dapat memengaruhi kualitas hasil analisis. Untuk menangani *missing values*, beberapa metode yang umum

digunakan meliputi penghapusan data yang tidak lengkap (jika proporsi data missing values kecil) atau imputasi menggunakan nilai rata-rata *(mean)*, nilai tengah *(median)*, atau nilai yang paling sering muncul *(mode)*. Metode yang dipilih bergantung pada sifat data dan tingkat missing values dalam dataset.

Selanjutnya, data *outlier* diidentifikasi menggunakan metode seperti boxplot, z-score, atau IQR *(interquartile range)*. *Outlier* yang signifikan dapat dihapus jika berpotensi mengganggu hasil analisis atau dapat ditangani dengan transformasi data untuk mengurangi pengaruhnya. Selain itu, data numerik sering kali dinormalisasi atau distandarisasi untuk menyamakan skala atribut.

Data kategorikal, seperti variabel jenis kelamin atau status perokok, juga perlu diubah menjadi format numerik agar dapat diproses oleh model *machine learning*. Proses ini dilakukan menggunakan metode encoding, seperti *one-hot encoding* untuk menciptakan variabel biner baru atau label encoding untuk memberikan nilai numerik pada setiap kategori.

Terakhir, jika terdapat ketidakseimbangan pada kelas target, metode penanganan seperti SMOTE, *undersampling*, atau SMOTE-ENN dapat digunakan untuk memperbaiki distribusi kelas. Teknik ini bertujuan untuk meningkatkan performa model pada kelas minoritas tanpa mengorbankan kualitas data.

3.2. Eksperimen Dengan Dataset PIMA

Dataset PIMA Indian Diabetes adalah dataset publik yang dikenal luas, dikumpulkan oleh National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Dataset ini terdiri dari 768 sampel dan 9 fitur, yaitu Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age, Outcome. Deskripsi dari masing-masing fitur dapat dilihat pada tabel 1. Namun, dataset ini menghadirkan tantangan seperti ketidakseimbangan kelas, dengan jumlah kasus diabetes positif yang lebih sedikit, nilai yang hilang, dan kemungkinan keberadaan outlier. Selain itu, ukuran dataset yang terbatas dapat mempengaruhi generalisasi model machine learning.

Tabel 1. Deskripsi Dataset PIMA

Fitur/Atribut	Deskripsi
Pregnancies	Jumlah kehamilan yang pernah dialami oleh individu
Glucose	Konsentrasi glukosa plasma setelah 2 jam dalam tes toleransi glukosa oral
BloodPressure	Tekanan darah diastolik (mm Hg)
SkinThickness	Ketebalan kulit pada area triceps (diukur dalam mm)
Insulin	Tingkat insulin dalam serum (μU/ml)
BMI	Indeks Massa Tubuh (berat dalam kg/(tinggi dalam m)^2)
DiabetesPedigreeFunction	Skor yang menunjukkan kemungkinan risiko diabetes

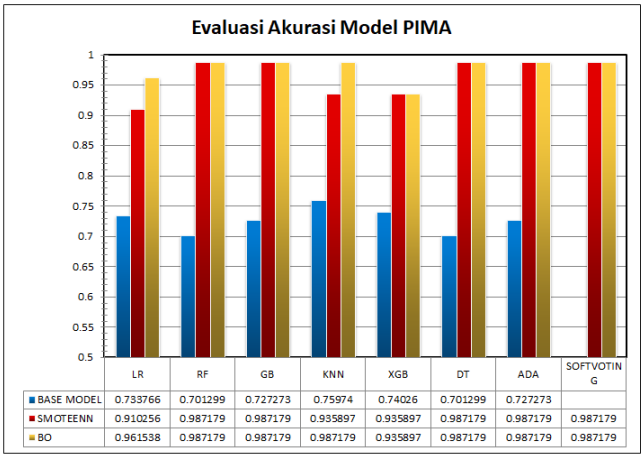
Fitur/Atribut	Deskripsi
	berdasarkan riwayat keluarga
Age	Umur individu (dalam tahun)
Outcome	Variabel target yang menunjukkan apakah individu menderita diabetes (1) atau tidak (0)

