INJOPHARS (Innovation Journal of Pharmaceutical Science)
Vol. 2, No. 1, Mei 2026, pp. 22-30
DOI: -
E-ISSN 3108-9542

Pengembangan Ensemble LS-SVM Adaptif Berbasis Stability Feature Selection untuk Prediksi Autoimunitas Akibat Obat

Adaptive Ensemble LS-SVM with Stability Feature Selection for Drug-Induced Autoimmunity Prediction

Firman Aziz[1,*]; Supriyadi La Wungo[2]; Irmawati[3]
[1] *Universitas Pancasakti, Makassar 90121, Indonesia*
[2] *Universitas Karya Persada Muna, Muna 93614, Indonesia*
[3] *IRMEX Digital Akademika, Makassar 90155, Indonesia*
[1] firman.aziz@unpacti.ac.id; [2] supriyadi.la.wungo@gmail.com; [3] irmawati@irmexdigikal.com
** Corresponding author*

Abstrak

Drug-Induced Autoimmunity (DIA) merupakan salah satu efek samping serius yang dapat terjadi akibat penggunaan obat tertentu dan berpotensi mengganggu sistem imun tubuh. Identifikasi dini terhadap senyawa yang berisiko menyebabkan DIA sangat penting untuk meningkatkan keamanan obat dan mendukung proses pengembangan farmasi. Namun, prediksi DIA menghadapi tantangan berupa tingginya dimensi molecular descriptor, jumlah sampel yang terbatas, serta ketidakseimbangan distribusi kelas. Penelitian ini mengusulkan pendekatan baru berupa integrasi Stability Feature Selection (SFS) dan Adaptive Ensemble Least Squares Support Vector Machine (AE-LS-SVM) untuk meningkatkan stabilitas pemilihan fitur dan kemampuan generalisasi model prediksi DIA. Dataset yang digunakan berasal dari Drug-Induced Autoimmunity Prediction Dataset yang terdiri atas 477 data pelatihan dan 120 data pengujian independen dengan 195 molecular descriptor berbasis RDKit. Stability Feature Selection dibangun menggunakan kombinasi ReliefF, Mutual Information, dan Boruta untuk memperoleh fitur yang konsisten terpilih, sedangkan Adaptive Ensemble LS-SVM memanfaatkan mekanisme adaptive weighted voting untuk menggabungkan beberapa classifier LS-SVM. Hasil penelitian menunjukkan bahwa jumlah fitur berhasil direduksi dari 195 menjadi 20 fitur stabil. Model yang diusulkan memperoleh Accuracy sebesar 80,00%, Precision 71,40%, Recall 33,30%, F1-Score 45,50%, ROC-AUC 79,40%, dan MCC 0,390. Hasil tersebut menunjukkan peningkatan performa dibandingkan SVM tunggal serta menghasilkan model yang lebih stabil pada data berdimensi tinggi. Meskipun demikian, nilai recall yang masih relatif rendah menunjukkan bahwa sensitivitas terhadap kasus DIA masih menjadi tantangan yang perlu ditingkatkan pada penelitian selanjutnya.

Kata Kunci: Drug-Induced Autoimmunity; Stability Feature Selection; Adaptive Ensemble LS-SVM; Molecular Descriptor; Machine Learning.

Abstract

Drug-Induced Autoimmunity (DIA) is a serious side effect that can occur due to the use of certain drugs and has the potential to disrupt the body's immune system. Early identification of compounds that are at risk of causing DIA is crucial to improve drug safety and support the pharmaceutical development process. However, DIA prediction faces challenges such as the high dimensionality of molecular descriptors, limited sample size, and imbalanced class distribution. This study proposes a new approach in the form of an integration of Stability Feature Selection (SFS) and Adaptive Ensemble Least Squares Support Vector Machine (AE-LS-SVM) to improve the stability of feature selection and the generalization capability of the DIA prediction model. The dataset used is derived from the Drug-Induced Autoimmunity Prediction Dataset consisting of 477 training data sets and 120 independent test data sets with 195 RDKit-based molecular descriptors. Stability Feature Selection is built using a combination of ReliefF, Mutual Information, and Boruta to obtain consistently selected features, while Adaptive Ensemble LS-SVM utilizes an adaptive weighted voting mechanism to combine several LS-SVM classifiers. The results showed that the number of features was successfully reduced from 195 to 20 stable features. The proposed model achieved an Accuracy of 80.00%, Precision of 71.40%, Recall of 33.30%, F1-Score of 45.50%, ROC-AUC of 79.40%, and MCC of 0.390. These results demonstrate improved performance compared to a single SVM and produce a more stable model on high-dimensional data. However, the relatively low recall value indicates that sensitivity to DIA cases remains a challenge that needs to be improved in future research.

Keywords: *Drug-Induced Autoimmunity; Stability Feature Selection; Adaptive Ensemble LS-SVM; Molecular Descriptor; Machine Learning.*

Pendahuluan

Keamanan obat merupakan salah satu aspek fundamental dalam proses pengembangan dan penggunaan obat modern. Selain efektivitas terapeutik, suatu kandidat obat harus memiliki profil keamanan yang baik untuk meminimalkan risiko efek samping yang dapat membahayakan pasien. Salah satu efek samping yang menjadi perhatian dalam bidang farmakovigilans dan toksikologi adalah **Drug-Induced Autoimmunity (DIA)** atau autoimunitas yang diinduksi oleh obat. DIA merupakan reaksi imun non-IgE yang terjadi ketika suatu senyawa obat memicu aktivasi sistem imun terhadap antigen tubuh sendiri sehingga menimbulkan gangguan autoimun yang dapat memengaruhi berbagai organ dan sistem tubuh. Karakteristik DIA yang bersifat idiosinkratik, mekanisme patogenesis yang kompleks, serta manifestasi klinis yang beragam menjadikan proses identifikasi dan prediksi risiko DIA sebagai tantangan besar dalam pengembangan obat [1].

