Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.202072857
Vol.7, No. 6, Desember 2020, hlm. 1221-1230 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No.36/E/KPT/2019 e-ISSN: 2528-6579

OPTIMASI DATA TIDAK SEIMBANG PADA INTERAKSI DRUG TARGET DENGAN *SAMPLING* DAN *ENSEMBLE SUPPORT VECTOR MACHINE*

Nabila Sekar Ramadhanti*[1], Wisnu Ananta Kusuma[2], Annisa[3]

[1,2,3]Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor
[2]Pusat Studi Biofarmaka Tropika, Institut Pertanian Bogor
Email: [1]nabila_skr@apps.ipb.ac.id, [2]ananta@apps.ipb.ac.id, [3]annisa@apps.ipb.ac.id
*Penulis Korespondensi

(Naskah masuk: 10 Desember 2019, diterima untuk diterbitkan: 26 November 2020)

Abstrak

Data tidak seimbang menjadi salah satu masalah yang muncul pada masalah prediksi atau klasifikasi. Penelitian ini memfokuskan untuk mengatasi masalah data tidak seimbang pada prediksi *drug-target interaction* (interaksi senyawa-protein). Ada banyak protein target dan senyawa obat yang terdapat pada basis data interaksi senyawa-protein yang belum divalidasi interaksinya secara eksperimen. Belum diketahuinya interaksi antar senyawa dan target tersebut membuat proporsi antara data yang diketahui interaksinya dan yang belum diketahui menjadi tidak seimbang. Data interaksi yang sangat tidak seimbang dapat menyebabkan hasil prediksi menjadi bias. Terdapat banyak cara untuk mengatasi data tidak seimbang ini, namun pada penelitian ini diimplementasikan metode yang menggabungkan *Biased Support Vector Machine* (BSVM), *oversampling,* dan *undersampling* dengan *Ensemble Support Vector Machine* (SVM). Penelitian ini mengeksplorasi pengaruh *sampling* yang digabungkan dalam metode tersebut pada data interaksi senyawa-protein. Metode ini sudah diuji pada dataset *Nuclear Receptor, G-Protein Coupled Receptor* dan *Ion Channel* dengan rasio ketidakseimbangan sebesar 14.6%, 32.36%, dan 28.2%. Hasil pengujian dengan menggunakan ketiga dataset tersebut menunjukkan nilai *area under curve* (AUC) secara berturut-turut sebesar 63.4%, 71.4%, 61.3% dan F-measure sebesar 54%, 60.7%, dan 39%. Meskipun nilai akurasi dari metode yang diusulkan lebih kecil dari metode SVM tanpa perlakuan apapun, hasil AUC dan F-measure menunjukkan bahwa metode yang diusulkan ini mampu menurunkan bias dari model yang menggunakan SVM tersebut. Hal ini ditunjukkan oleh peningkatan nilai AUC dan f-measure sekitar 5%-20%.

Kata kunci: data tidak seimbang, *ensemble*, over*sampling*, *undersampling*, SVM

IMBALANCED DATA OPTIMIZATION WITH SAMPLING AND SUPPORT VECTOR MACHINE ENSEMBLE

Abstract

Imbalanced data has been one of the problems that arise in processing data. This research is focusing on handling imbalanced data problem for drug-target (compound-protein) interaction data. There are many target proteins and drug compounds existed in compound-protein interaction databases, which many interactions are not validated yet by experiment. This unknown interaction led drug target interaction to become imbalanced data. A really imbalanced data may cause bias to prediction result. There are many ways of handling imbalanced data, but this research implemented some methods such as BSVM, oversampling, undersampling with SVM ensemble. This research is focusing on exploration of effect on the sampling that used in these method for compound-protein interaction data. This method had been tested on compound-protein interaction Nuclear Receptor, GPCR and Ion Channel with 14.6%, 32.36% and 28.2% of imbalance ratio. The evaluation result using these three dataset show the value of AUC respectively 63.4%, 71.4%, 61.3% and F-measure of 54%, 60.7% and 39%. Even though the score of accuracy of the proposed method is smaller than the SVM, the AUC and F-measure score shows that the proposed method can decrease the bias from the model which use SVM. This was shown from the increase of AUC and F-measure score from 5% to 20%.