Hasil evaluasi dari eksperimen dengan menggunakan data PIMA menunjukan bahwa nilai akurasi yang paling tinggi terdapat pada eksperimen yang telah dilakukan SMOTE-ENN dan Bayesian Optimation pada model RF, GB, K-NN, DT, ADA, dan *Softvoting* dengan nilai akurasi sebesar 0.987179. Detail dari hasil evaluasi dari eksperimen dengan menggunakan data PIMA dapat dilihat pada tabel 2**,** dan visualisasi dapat dilihat pada gambar 5.

Peningkatan signifikan terlihat pada model yang dioptimalkan dibandingkan dengan versi dasarnya, menunjukkan bahwa teknik pemrosesan fitur dan tuning parameter berperan penting dalam meningkatkan performa. Model dengan akurasi terbaik dapat diandalkan untuk implementasi lebih lanjut, sementara model dengan performa rendah memerlukan analisis tambahan untuk memahami penyebab ketidakefektifannya. Secara keseluruhan, eksperimen ini menegaskan bahwa metode optimasi dan kombinasi model menghasilkan akurasi yang lebih baik dibandingkan model sederhana.

Tabel 2 Evaluasi Model Pada Data PIMA dengan SMOTE-ENN dan Bayesian Optimization

Model	Accuracy	Precision	Recall	f1-score
LR+SMOTE-ENN+BO	0.961538	0.961817	0.961538	0.961519
RF+SMOTE-ENN+BO	**0.987179**	**0.987508**	**0.987179**	**0.987182**
GB+SMOTE-ENN+BO	**0.987179**	**0.987508**	**0.987179**	**0.987182**
K-NN+SMOTE-ENN+BO	**0.987179**	**0.987508**	**0.987179**	**0.987182**
XGB+SMOTE-ENN+BO	0.935897	0.943351	0.935897	0.935739
DT+SMOTE-ENN+BO	**0.987179**	**0.987508**	**0.987179**	**0.987182**
ADA+SMOTE-ENN+BO	**0.987179**	**0.987508**	**0.987179**	**0.987182**
SOFTVOTING	**0.987179**	**0.987508**	**0.987179**	**0.987182**



Gambar 6. Hasil Evaluasi Menggunakan Data PIMA

3.3. Eksperimen Dengan Dataset DPD

Dataset yang kedua yaitu, Diabetes Prediction Dataset (DPD) yang merupakan dataset lebih besar dan komprehensif, sering kali berasal dari sistem klinis atau kesehatan yang mencakup beragam demografi. Meskipun memiliki fitur serupa dengan PIMA, DPD biasanya menyertakan atribut tambahan seperti faktor gaya hidup dan hasil laboratorium, memberikan perspektif yang lebih lengkap untuk prediksi diabetes. Ukurannya yang lebih besar dan

ketidakseimbangan kelas yang lebih kecil membuat dataset ini sangat berguna untuk melatih model yang lebih andal. Namun, kompleksitas dan kemungkinan adanya noise pada data memerlukan pra-pemrosesan dan seleksi fitur yang lebih canggih. Dengan memanfaatkan PIMA dan DPD, penelitian ini mendapatkan manfaat dari kondisi terkendali pada dataset yang lebih kecil dan representasi yang lebih luas dari dataset yang lebih besar, memastikan evaluasi yang komprehensif terhadap model ensemble yang diusulkan. tabel 3 menampilkan deskripsi dari fitur yang digunakan untuk pengolahan data eksperimen.