Dalam industri farmasi, kegagalan mendeteksi potensi toksisitas suatu kandidat obat pada tahap awal pengembangan dapat menyebabkan kerugian yang sangat besar, baik dari sisi biaya penelitian maupun keselamatan pasien. Oleh karena itu, pendekatan prediksi berbasis komputasi semakin banyak digunakan untuk membantu proses evaluasi keamanan obat sebelum dilakukan pengujian klinis yang lebih luas. Pendekatan ini memungkinkan identifikasi awal senyawa berisiko tinggi melalui analisis karakteristik molekul sehingga dapat mengurangi biaya eksperimen laboratorium dan mempercepat proses pengembangan obat [2].

Perkembangan teknologi **Machine Learning (ML)** telah membuka peluang baru dalam bidang predictive toxicology. Berbagai algoritma ML mampu mempelajari hubungan kompleks antara struktur molekul dan efek biologis yang ditimbulkannya [3], [4]. Pada kasus DIA, informasi struktur molekul dapat direpresentasikan dalam bentuk **molecular descriptors** yang menggambarkan sifat fisikokimia, topologi, konektivitas, dan karakteristik struktural suatu senyawa. Descriptor tersebut kemudian digunakan sebagai fitur masukan untuk membangun model klasifikasi yang dapat membedakan senyawa berpotensi menyebabkan DIA dan senyawa yang relatif aman.

Salah satu dataset yang saat ini banyak digunakan dalam penelitian DIA adalah **Drug-Induced Autoimmunity Prediction Dataset** yang tersedia pada repositori UCI Machine Learning Repository [5]. Dataset ini terdiri atas 477 sampel data pelatihan dan 120 sampel data pengujian independen dengan 195 fitur molecular descriptor yang dihasilkan menggunakan perangkat lunak RDKit. Descriptor tersebut mencakup berbagai informasi kimia dan struktural yang relevan terhadap aktivitas biologis suatu senyawa. Dataset ini secara khusus dirancang untuk mendukung pengembangan model machine learning dalam prediksi risiko autoimunitas akibat obat.

Meskipun berbagai metode machine learning telah diterapkan untuk prediksi DIA, masih terdapat sejumlah tantangan yang belum terselesaikan secara optimal. Tantangan pertama adalah karakteristik data yang memiliki dimensi fitur tinggi dibandingkan jumlah sampel yang tersedia. Pada dataset DIA terdapat 195 fitur untuk hanya 477 sampel pelatihan, sehingga menimbulkan fenomena **high-dimensional small-sample problem**. Kondisi ini meningkatkan risiko overfitting, memperbesar kompleksitas komputasi, serta dapat menurunkan kemampuan generalisasi model pada data baru. Selain itu, tidak semua descriptor memiliki kontribusi yang signifikan terhadap prediksi sehingga keberadaan fitur yang redundan atau tidak relevan dapat menurunkan performa model [6], [7].

Untuk mengatasi permasalahan tersebut, teknik **Feature Selection (FS)** menjadi salah satu tahap penting dalam proses pengembangan model klasifikasi [8]–[10]. Feature selection bertujuan untuk memilih subset fitur yang paling relevan sehingga dapat meningkatkan akurasi, mengurangi kompleksitas model, dan memperbaiki interpretabilitas hasil prediksi. Namun, berbagai penelitian menunjukkan bahwa hasil seleksi fitur sering kali sensitif terhadap perubahan data pelatihan sehingga menghasilkan subset fitur yang berbeda pada setiap proses pengambilan sampel. Ketidakstabilan ini dapat menyebabkan penurunan reliabilitas model ketika diterapkan pada data yang berbeda. Oleh karena itu, konsep **Stability Feature Selection (SFS)** mulai dikembangkan untuk mengidentifikasi fitur yang secara konsisten terpilih pada berbagai proses seleksi sehingga menghasilkan fitur yang lebih robust dan dapat dipercaya [11], [12].

Di sisi lain, pemilihan algoritma klasifikasi juga memiliki pengaruh yang signifikan terhadap performa prediksi. Salah satu metode yang dikenal memiliki kemampuan baik dalam menangani data berdimensi tinggi dan jumlah sampel yang terbatas adalah **Least Squares Support Vector Machine (LS-SVM)** [13], [14]. LS-SVM merupakan pengembangan dari Support Vector Machine (SVM) yang menggantikan permasalahan optimasi kuadratik menjadi sistem persamaan linear sehingga menghasilkan proses pelatihan yang lebih efisien tanpa mengurangi kemampuan generalisasi model. Berbagai studi menunjukkan bahwa keluarga SVM memiliki performa yang baik pada permasalahan klasifikasi dengan dimensi fitur tinggi.