Keywords: : *imbalanced data*, *ensemble*, over*sampling*, *undersampling*, SVM

1. PENDAHULUAN

Proses penemuan obat baru (*drug discovery)* memerlukan tahapan yang panjang yang berimplikasi pada tingginya biaya yang dibutuhkan dan waktu yang lama, berkisar antara 10 – 17 tahun (Roder & Thomson, 2015). Untuk mengatasi hal tersebut, dikenalkanlah pendekatan baru, yaitu *drug repurposing* yang merupakan pendekatan dengan menemukan manfaat khasiat baru dari obat yang sudah diteliti sebelumnya. Keuntungan dengan menggunakan pendekatan ini, profil farmakokinetik, farmakodinamik, dan toksisitas sudah diketahui secara umum dari tahap awal sehingga obat lebih cepat masuk ke uji klinis (Gupta dkk. 2013). Proses ini memudahkan penelitian pengobatan baru serta memungkinkan akademisi, pemerintah atau pihak lain untuk ikut berpartisipasi karena biaya dan waktu yang lebih terjangkau (Chong & Sullivan, 2007). Kemudahan mendapatkan obat baru dari proses *drug repurposing* ini tentu juga mempermudah pasien untuk mendapat akses terhadap obat tersebut (Pessetto 2013). Pendekatan *drug repurposing,* seperti halnya dengan *drug discovery,* memerlukan data interaksi senyawa dan protein untuk membangun model prediksinya.

Salah satu problem yang dihadapi ketika melakukan prediksi interaksi senyawa-protein adalah adanya ketidakseimbangan jumlah sampel setiap kelas pada data latih. Kondisi ini mempengaruhi kinerja model dalam melakukan prediksi interaksi sehingga hasil prediksinya menjadi bias. Oleh karena itu diperlukan tahap praproses data untuk memperkecil ketidakseimbangan data tersebut. Secara umum solusi untuk mengatasi masalah ketidakseimabangan data ini dapat dikelompokkan menjadi teknik *sampling*, *cost-sensitive learning, ensemble learning,* seleksi fitur, dan modifikasi algoritma (Choi 2010).

Pendekatan melalui data dapat dilakukan dengan metode *sampling* dan seleksi fitur pada bagian praproses data. Pada metode *sampling*, yaitu *oversampling*, *undersampling*, atau gabungan dari keduanya, terdapat banyak metode seperti *Synthetic Minority Oversampling Technique* (SMOTE) (Blagus dan Lusa, 2012), *Random Undersampling* (RUS) ataupun *Cluster Centroid Undersampling* (CCU). Pendekatan melalui data biasa digabung dengan metode *machine learning* untuk mengatasi kekurangannya yaitu seperti pada penelitian yang melakukan prediksi interaksi senyawa-protein dengan *ensemble learning* dan *oversampling* data. Pendekatan ini menghasilkan nilai *area under the curve* (AUC) sebesar 0.9 yang lebih besar dari metode *Decision Tree*, SVM, *Nearest Neighbor* dan *Random Forest* (Ezzat dkk. 2016). Selain itu Tomek Link juga digunakan sebagai proses reduksi data pada data tidak seimbang untuk data medis (Elhassan & Aljurf, 2016).

Penelitian pada data interaksi senyawa-protein dengan kombinasi dari *Complementary Fuzzy Support Vector Machine* (CMTFSVM) dengan SMOTE dan menghasilkan nilai akurasi, Gmean, AUC secara berturut-turut 83.46%, 68.12%, dan 53.19% (Kusuma dkk. 2019). Penelitian tersebut ditujukan untuk data interaksi senyawa-protein pada dataset IJAH (Indonesia Jamu *Herbs*) yang dapat digunakan untuk memprediksi hubungan dari interaksi tanaman-senyawa-protein-penyakit. Dataset IJAH memiliki ketidak-seimbangan yang sangat tinggi dan perlu dilakukan optimasi pada data tersebut. Nilai AUC yang merupakan indikasi penting pemodelan data tidak seimbang pada penelitian sebelumnya (Kusuma dkk. 2019) masih kurang optimal karena data tersebut menyebabkan hasil akurasi prediksi interaksi menjadi bias.