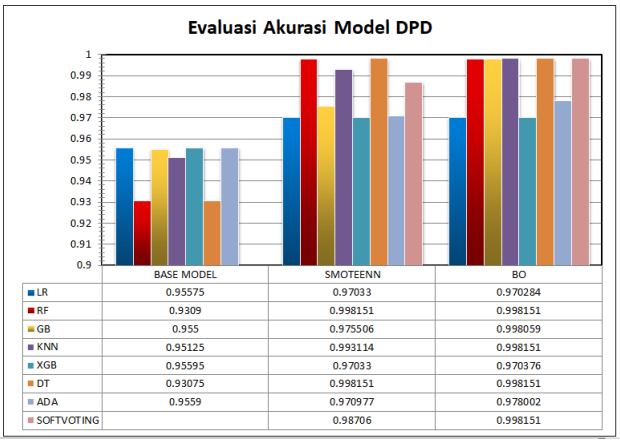
Tabel 3. Deskripsi dataset DPD

Fitur/Atribut	Deskripsi
Gender	Jenis kelamin individu (Male/Female)
Age	Usia individu (dalam tahun)
Hypertension	Apakah individu memiliki hipertensi (tekanan darah tinggi) atau tidak (1 = Ya, 0 = Tidak)
Heart_disease	Apakah individu memiliki riwayat penyakit jantung atau tidak (1 = Ya, 0 = Tidak)
Smoking_history	Riwayat merokok individu
BMI	Indeks Massa Tubuh (Body Mass Index), dihitung sebagai berat badan (kg) dibagi tinggi badan kuadrat (m^2)
HbA1c_level	Level hemoglobin terglikasi (HbA1c) dalam darah, yang menunjukkan kadar gula darah rata-rata dalam beberapa bulan terakhir
Blood_glucose_level	Tingkat gula darah (kemungkinan dalam mg/dL)
Diabetes	Variabel target yang menunjukkan apakah individu menderita diabetes (1 = Ya, 0 = Tidak)

Hasil evaluasi dari eksperimen dengan menggunakan data DPD menunjukan bahwa nilai akurasi yang paling tinggi terdapat pada eksperimen yang telah dilakukan SMOTE-ENN dan Bayesian Optimation pada model Gradian Boosting dengan nilai akurasi sebesar 0.998059. Detail hasil evaluasi dari eksperimen dengan menggunakan data DPD dapat dilihat pada tabel tabel 4, dan visualisasi dapat dilihat pada gambar 7.

Tabel 4. Evaluasi Model Pada Data DPD dengan SMOTE-ENN dan Bayesian Optimization

Model	Accuracy	Precision	Recall	f1-score
LR+SMOTE-ENN+BO	0.970284	0.970291	0.970284	0.970283
RF+SMOTE-ENN+BO	0.998151	0.998153	0.998151	0.998151
GB+SMOTE-ENN+BO	**0.998059**	0.99806	0.998059	0.998059
K-NN+SMOTE-ENN+BO	0.998151	0.998153	0.998151	0.998151
XGB+SMOTE-ENN+BO	0.970376	0.970397	0.970376	0.970375
DT+SMOTE-ENN+BO	0.998151	0.998153	0.998151	0.998151
ADA+SMOTE-ENN+BO	0.978002	0.978091	0.978002	0.978002
SOFTVOTING	0.998151	0.998153	0.998151	0.998151



	BASE MODEL	SMOTEENN	BO
LR	0.95575	0.97033	0.970284
RF	0.9309	0.998151	0.998151
GB	0.955	0.975506	0.998059
KNN	0.95125	0.993114	0.998151
XGB	0.95595	0.97033	0.970376
DT	0.93075	0.998151	0.998151
ADA	0.9559	0.970977	0.978002
SOFTVOTING		0.98706	0.998151

Gambar 8. Hasil Evaluasi Menggunakan Data DPD

Peneliti melakukan perbandingan antara penelitian yang dilakukan dengan penelitian yang sebelumnya telah dilakukan terkait dengan prediksi penyakit diabetes yang disajikan pada tabel 5.