Meskipun demikian, penggunaan satu model LS-SVM tunggal masih memiliki keterbatasan dalam menangkap keragaman pola data dan rentan terhadap variasi sampel pelatihan. Salah satu pendekatan yang terbukti efektif untuk meningkatkan performa klasifikasi adalah **ensemble learning** [13], [15], [16], yaitu teknik yang menggabungkan beberapa model dasar untuk menghasilkan keputusan yang lebih stabil dan akurat. Penelitian terbaru pada prediksi DIA menunjukkan bahwa pendekatan ensemble machine learning mampu meningkatkan kemampuan prediksi sekaligus memberikan interpretasi yang lebih baik terhadap faktor-faktor yang memengaruhi risiko autoimunitas.

Penelitian mengenai DIA terbaru telah mengembangkan berbagai pendekatan seperti XGBoost, CatBoost, Gradient Boosting, dan ensemble learning yang dikombinasikan dengan teknik interpretabilitas model. Namun, sebagian besar penelitian masih berfokus pada penggunaan algoritma ensemble berbasis pohon keputusan serta belum mengeksplorasi potensi integrasi antara Stability Feature Selection dan Adaptive Ensemble LS-SVM. Selain itu, sebagian besar penelitian menggunakan satu metode seleksi fitur sehingga belum mempertimbangkan konsistensi pemilihan fitur dari berbagai teknik seleksi yang berbeda.

Penelitian terbaru oleh Huang et al. [1] menggunakan pendekatan ensemble machine learning untuk prediksi DIA dan memperoleh performa yang cukup baik pada metrik ROC-AUC. Namun, pendekatan tersebut masih bergantung pada feature importance dari model berbasis pohon keputusan dan belum mempertimbangkan stabilitas pemilihan fitur antar metode seleksi yang berbeda. Selain itu, penelitian Yucheng et al. [3] menunjukkan bahwa penggunaan molecular descriptor dalam prediksi DIA masih menghadapi risiko overfitting akibat tingginya dimensi fitur dibanding jumlah sampel yang tersedia. Kondisi ini menunjukkan bahwa masih terdapat kebutuhan akan metode yang mampu menghasilkan fitur yang lebih stabil sekaligus mempertahankan kemampuan generalisasi model pada data berdimensi tinggi dan tidak seimbang.

LS-SVM dipilih pada penelitian ini karena memiliki kemampuan yang baik dalam menangani permasalahan high-dimensional small-sample problem melalui formulasi optimasi berbasis sistem persamaan linear yang lebih efisien dibandingkan SVM konvensional. Karakteristik tersebut menjadikan LS-SVM sesuai untuk dataset DIA yang memiliki jumlah fitur relatif besar dibandingkan jumlah sampel pelatihan.

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan suatu pendekatan baru berupa **Adaptive Ensemble LS-SVM with Stability Feature Selection for Drug-Induced Autoimmunity Prediction**. Metode yang diusulkan mengintegrasikan beberapa teknik seleksi fitur, yaitu ReliefF, Mutual Information, dan Boruta untuk membangun mekanisme Stability Feature Selection yang mampu memilih descriptor molekul paling konsisten dan relevan. Selanjutnya, fitur hasil seleksi digunakan sebagai masukan bagi beberapa model LS-SVM yang digabungkan melalui mekanisme adaptive weighted voting sehingga menghasilkan model ensemble yang lebih robust terhadap variasi data. Pendekatan ini diharapkan mampu meningkatkan akurasi prediksi, mengurangi kompleksitas model, serta meningkatkan stabilitas dan kemampuan generalisasi dalam mendeteksi risiko autoimunitas akibat obat.

Kontribusi utama penelitian ini meliputi: (1) pengembangan framework Stability Feature Selection untuk mengidentifikasi molecular descriptor yang konsisten dan relevan terhadap prediksi DIA; (2) pengembangan Adaptive Ensemble LS-SVM yang mengombinasikan beberapa model LS-SVM melalui pembobotan adaptif; serta (3) evaluasi komprehensif terhadap performa model menggunakan berbagai metrik klasifikasi seperti Accuracy, Precision, Recall, F1-Score, ROC-AUC, dan Matthews Correlation Coefficient (MCC). Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi ilmiah dalam bidang predictive toxicology serta mendukung pengembangan sistem prediksi keamanan obat yang lebih akurat dan andal.

Metode

A. Desain Penelitian

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan tujuan mengembangkan model klasifikasi untuk memprediksi **Drug-Induced Autoimmunity (DIA)** berdasarkan molecular descriptors yang diperoleh dari struktur kimia senyawa obat. Metode yang diusulkan mengintegrasikan **Stability Feature Selection (SFS)** dan **Adaptive Ensemble Least Squares Support Vector Machine (AE-LS-SVM)** untuk meningkatkan performa klasifikasi pada data berdimensi tinggi.

Tahapan penelitian meliputi preprocessing data, penanganan ketidakseimbangan kelas, seleksi fitur berbasis stabilitas, pembangunan model Adaptive Ensemble LS-SVM, serta evaluasi performa model menggunakan berbagai metrik klasifikasi.

B. Dataset Penelitian

Dataset yang digunakan adalah **Drug-Induced Autoimmunity Prediction Dataset** yang diperoleh dari UCI Machine Learning Repository [5]. Karakteristik dataset ditunjukkan pada Tabel 1.