Selanjutnya terdapat penelitian dengan metode *Different Contribution Sampling* (DCS) yang berbasiskan *ensemble* SVM yang diusulkan oleh Jian dkk. (2016). Metode ini diimplementasikan pada beberapa macam dataset yang berasal dari repository data *machine learning University of Califormia, Irvine* (UCI) yang memiliki rasio ketidakseimbangan yang beragam. Pada penelitian tersebut digunakan sepuluh data dua kelas dan sembilan data dengan banyak kelas. Data tersebut biasa digunakan dalam diagnosis medis, pengenalan huruf, bioinformatika dan data lalu lintas. Penelitian tersebut mendapatkan hasil yang memuaskan dengan nilai *recall* dan AUC yang secara rata-rata mencapai 67.49% dan 95.068%. Nilai tersebut lebih baik dibanding nilai *recall* dan AUC dari model *Support Vector Machine* (SVM) tanpa *resampling*, SVM dengan *undersampling,* dan SVM dengan SMOTE. Metode *Different Contribution Sampling* (DCS) (Jian dkk. 2016) tersebut dengan dasar SVM yang juga digunakan pada metode CMTFSVM (Kusuma dkk. 2019) diterapkan pada beberapa dataset, tetapi belum pernah diterapkan pada data interaksi senyawa-protein. Secara umun nilai AUC dengan metode DCS lebih baik (95.068%) dari pada CMTFSVM (53.19%).

Penelitian ini berusaha memberikan kontribusi dengan mengusulkan pendekatan baru, yaitu dengan mengombinasikan CCU dan DCS (Jian dkk. 2016) untuk mengeksplorasi solusi masalah ketidak-seimbangan data seperti pada data interaksi senyawa-protein yang menghasilkan akurasi prediksi kurang optimal karena bias (Kusuma dkk. 2019). Penggunaan CCU menggantikan RUS yang dilakukan pada penelitian Jian dkk. 2016 karena memiliki performa lebih baik pada *imbalance class learning* (Zhang dkk. 2010).

2. METODE PENELITIAN

2.1. Data Penelitian

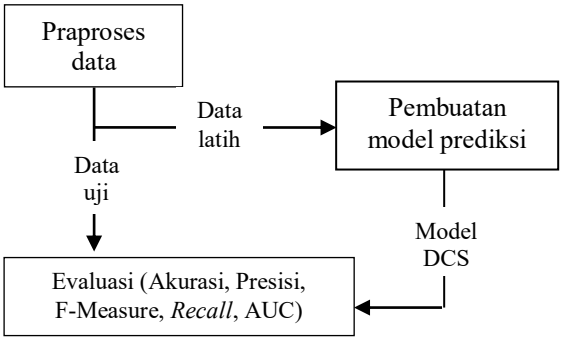
Dataset yang digunakan pada penelitian ini berasal dari *dataset* penelitian Yamanishi dkk. (2008) yang sudah banyak digunakan sebagai data pembanding dalam penelitian interaksi senyawa-protein. *Dataset* Yamanishi yang digunakan terdiri atas *dataset ion channel* (IC), *G-Protein Coupled Receptor* (GPCR) dan *nuclear receptor* (NR). Keterangan *dataset* Yamanishi dilihat pada Tabel 1.

Tabel 1. Statistik dataset Yamanishi dkk. (2008)

Dataset	#*Drug*	#*Target*	#Interaksi	Rasio tidak seimbang (%)
Ion channel	204	210	1476	28.02
GCPR	223	95	635	32.36
Nuclear receptor	26	54	90	14.6

2.2. Tahapan Penelitian

Penelitian yang dilakukan terdiri atas beberapa tahap (Gambar 1) yang dimulai dari proposes data yang terdiri atas akuisisi fitur data, praproses dan reduksi dimensi. Selanjutnya data dipisahkan menjadi data uji dan data latih. Data latih diproses dengan *Different Contribution* Sampling (DCS) yang melakukan B-SVM, *sampling*, dan *ensemble*. Selanjutnya dilakukan evaluasi dengan menghitung nilai akurasi, presisi, F-measure, *recall*, dan AUC.