Tabel 5. Perbandingan Penelitian terkait prediksi Penyakit diabetes

Peneliti	Tahun	Metode	Hasil	Dataset
Zhou et al (Gündoğdu, 2023)	2023	stacking ensemble	Akurasi 98% pada dataset PIMA	PIMA
Reza et al (Alzubaidi, Halawani, & Jarrah, 2024)	2024	stacking ensemble	Akurasi 77.10% pada dataset PIMA Akurasi 96.91%, pada dataset klinis lokal pabna	PIMA & klinis lokal pabna
Phongying et al (Smita Panigrahy, 2024)	2023	Random Forest	Akurasi 97.5%	Departemen Layanan Medis Bangkok (2019–2021)
Gündoğdu et al (An, Rahman, Zhou, & Kang, 2023)	2023	MLR dan RF	Akurasi 99.23%	Sylhet Diabetes Hospital di Bangladesh
Panigrahy et al. (Endut et al., 2022)	2024	SMOTE-based Deep LSTM dengan GridSearchCV	Akurasi 97.12%	PIMA
Albadri et al.(Albadri, Awad, Hameed, Mandeel, & Jabbar, 2024)	2024	Hybrid (SVM, DT, RF)	Akurasi 88,5% (holdout) dan 90,1% (k-fold cross-validation)	PIMA
Kumari et al (Sarita Kumari, 2024)	2024	Logistic Regression (LR), K-NN, XGBoost, SVM	LR mencapai AUC 0,95, XGBoost AUC 0,84	Kaggle
Sheta et al. (Sheta, Elashmawi, Al-Qerem, & Othman, 2024)	2024	ANN, LR, SVM, DT	ANN dan LR memiliki akurasi masing-masing 96,1% dan 97%	Data lokal
Yang [59]	2024	Random Forest (RF)	Akurasi 94,5% dengan SMOTE	PIMA
Alzubaidi et al (Reza, Amin, Yasmin, Kulsum, & Ruhi, 2024)	2024	Stacking Ensemble	Akurasi 99% pada dataset PIMA Akurasi 97%, pada dataset DPD	PIMA & DPD
Proposed Model	2025	Softvoting Ensemble	Akurasi 98.72% pada dataset PIMA Akurasi 99.82%, pada dataset DPD	PIMA & DPD

4. Kesimpulan

Penelitian ini mengusulkan pendekatan *ensemble Softvoting* yang dikombinasikan dengan SMOTE-ENN untuk menangani ketidakseimbangan data serta optimasi Bayesian guna meningkatkan akurasi klasifikasi penyakit diabetes. Dengan menggunakan dataset PIMA Indian

Diabetes dan Diabetes Prediction Dataset (DPD), serta berbagai algoritma klasifikasi seperti *Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), k-Nearest Neighbors (K-NN), Extreme Gradient Boosting (XGB), Decision Tree (DT),* dan *AdaBoost (ADA),* metode yang diusulkan berhasil meningkatkan kinerja model dalam mendeteksi diabetes. Hasil eksperimen menunjukkan bahwa metode *Softvoting* dengan SMOTE-ENN dan optimasi Bayesian mampu meningkatkan akurasi dan ketahanan model dalam kondisi data yang tidak seimbang. Pada dataset PIMA Indian Diabetes, nilai terbaik terdapat pada eksperimen yang telah dilakukan SMOTE-ENN dan Bayesian Optimation pada model RF, GB, K-NN, DT, ADA, dan Softvoting dengan nilai akurasi sebesar 0.987179 dengan menggunakan 10-fold cross-validation. Sementara itu, pada dataset DPD, model terbaik memperoleh akurasi 0.998059 pada eksperimen yang telah dilakukan SMOTE-ENN dan Bayesian Optimation pada model *Gradiant Boosting.* Hasil ini menunjukkan bahwa pendekatan yang diusulkan dapat menjadi solusi efektif untuk meningkatkan deteksi dini penyakit diabetes, yang berpotensi membantu tenaga medis dalam mengambil keputusan lebih cepat dan akurat. Penelitian ini juga membuka peluang untuk eksplorasi lebih lanjut dengan metode *deep learning* seperti LSTM, MLP, TabNet, kemudian dari ketiga model tersebut dapat dilakukan ensemble learning, serta penerapan model dalam sistem berbasis IoT atau aplikasi kesehatan berbasis AI guna meningkatkan aksesibilitas dan efisiensi diagnosis penyakit diabetes di dunia nyata.