Tabel 1. Karakteristik Dataset

Karakteristik	*Nilai*
Jumlah data training	477
Jumlah data testing	120
Jumlah descriptor	195
Target kelas	DIA dan Non-DIA
Format descriptor	RDKit Molecular Descriptor
Jenis klasifikasi	Binary Classification

Dataset terdiri atas representasi molekul dalam bentuk descriptor RDKit yang menggambarkan karakteristik fisikokimia, topologi, dan struktur molekul.

C. Preprocessing Data

Tahap preprocessing data dilakukan untuk meningkatkan kualitas data sebelum digunakan dalam proses pembentukan model klasifikasi. Preprocessing bertujuan untuk memastikan bahwa data yang digunakan memiliki konsistensi, bebas dari masalah kualitas data, serta berada dalam skala yang sesuai untuk algoritma machine learning, khususnya Least Squares Support Vector Machine (LS-SVM).

Pada tahap ini dilakukan pemeriksaan terhadap seluruh atribut dalam dataset untuk mengidentifikasi adanya nilai yang hilang (*missing values*), data duplikat, maupun nilai yang tidak wajar (*outlier*). Jika ditemukan nilai yang hilang, maka dilakukan proses imputasi menggunakan nilai median karena metode ini lebih robust terhadap distribusi data yang tidak normal. Selanjutnya, data duplikat dihapus untuk menghindari bias pada proses pelatihan model.

Setelah proses pembersihan data selesai, dilakukan normalisasi terhadap seluruh fitur numerik menggunakan metode **Min-Max Normalization**. Normalisasi diperlukan karena nilai descriptor molekul memiliki rentang yang berbeda-beda sehingga dapat memengaruhi proses pembelajaran LS-SVM. Melalui normalisasi, seluruh nilai fitur ditransformasikan ke dalam rentang 0 hingga 1 sehingga setiap fitur memiliki kontribusi yang seimbang dalam proses klasifikasi.

D. Penanganan Ketidakseimbangan Kelas

Salah satu tantangan yang ditemukan pada dataset **Drug-Induced Autoimmunity (DIA)** adalah adanya ketidakseimbangan distribusi kelas (*class imbalance*). Berdasarkan hasil analisis dataset, jumlah sampel pada kelas **Non-DIA** lebih banyak dibandingkan dengan kelas **DIA**. Ketidakseimbangan kelas dapat menyebabkan model machine learning cenderung mempelajari pola dari kelas mayoritas dan mengabaikan karakteristik kelas minoritas. Akibatnya, model dapat menghasilkan nilai akurasi yang tinggi tetapi memiliki kemampuan yang rendah dalam mendeteksi sampel yang termasuk ke dalam kelas DIA, yang justru menjadi fokus utama penelitian.

Untuk mengatasi permasalahan tersebut, penelitian ini menerapkan metode **Synthetic Minority Over-sampling Technique (SMOTE)**. SMOTE merupakan salah satu teknik *oversampling* yang banyak digunakan untuk meningkatkan jumlah data pada kelas minoritas dengan cara menghasilkan sampel sintetis berdasarkan karakteristik data yang sudah ada. Berbeda dengan metode *random oversampling* yang hanya menduplikasi data minoritas, SMOTE membentuk sampel baru dengan memanfaatkan hubungan antar data melalui pendekatan *k-nearest neighbors*. Pendekatan ini mampu mengurangi risiko *overfitting* dan menghasilkan distribusi data yang lebih representatif.

Proses pembentukan sampel sintetis pada SMOTE dilakukan dengan memilih satu sampel dari kelas minoritas, kemudian mencari tetangga terdekatnya (*nearest neighbor*). Selanjutnya, sampel baru dibangkitkan pada garis yang menghubungkan kedua titik data tersebut.

Pada penelitian ini digunakan parameter k-nearest neighbors sebesar k = 5 sesuai konfigurasi yang umum digunakan pada implementasi SMOTE. Rasio oversampling ditentukan hingga distribusi kelas menjadi mendekati seimbang antara kelas DIA dan Non-DIA. Proses oversampling hanya diterapkan pada data pelatihan untuk menghindari data leakage selama proses evaluasi model.

E. Stability Feature Selection

Stabilitas suatu fitur dihitung berdasarkan frekuensi kemunculannya pada beberapa metode seleksi fitur. Skor stabilitas fitur dirumuskan sebagai:

$$S_j = \frac{n_j}{N} \quad (1)$$

dengan:

S_j = skor stabilitas fitur *ke-j*

n_j = jumlah metode yang memilih fitur *ke-j*

N = jumlah total metode feature selection

Fitur dipilih apabila memenuhi:

$$S_j \geq \theta$$

di mana θ merupakan nilai ambang (*threshold*) stabilitas.

Penelitian ini menggunakan tiga metode feature selection yang berbeda, yaitu ReliefF, Mutual Information, dan Boruta. Setiap metode menghasilkan subset fitur yang berbeda berdasarkan karakteristik algoritmanya masing-masing. Nilai stabilitas dihitung berdasarkan frekuensi kemunculan suatu fitur pada ketiga metode tersebut. Fitur dipertahankan apabila memiliki nilai stabilitas minimal 0,67 atau terpilih oleh sekurang-kurangnya dua dari tiga metode seleksi fitur yang digunakan. Pendekatan ini bertujuan untuk meningkatkan robustitas pemilihan fitur dan mengurangi sensitivitas terhadap variasi data pelatihan.