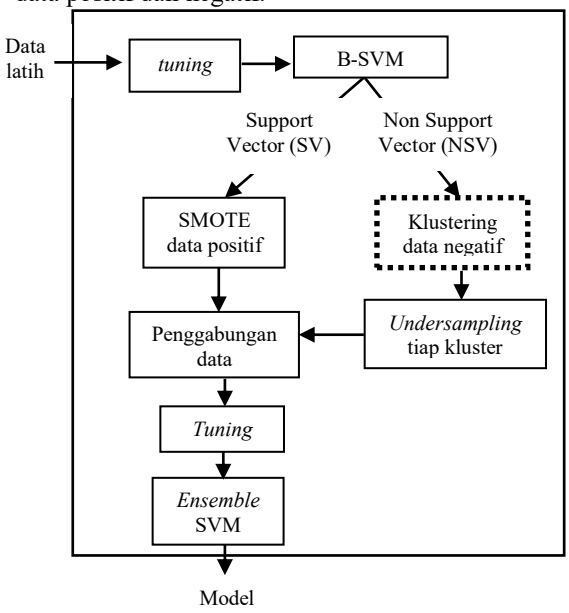
Gambar 1. Tahapan penelitian

2.3. Praproses Data

Data Yamanishi yang digunakan terdiri atas interaksi antar senyawa obat dan protein atau biasa disebut *drug-target interaction* (DTI) *data* dengan *Id* KEGG dan *Id* Uniprot dari senyawa dan protein. Fitur yang digunakan berdasarkan penelitian Mousavian dkk. (2016) menggunakan *package* Rcpi (Cao dkk. 2014) pada R di mana data *drug* yang berupa senyawa di representasikan dengan format SMILE yang diubah menjadi 881 fitur. Selanjutnya data target yang berupa protein direpresentasikan dengan *position specific scoring matrix* (PSSM) yang digabung dengan data senyawa menjadi satu vektor dengan total 1281 fitur dan 1 label kelas. Data yang sudah dilengkapi dengan fiturnya kemudian dinormalisasi, setelah itu data yang memiliki nilai yang sama di setiap barisnya dihapus dan dipisahkan menjadi data latih dan data uji.

2.4. Pembuatan Model Prediksi

Rincian langkah yang dilakukan pada pembuatan model prediksi dapat dilihat pada Gambar 2. Berbeda dari metode Different Contribution Sampling (DCS) yang dilakukan oleh Jian dkk. 2016, penelitian ini melakukan *clustering* sebelum *undersampling* dengan harapan hasil *undersampling* tersebar merata. Pada hasil penelitian ini metode diimplementasikan pada salah satu dataset Yamanishi, yaitu dataset *Nuclear Receptor.* Semula data awal dipisah menjadi data latih dan data uji. Selanjutnya dilakukan *tuning parameter* terhadap data latih kemudian dengan tujuan mendapat nilai parameter yang terbaik sebelum proses Biased Support Vector Machine (B-SVM). Kernel yang digunakan pada B-SVM ini adalah kernel RBF. Proses *tuning* menghasilkan nilai C, *gamma* dengan nilai *F-measure* yang tertinggi. Proses B-SVM menghasilkan data *support vector* data positif dan negatif.



Gambar 2. Pembuatan model prediksi

Modifikasi yang dilakukan untuk DCS pada penelitian ini, yaitu proses *undersampling* pada non-*support vector* random under sampling (NSV-RUS) yang dilakukan diganti menjadi *cluster centroid undersampling* (CCU). Sebelum dilakukan CCU, data non-*support vector* (NSV) dipisahkan dengan k-means *clustering* dengan tujuan *sampling* data yang dilakukan tersebar dengan lebih baik. Penelitian yang dilakukan oleh Zhang dkk. pada 2010 mendukung pemilihan CCU dengan menyatakan bahwa pendekatan *undersampling* kelas mayoritas berbasis kluster lebih unggul dibandingkan pendekatan lainnya pada data tidak seimbang.

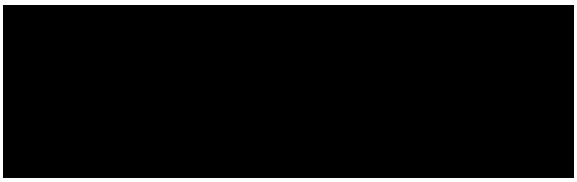
2.5. Rancangan percobaan

Rancangan percobaan yang dilakukan dapat dilihat pada Tabel 2. Pada sekanario satu digunakan metode Support Vector Machine (SVM) pada data yang belum dilakukan samping untuk melihat evaluasi dari model tanpa penanganan ketidakseimbangan data. Selanjutnya untuk melihat evaluasi model dengan *oversampling*, data positif diperbanyak dengan Synthetic Minority Oversampling Technique (SMOTE) lalu dibuat model prediksinya dengan SVM. Pada skenario 3 untuk melihat model dengan *undersampling*, data non-support vector minoritas direduksi dengan metode *undersampling* RUS dan Cluster Centroid Undersampling (CCU) kemudian model prediksinya dibuat dengan SVM. Terakhir menggunakan DCS yang menggabungkan *undersampling* dan *over-sampling* dengan *ensemble* SVM.