Acknowledgments

Peneliti mengucapkan terima kasih kepada Universitas Nusa Mandiri yang telah memberikan fasilitas serta dukungan akademik selama proses penelitian berlangsung. Selain itu, Peneliti menghargai kontribusi dan masukan berharga dari tim, dan rekan sejawat yang turut serta dalam diskusi dan pengembangan penelitian ini. Semoga hasil penelitian ini dapat memberikan kontribusi yang bermanfaat bagi pengembangan ilmu pengetahuan.

References

[1] Abdulsadig, R. S., & Rodriguez-Villegas, E. (2024). A comparative study in class imbalance mitigation when working with physiological signals. *Frontiers in Digital Health*, *6*(March), 1–11. https://doi.org/10.3389/fdgth.2024.1377165

[2] Albadri, R. F., Awad, S. M., Hameed, A. S., Mandeel, T. H., & Jabbar, R. A. (2024). A Diabetes Prediction Model Using Hybrid Machine Learning Algorithm. *Mathematical Modelling of Engineering Problems*, *11*(8), 2119–2126. https://doi.org/10.18280/mmep.110813

[3] Alpiansah, A. B., & Ramdhani, Y. (2023). Optimasi Fitur dengan Forward Selection pada Estimasi Tingkat Obesitas menggunakan Random Forest Feature Optimization with Forward Selection on Obesity Rate Estimation using Random Forest. *SISTEMASI: Jurnal Sistem Informasi*, *12*(September), 860–873.

[4] Alzubaidi, A. A., Halawani, S. M., & Jarrah, M. (2024). Integrated Ensemble Model for Diabetes Mellitus Detection. *International Journal of Advanced Computer Science and Applications*, *15*(4), 223–233. https://doi.org/10.14569/IJACSA.2024.0150423

[5] Amou, M. A., Xia, K., Kamhi, S., & Mouhafid, M. (2022). A Novel MRI Diagnosis Method for Brain Tumor Classification Based on CNN and Bayesian Optimization. *Healthcare (Switzerland)*, *10*(3), 1–21. https://doi.org/10.3390/healthcare10030494

[6] An, Q., Rahman, S., Zhou, J., & Kang, J. J. (2023). A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. *Sensors*, *23*(9). https://doi.org/10.3390/s23094178

[7] Anharudin, & Mila Tejamaya. (2022). Perbandingan Risiko Kardiovaskuler Menggunakan Metode Framingham, WHO Chart dan ASCVD Pada Pekerja PT. X Tahun 2021. *Promotif : Jurnal Kesehatan Masyarakat*, *12*(1), 77–84. https://doi.org/10.56338/pjkm.v12i1.2465

[8] Anisa, D. N., & Jumanto, J. (2022). Klasifikasi Penyakit Diabetes Menggunakan Algoritma Naive Bayes. *Jurnal Dinamika Informatika*, *14*(1), 33–42. https://doi.org/10.35315/informatika.v14i1.9135

[9] Das, M. N., Panda, N., & Rautray, R. (2023). Enhancing lung cancer disease diagnosis by employing ensemble deep learning approaches. *Indonesian Journal of Electrical Engineering and Computer Science*, *32*(3), 1766–1773. https://doi.org/10.11591/ijeecs.v32.i3.pp1766-1773

[10] Desiani, A., Kresnawati, E. S., Arhami, M., Resti, Y., Eliyati, N., Yahdin, S., … Nawawi, M. (2023). Majority Voting as Ensemble Classifier for Cervical Cancer Classification. *Science and Technology Indonesia*, *8*(1), 84–92. https://doi.org/10.26554/sti.2023.8.1.84-92

[11] Desiani, A., Primartha, R., Hanum, H., Dewi, S. R. P., Al-Filambany, M. G., Suedarmin, M., & Suprihatin, B. (2024). Weighted Voting Ensemble Learning of CNN Architectures for Diabetic Retinopathy Classification. *Jurnal Infotel*, *16*(1), 136–155. https://doi.org/10.20895/infotel.v16i1.999