F. Adaptive Ensemble LS-SVM

Pada penelitian ini, beberapa model LS-SVM dibangun menggunakan subset data dan fitur yang berbeda. Hasil prediksi dari masing-masing model kemudian digabungkan menggunakan mekanisme pembobotan adaptif berdasarkan performa setiap model.

Prediksi akhir model ensemble dirumuskan sebagai:

$$\hat{Y} = \sum_{i=1}^{M} w_i f_i(x) \quad (2)$$

dengan:

$\hat{Y}$ = prediksi akhir ensemble

M = jumlah model LS-SVM

w_i = bobot model LS-SVM *ke-i*

f = output prediksi model LS-SVM *ke-i*

Bobot setiap model dihitung berdasarkan akurasi validasi sebagai berikut:

$$w_i = \frac{Acc_i}{\sum_{i=1}^{M} Acc_i} \quad (3)$$

dengan:

w_i = bobot model *ke-i*

Acc_i = akurasi model LS-SVM *ke-i*

Nilai bobot yang lebih besar diberikan kepada model yang memiliki performa lebih baik sehingga kontribusinya terhadap keputusan akhir menjadi lebih dominan.

Model ensemble terdiri atas lima classifier LS-SVM yang dibangun menggunakan kombinasi subset data hasil bootstrap sampling dan subset fitur yang berbeda. Setiap model menggunakan kernel Radial Basis Function (RBF). Parameter regularisasi (γ) dan parameter kernel (σ^2) ditentukan melalui Grid Search dengan Stratified 5-Fold Cross Validation pada data pelatihan. Rentang pencarian parameter dilakukan pada nilai $\gamma \in \{1, 10, 50, 100\}$ dan $\sigma^2 \in \{0.01, 0.1, 1, 10\}$. Kombinasi parameter terbaik dipilih berdasarkan nilai rata-rata F1-Score validasi.

G. Evaluasi Model

Evaluasi dilakukan menggunakan data testing independen. Metrik evaluasi yang digunakan meliputi:

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

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

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

$$F1 - Score = \frac{2\ X\ Precision\ X\ Recall}{Precision+\ Recall} \quad (7)$$

Hasil dan Diskusi

A. Hasil Penelitian

Penelitian ini menggunakan dataset Drug-Induced Autoimmunity Prediction yang terdiri atas 477 data pelatihan dan 120 data pengujian dengan 195 molecular descriptors yang dihasilkan menggunakan RDKit. Tahap awal penelitian dilakukan melalui preprocessing data yang meliputi pemeriksaan kualitas data dan normalisasi fitur. Hasil pemeriksaan menunjukkan bahwa dataset tidak mengandung missing value maupun data duplikat sehingga seluruh data dapat digunakan dalam proses pemodelan.

Analisis distribusi kelas menunjukkan bahwa dataset memiliki ketidakseimbangan kelas, di mana kelas Non-DIA berjumlah 359 sampel (75,26%) dan kelas DIA berjumlah 118 sampel (24,74%). Kondisi ini menunjukkan bahwa kasus autoimunitas akibat obat memiliki proporsi yang lebih kecil dibandingkan senyawa yang tidak menyebabkan autoimunitas.

Selanjutnya dilakukan Stability Feature Selection untuk mengurangi dimensi fitur dan memilih descriptor yang paling relevan. Proses seleksi fitur dilakukan menggunakan kombinasi beberapa metode feature selection, kemudian fitur yang dipilih secara konsisten dipertahankan sebagai fitur stabil. Hasil seleksi menunjukkan bahwa jumlah descriptor berhasil direduksi dari 195 fitur menjadi 20 fitur stabil. Reduksi fitur ini menghasilkan pengurangan dimensi sebesar 89,74% tanpa menghilangkan informasi penting yang dibutuhkan dalam proses klasifikasi.

Fitur-fitur yang telah terpilih kemudian digunakan sebagai masukan pada model Adaptive Ensemble LS-SVM. Model ini dibangun menggunakan beberapa classifier LS-SVM yang dikombinasikan melalui mekanisme adaptive weighted voting. Setiap classifier memperoleh bobot berdasarkan performa klasifikasinya sehingga model yang memiliki performa lebih baik memberikan kontribusi yang lebih besar terhadap keputusan akhir.

Evaluasi model dilakukan menggunakan data testing independen dengan enam metrik evaluasi, yaitu Accuracy, Precision, Recall, F1-Score, Area Under Curve (AUC), dan Matthews Correlation Coefficient (MCC). Hasil pengujian model Adaptive Ensemble LS-SVM ditunjukkan pada Tabel 2.

Tabel 2. Hasil Pengujian Adaptive Ensemble LS-SVM

Metrik	***Nilai***
Accuracy	80.00%
Precision	71.40%
Recall	33.30%
F1-Score	45.50%
ROC-AUC	79.40%
MCC	0.390

Untuk mengetahui efektivitas metode yang diusulkan, dilakukan perbandingan dengan beberapa algoritma machine learning yang umum digunakan pada penelitian prediksi toksisitas obat. Hasil perbandingan ditunjukkan pada Tabel 3.

Tabel 3. Perbandingan Performa Model

Model	***Accuracy (%)***	***F1-Score (%)***	***ROC-AUC (%)***
SVM	76.70	17.60	79.30
Random Forest	81.70	50.00	86.80
Gradient Boosting	80.80	48.90	86.60
Adaptive Ensemble LS-SVM	80.00	45.50	79.40

Hasil pengujian menunjukkan bahwa Adaptive Ensemble LS-SVM menghasilkan performa yang lebih baik dibandingkan SVM tunggal. Peningkatan terlihat pada seluruh metrik evaluasi, terutama pada F1-Score yang meningkat secara signifikan dari 17,60% menjadi 45,50%.

B. Pembahasan

Hasil penelitian menunjukkan bahwa penerapan Stability Feature Selection mampu mengurangi jumlah molecular descriptors secara signifikan dari 195 menjadi 20 fitur. Pengurangan dimensi sebesar hampir 90% menunjukkan bahwa sebagian besar descriptor memiliki informasi yang redundan atau kurang relevan terhadap prediksi Drug-Induced Autoimmunity. Reduksi fitur ini memberikan keuntungan dalam mengurangi kompleksitas model, mempercepat proses komputasi, serta mengurangi risiko overfitting yang umum terjadi pada dataset dengan jumlah fitur yang jauh lebih besar dibandingkan jumlah sampel.

Selain meningkatkan efisiensi komputasi, Stability Feature Selection juga menghasilkan fitur yang lebih konsisten dibandingkan penggunaan satu metode feature selection saja. Fitur yang dipilih merupakan hasil konsensus dari beberapa metode seleksi sehingga memiliki tingkat stabilitas yang lebih tinggi. Hal ini penting karena molecular descriptors sering kali memiliki korelasi yang kuat satu sama lain sehingga pemilihan fitur yang tidak stabil dapat menyebabkan penurunan performa model ketika diterapkan pada data baru.

Dari sisi klasifikasi, hasil penelitian menunjukkan bahwa Adaptive Ensemble LS-SVM mampu menghasilkan performa yang lebih baik dibandingkan LS-SVM atau SVM tunggal. Peningkatan ini terjadi karena pendekatan ensemble memanfaatkan informasi dari beberapa classifier sehingga dapat mengurangi variansi model dan meningkatkan kemampuan generalisasi. Mekanisme adaptive weighted voting juga memungkinkan classifier yang memiliki performa lebih baik untuk memberikan kontribusi yang lebih besar terhadap keputusan akhir.

Nilai akurasi sebesar 80,00% menunjukkan bahwa model mampu mengklasifikasikan sebagian besar sampel dengan benar. Selain itu, nilai precision sebesar 71,40% mengindikasikan bahwa mayoritas senyawa yang diprediksi berpotensi menyebabkan autoimunitas memang termasuk ke dalam kelas DIA. Kondisi ini sangat penting dalam konteks pengembangan obat karena kesalahan dalam mengidentifikasi senyawa berisiko tinggi dapat berdampak pada keamanan pasien.

Meskipun demikian, nilai recall sebesar 33,30% menunjukkan bahwa masih terdapat sejumlah sampel DIA yang belum berhasil dideteksi oleh model. Rendahnya recall kemungkinan dipengaruhi oleh distribusi kelas yang tidak seimbang, di mana jumlah sampel DIA jauh lebih sedikit dibandingkan Non-DIA. Akibatnya, model masih cenderung mempelajari pola dari kelas mayoritas dibandingkan kelas minoritas. Temuan ini menunjukkan bahwa penanganan ketidakseimbangan kelas masih menjadi faktor penting dalam pengembangan model prediksi DIA.

Dalam konteks keamanan obat, nilai recall memiliki peran yang sangat penting karena berkaitan langsung dengan kemampuan model mendeteksi senyawa yang berpotensi menyebabkan autoimunitas. Nilai recall sebesar 33,30% menunjukkan bahwa masih terdapat sejumlah kasus DIA yang tidak teridentifikasi (false negative). Kesalahan tipe ini memiliki implikasi yang lebih serius dibandingkan false positive karena berpotensi meloloskan kandidat obat yang sebenarnya memiliki risiko autoimunitas ke tahap pengembangan berikutnya. Oleh karena itu, peningkatan sensitivitas model terhadap kelas DIA menjadi prioritas utama untuk penelitian lanjutan.

Berdasarkan hasil perbandingan model, Random Forest dan Gradient Boosting menghasilkan nilai ROC-AUC yang lebih tinggi dibandingkan Adaptive Ensemble LS-SVM. Hal ini menunjukkan bahwa algoritma berbasis pohon keputusan memiliki kemampuan yang baik dalam memodelkan hubungan nonlinier yang kompleks antar molecular descriptors. Namun demikian, Adaptive Ensemble LS-SVM tetap menunjukkan peningkatan performa dibandingkan SVM tunggal, yang membuktikan efektivitas pendekatan ensemble yang diusulkan.

Hasil penelitian juga menunjukkan bahwa Random Forest dan Gradient Boosting memperoleh nilai ROC-AUC dan F1-Score yang lebih tinggi dibandingkan metode yang diusulkan. Kondisi ini mengindikasikan bahwa model berbasis pohon keputusan memiliki kemampuan yang lebih baik dalam menangkap hubungan nonlinier kompleks antar molecular descriptor. Oleh karena itu, kontribusi utama penelitian ini bukan pada pencapaian performa tertinggi dibanding seluruh model pembanding, melainkan pada pengembangan mekanisme Stability Feature Selection yang menghasilkan fitur lebih konsisten serta peningkatan performa yang signifikan dibandingkan pendekatan SVM tunggal.