[12] Dewi Nasien, Darwin, R., Cia, A., Leo Winata, A., Go, J., M.C, R., … Charles Lo, K. (2024). Perbandingan Implementasi Machine Learning Menggunakan Metode KNN, Naive Bayes, dan Logistik Regression Untuk Mengklasifikasi Penyakit Diabetes. *JEKIN - Jurnal Teknik Informatika*, *4*(1), 10–17. https://doi.org/10.58794/jekin.v4i1.640

[13] Endut, N., Hamzah, W. M. A. F. W., Ismail, I., Yusof, M. K., Baker, Y. A., & Yusoff, H. (2022). A Systematic Literature Review on Multi-Label Classification based on Machine Learning Algorithms. *TEM Journal*, *11*(2), 658–666. https://doi.org/10.18421/TEM112-20

[14] Ernawati, S., & Wati, R. (2024). Evaluasi Performa Kernel SVM dalam Analisis Sentimen Review Aplikasi ChatGPT Menggunakan Hyperparameter dan VADER

Lexicon. *Jurnal Buana Informatika*, *15*(01), 40–49. https://doi.org/10.24002/jbi.v15i1.7925

[15] Ferdina, F., Satyahadewi, N., & Kusnandar, D. (2023). Penerapan Algoritma Iterative Dichotomiser 3 (Id3) Dalam Klasifikasi Faktor Risiko Penyakit Diabetes Melitus. *VARIANCE: Journal of Statistics and Its Applications*, *5*(2), 139–146. https://doi.org/10.30598/variancevol5iss2page139-146

[16] Gündoğdu, S. (2023). Efficient prediction of early-stage diabetes using XGBoost classifier with random forest feature selection technique. *Multimedia Tools and Applications*, *82*(22), 34163–34181. https://doi.org/10.1007/s11042-023-15165-8

[17] Hatmanti, N. M., Winoto, P. M. P., Salamy, N. F. W., Rusdianingseh, R., Septianingrum, Y., Maimunah, S., & Wardani, E. M. (2023). Pemberdayaan Kader Kesehatan dalam Penatalaksanaan Penyakit Diabetes Mellitus. *Jurnal ABDINUS : Jurnal Pengabdian Nusantara*, *7*(3), 830–838. https://doi.org/10.29407/ja.v7i3.20160

[18] Helmut, A., & Murdiansyah, D. T. (2023). Multiclass Email Classification by Using Ensemble Bagging and Ensemble Voting. *JIKO (Jurnal Informatika Dan Komputer)*, *6*(2), 144–149. https://doi.org/10.33387/jiko.v6i2.6394

[19] Izraiq, M., Almousa, E., Hammoudeh, S., Sudqi, M., Ahmed, Y., Abu-Dhaim, O., … Abu-Hantash, H. (2024). Impact of Diabetes Mellitus on Heart Failure Patients: Insights from a Comprehensive Analysis and Machine Learning Model Using the Jordanian Heart Failure Registry [Corrigendum]. *International Journal of General Medicine*, *Volume 17*(May), 3371–3372. https://doi.org/10.2147/ijgm.s487404

[20] Lukitaningtyas, D., Kurniasih, E., & Pariyem, P. (2022). Deteksi Dini dan Monitoring Penyakit Degeneratif Diabetes Melitus di Dusun Pilangpayung I , Desa Geneng Kec. Geneng Kabupaten Ngawi. *Jurnal Kreativitas Pengabdian Kepada Masyarakat (PKM)*, *5*(12), 4551–4557. https://doi.org/10.33024/jkpm.v5i12.8148

[21] Nistiandani, A., Siswoyo, S., Sakinah, E. N., Rahmah, T. A., & Lisar, A. F. (2024). Edukasi Kader Kesehatan Tentang Deteksi Resiko Luka Kaki Pada Pasien Diabetes Mellitus Di Desa Mayang. *Abdimasku : Jurnal Pengabdian Masyarakat*, *7*(1), 381. https://doi.org/10.62411/ja.v7i1.1867

[22] Nuanmeesri, S., & Poomhiran, L. (2023). Improving the Avoidant Personality Disorder Prediction for Higher Education Using SMOTE-ENN and Multi-Layer Perceptron Neural Network. *TEM Journal*, *12*(2), 1008–1022. https://doi.org/10.18421/TEM122-47