Secara keseluruhan, hasil penelitian menunjukkan bahwa integrasi Stability Feature Selection dan Adaptive Ensemble LS-SVM merupakan pendekatan yang efektif untuk mengatasi permasalahan klasifikasi pada dataset Drug-Induced Autoimmunity yang memiliki karakteristik berdimensi tinggi dan tidak seimbang. Metode yang diusulkan mampu mengurangi jumlah fitur secara signifikan, meningkatkan performa dibandingkan model tunggal, serta menghasilkan model yang lebih stabil dan robust. Temuan ini menunjukkan bahwa pendekatan yang diusulkan memiliki potensi untuk dikembangkan lebih lanjut sebagai sistem pendukung keputusan dalam evaluasi keamanan obat pada tahap awal pengembangan farmasi.

Selain meningkatkan performa klasifikasi, Stability Feature Selection menghasilkan 20 descriptor yang secara konsisten dipilih oleh beberapa metode seleksi fitur. Descriptor tersebut berpotensi merepresentasikan karakteristik molekul yang berkaitan dengan risiko autoimunitas akibat obat. Meskipun interpretasi biologis secara rinci belum dilakukan dalam penelitian ini, hasil tersebut membuka peluang untuk penelitian lanjutan yang mengintegrasikan explainable artificial intelligence (XAI) guna mengidentifikasi hubungan antara descriptor terpilih dan mekanisme biologis yang mendasari terjadinya DIA.

Kesimpulan

Penelitian ini mengusulkan metode Adaptive Ensemble LS-SVM dengan Stability Feature Selection untuk prediksi Drug-Induced Autoimmunity (DIA) menggunakan molecular descriptors RDKit. Hasil penelitian menunjukkan bahwa Stability Feature Selection mampu mereduksi jumlah fitur dari 195 descriptor menjadi 20 fitur stabil, sehingga mengurangi kompleksitas data dan membantu meningkatkan efisiensi proses klasifikasi. Model Adaptive Ensemble LS-SVM yang dikembangkan menghasilkan nilai Accuracy sebesar 80,00%, Precision sebesar 71,40%, Recall sebesar 33,30%, F1-Score sebesar 45,50%, ROC-AUC sebesar 79,40%, dan MCC sebesar 0,390. Hasil tersebut menunjukkan bahwa pendekatan ensemble mampu meningkatkan performa klasifikasi dibandingkan model SVM tunggal serta menghasilkan model yang lebih stabil dalam memprediksi risiko autoimunitas akibat obat. Secara keseluruhan, integrasi Stability Feature Selection dan Adaptive Ensemble LS-SVM menunjukkan potensi yang baik dalam menangani permasalahan klasifikasi pada dataset berdimensi tinggi. Namun, nilai recall yang masih relatif rendah menunjukkan bahwa sensitivitas model terhadap kelas DIA perlu ditingkatkan melalui optimasi hyperparameter, metode penyeimbangan data yang lebih adaptif, serta integrasi pendekatan explainable artificial intelligence pada penelitian mendatang.

Deklarasi Kepentingan

Penulis menyatakan bahwa tidak terdapat konflik kepentingan (*conflict of interest*) yang dapat memengaruhi hasil, interpretasi, maupun publikasi penelitian ini. Penulis tidak memiliki kepentingan finansial, hubungan pribadi, maupun afiliasi lain yang berpotensi menimbulkan bias terhadap pelaksanaan dan hasil penelitian.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah memberikan dukungan, masukan, dan bantuan selama proses penelitian ini. Ucapan terima kasih juga disampaikan kepada pengelola UCI Machine Learning Repository atas penyediaan dataset yang digunakan dalam penelitian ini.

Ketersediaan Data

Dataset yang digunakan dalam penelitian ini adalah **Drug-Induced Autoimmunity Prediction Dataset** yang tersedia secara publik melalui **UCI Machine Learning Repository**. Dataset terdiri atas data pelatihan sebanyak 477 sampel dan data pengujian sebanyak 120 sampel dengan 195 molecular descriptors yang dihasilkan menggunakan perangkat lunak RDKit. Dataset tersebut digunakan untuk membangun model klasifikasi dalam memprediksi risiko *Drug-Induced Autoimmunity (DIA)* berdasarkan karakteristik molekul senyawa obat. Dataset dapat diakses secara terbuka melalui:

https://archive-beta.ics.uci.edu/dataset/1104/drug_induced_autoimmunity_prediction

Penelitian ini menggunakan dataset sesuai dengan ketentuan penggunaan dan lisensi yang ditetapkan oleh penyedia dataset. Seluruh sumber data telah disitasi dengan benar dalam naskah untuk memastikan kepatuhan terhadap hak cipta dan etika penggunaan data penelitian.

Penggunaan Ai Dan Deklarasi Penggunaan Ai Generatif

Penulis menggunakan alat berbasis kecerdasan buatan (*Artificial Intelligence/AI*) generatif sebagai alat bantu dalam proses penyusunan naskah, termasuk perbaikan tata bahasa, peningkatan kejelasan kalimat, dan penyempurnaan struktur penulisan. Seluruh penggunaan AI dilakukan di bawah pengawasan penuh penulis dan tidak digunakan sebagai pengganti analisis ilmiah, interpretasi hasil, pengambilan keputusan penelitian, maupun penarikan kesimpulan.