[23] Oei, C. W., Ng, E. Y. K., Ng, M. H. S., Tan, R. S., Chan, Y. M., Chan, L. G., & Acharya, U. R. (2023). Explainable Risk Prediction of Post-Stroke Adverse Mental Outcomes Using Machine Learning Techniques in a Population of 1780 Patients. *Sensors*, *23*(18), 1–12. https://doi.org/10.3390/s23187946

[24] Prabha, R. M., & Sasikala, S. (2024). Data Analytics for Imbalanced Dataset. *Journal of Computer Science*, *20*(2), 207–217. https://doi.org/10.3844/jcssp.2024.207.217

[25] Prasetyo, S. Y., & Nabiilah, G. Z. (2023). Perbandingan Model Machine Learning pada Klasifikasi Tumor Otak Menggunakan Fitur Discrete Cosine Transform. *Jurnal Teknologi Terpadu*,

9(1), 29–34. https://doi.org/10.54914/jtt.v9i1.605

[26] Reza, M. S., Amin, R., Yasmin, R., Kulsum, W., & Ruhi, S. (2024). Improving diabetes disease patients classification using stacking ensemble method with PIMA and local healthcare data. *Heliyon*, *10*(2), e24536. https://doi.org/10.1016/j.heliyon.2024.e24536

[27] Sarita Kumari. (2024). Investigating Role of Supervised Machine Learning Approach in Classification of Diabetic Patient. *Journal of Electrical Systems*, *20*(4s), 2539–2556. https://doi.org/10.52783/jes.2987

[28] Sheta, A., Elashmawi, W. H., Al-Qerem, A., & Othman, E. S. (2024). Utilizing Various Machine Learning Techniques for Diabetes Mellitus Feature Selection and Classification. *International Journal of Advanced Computer Science and Applications*, *15*(3), 1372–1384. https://doi.org/10.14569/IJACSA.2024.01503134

[29] Smita Panigrahy. (2024). SMOTE-based Deep LSTM System with GridSearchCV Optimization for Intelligent Diabetes Diagnosis. *Journal of Electrical Systems*, *20*(7s), 804–815. https://doi.org/10.52783/jes.3455

[30] Ullah, Z., Saleem, F., Jamjoom, M., Fakieh, B., Kateb, F., Ali, A. M., & Shah, B. (2022). Detecting High-Risk Factors and Early Diagnosis of Diabetes Using Machine Learning Methods. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/2557795

[31] Xu, Y., Ye, W., Song, Q., Shen, L., Liu, Y., Guo, Y., … Xu, Z. (2022). Using machine learning models to predict the duration of the recovery of COVID-19 patients hospitalized in Fangcang shelter hospital during the Omicron BA. 2.2 pandemic. *Frontiers in Medicine*, *9*. https://doi.org/10.3389/fmed.2022.1001801

[32] Yang, F., Wang, K., Sun, L., Zhai, M., Song, J., & Wang, H. (2022). A hybrid sampling algorithm combining synthetic minority over-sampling technique and edited nearest neighbor for missed abortion diagnosis. *BMC Medical Informatics and Decision Making*, *22*(1), 1–14. https://doi.org/10.1186/s12911-022-02075-2

[33] Ye, F., Li, X., Zhang, N., & Xu, F. (2024). Prediction of Single-Well Production Rate after Hydraulic Fracturing in Unconventional Gas Reservoirs Based on Ensemble Learning Model. *Processes*, *12*(6). https://doi.org/10.3390/pr12061194

[34] Yusa, M., Coastera, F. F., & Yandika, M. R. (2022). Reduksi Dimensi Data menggunakan Metode Wrapper Sequential Feature Selection untuk Peningkatan Performa Algoritma Naïve Bayes terhadap Dataset Medis. *Jurnal Edukasi Dan Penelitian Informatika (JEPIN)*, *8*(2), 364. https://doi.org/10.26418/jp.v8i2.54328

[35]