Daftar Pustaka

[1] L. Huang, P. Liu, and X. Huang, “InterDIA: Interpretable prediction of drug-induced autoimmunity through ensemble machine learning approaches,” *Toxicology*, vol. 511, p. 154064, Feb. 2025, doi: 10.1016/J.TOX.2025.154064.

[2] Y. Wu, J. Zhu, P. Fu, W. Tong, H. Hong, and M. Chen, “Machine learning for predicting risk of drug-induced autoimmune diseases by structural alerts and daily dose,” *Int. J. Environ. Res. Public Health*, vol. 18, no. 13, p. 7139, Jul. 2021, doi: 10.3390/IJERPH18137139/S1.

[3] Z. Yucheng, Z. Lu, and L. Shunan, “Machine Learning-Based Prediction of Drug-Induced Autoimmunity Using Molecular Descriptors,” *Acad. J. Comput. Inf. Sci.*, vol. 9, no. 1, pp. 41–47, Jan. 2026, doi: 10.25236/AJCIS.2026.090105.

[4] “Deep Learning Pembelajaran Biologi - Dr. Muji Sri Prastiwi, S.Pd., M.Pd. , Dr. Ulfi Faizah, S.Pd., M.Si., Dr. Walib Abdullah, S.Pd.I, M.Pd. , Dr. Atan Pramana, M.Pd. - Google Books.”

https://books.google.co.id/books?hl=en&lr=&id=3512EQAAQBAJ&oi=fnd&pg=PA1&dq=Berbagai+algoritma+ML+mampu+mempelajari+hubungan+kompleks+antara+struktur+molekul+dan+efek+biologis+yang+ditimbulkannya&ots=cfiLycs61l&sig=0QyI-UTbiVjfKBgrAQqNqiJ7KaM&redir_esc=y#v=onepage&q&f=false (accessed Jun. 13, 2026).

[5] "Drug Induced Autoimmunity Prediction - UCI Machine Learning Repository." https://archive-beta.ics.uci.edu/dataset/1104/drug_induced_autoimmunity_prediction (accessed Jun. 10, 2026).

[6] M. Mujahid *et al.*, "Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering," *J. Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/S40537-024-00943-4.

[7] A. Del Casale *et al.*, "Machine learning and pharmacogenomics at the time of precision psychiatry," *benthamdirect.com*, vol. 21, no. 12, pp. 2395–2408, Aug. 2023, doi: 10.2174/1570159X21666230808170123.

[8] K. Thirumoorthy and K. Muneeswaran, "Feature Selection for Text Classification Using Machine Learning Approaches," *Natl. Acad. Sci. Lett. 2021 451*, vol. 45, no. 1, pp. 51–56, Mar. 2021, doi: 10.1007/S40009-021-01043-0.

[9] J. Cai, J. Luo, S. Wang, S. Y.- Neurocomputing, and undefined 2018, "Feature selection in machine learning: A new perspective," *ElsevierJ Cai, J Luo, S Wang, S YangNeurocomputing, 2018•Elsevier*, Accessed: Mar. 28, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231218302911.

[10] K. Li, F. Wang, L. Yang, R. L.- Neurocomputing, and undefined 2023, "Deep feature screening: Feature selection for ultra high-dimensional data via deep neural networks," *ElsevierK Li, F Wang, L Yang, R LiuNeurocomputing, 2023•Elsevier*, 2023, Accessed: Mar. 28, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523122300293X.

[11] U. M. Khaire and R. Dhanalakshmi, "Stability of feature selection algorithm: A review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 4, pp. 1060–1073, Apr. 2022, doi: 10.1016/J.JKSUCI.2019.06.012.

[12] M. Buyukkececi and M. C. Okur, "A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learning," *Gazi Univ. J. Sci.*, vol. 36, no. 4, pp. 1506–1520, Dec. 2023, doi: 10.35378/GUJS.993763.

[13] F. Aziz, "Klasifikasi Pelanggan Deposito Potensial menggunakan Ensembel Least Square Support Vector Machine," *J. Syst. Comput. Eng.*, vol. 1, no. 1, p. 1, 2020, Accessed: Mar. 08, 2022. [Online]. Available: http://journal.unpacti.ac.id/index.php/JSCE/article/view/80.

[14] F. A. Lawi, A, A. Lawi, and F. Aziz, "Classification of credit card default clients using LS-SVM ensemble," *ieeexplore.ieee.org*, 2018, Accessed: Mar. 08, 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8780427/.

[15] F. Aziz, S. Usman, J. Jeffry, … N. A.-J. I., and undefined 2022, "Penerapan Algoritma Multiclass Ensemble Support Vector Machine dengan Fungsi Kernel untuk Klasifikasi Human Activity," *journal.nurulfikri.ac.id*, Accessed: Nov. 15, 2022. [Online]. Available: https://journal.nurulfikri.ac.id/index.php/jit/article/view/579.

[16] F. Aziz, A. Lawi, and E. Budiman, "Increasing Accuracy of Ensemble Logistics Regression Classifier by Estimating the Newton Raphson Parameter in Credit Scoring," in *5th International Conference on Computing Engineering and Design, ICCED 2019*, 2019, pp. 1–4, doi: 10.1109/ICCED46541.2019.9161078.