Tabel 2.Rancangan percobaan

No.	Sampling	Metode	Keterangan
1	-	SVM	Pembuatan model menggunakan SVM tanpa penanganan ketidakseimbangan data
2	Over-sampling	SMOTE-SVM	Pembuatan model menggunakan SVM dengan *oversampling* menggunakan SMOTE
3	Under-sampling	RUS-SVM	Pembuatan model menggunakan SVM dengan *undersampling* menggunakan RUS
		CCU-SVM	Pembuatan model menggunakan SVM dengan *undersampling* menggunakan CCU
4	Over-sampling + Under-sampling	DCS-RUS	Pembuatan model menggunakan DCS
		DCS-CCU	Pembuatan model menggunakan DCS dengan modifikasi *undersampling* menggunakan CCU

Hasil prediksi dari model ditabulasikan ke dalam *confusion matrix* (Tabel 3). Berdasarkan *confusion matrix* pada Tabel 2*,* pengujian klasifikasi *imbalanced data* dapat dilakukan dengan menghitung nilai *precission and recall* untuk masalah dengan kelas negatif dan positif. Nilai *precission* didapat dari persamaan (1) dan nilai *recall* didapat dari persamaan (2). Pada pengujian dengan *precission* dan *recall*, *imbalanced data learning* yang baik mampu meningkatkan nilai *recall* tanpa memengaruhi nilai *precission*. Pada beberapa kasus ketika TP meningkat, FP juga bisa meningkat yang mengakibatkan nilai *precission* turun, sehingga diperlukan F-measure (*f1*) pada persamaan (3) yang mengombinasikan *precission* dan *recall* dengan menghitung performa kelas minoritas secara menyeluruh yang menghindari masalah ketika TP dan FP meningkat secara bersamaan (Lin dkk. 2014).



Kurva *receiving operator characteristic* (ROC) digunakan untuk menguji *classifier* dengan melihat *true positive rate* (TPR) dan *false positive rate* (FPR). Nilai ideal dari ROC ada pada titik (0,100) di mana seluruh data minoritas ataupun mayoritas berhasil diklasifikasi dengan benar. Untuk melihat performa ROC ,nilai *Area Under the Curve* (AUC) dapat digunakan. Semakin besar nilai AUC, semakin bagus *classifier* yang digunakan. (Jian dkk. 2016)

Tabel 3. *Confusion matrix*

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

3. OPTIMASI DATA TIDAK SEIMBANG

3.1. Data Tidak Seimbang

Sebuah dataset dapat dikatakan tidak seimbang atau *imbalanced* apabila setiap kelas yang ada tidak merepresentasikan data dengan simbang (Chawla dkk. 2009). Secara umum data yang dimaksudkan sebagai data tak seimbang merupakan data yang salah satu kelasnya jauh lebih banyak dari kelas lain dengan skala perbedaannya bisa mencapai 1:1000 atau bahkan 1:10000. Pada dataset yang tidak seimbang terdapat sebutan kelas minoritas dan mayoritas. Untuk melakukan klasifikasi pada dataset dibutuhkan *classifier* yang akurat, namun pada data tidak seimbang seringkali akurasi prediksi oleh *classifier* ini juga tidak seimbang.

Untuk mengatasi masalah ini dapat dilakukan perlakuan dengan beberapa metode, antara lain: metode *sampling*, metode*cost-sensitive* serta metode berbasis *kernel* dan *active learning*. Pada metode *sampling*, data yang digunakan dimodifikasi sehingga menghasilkan distribusi yang seimbang. Ada beberapa sampling yang dapat digunakan, seperti *random oversampling*, *undersampling* atau *sampling* dengan melakukan generasi dan reduksi data. Metode *cost-sensitive* melakukan pengolahan data dengan mempertimbangkan biaya pada kelas positif dan negatif. Sampel yang termasuk kelas positif mendapat *cost* yang lebih besar.

3.2. *Support Vector Machine*

Support Vector Machine (SVM) dikembangkan oleh Vapnik dan timnya di AT&T Bell Laboratories pada 1979. SVM adalah suatu supervised learning untuk mencari hyperplane terbaik yang berfungsi sebagai pemisah dua kelas. SVM dibuat untuk menyelesaikan masalah klasifikasi biner sehingga tidak terlalu efektif jika digunakan pada data *multi-class*. Pada data yang besar penggunaan SVM membutuhkan waktu yang sangat lama. Menggunakan algoritma yang tepat bisa mengurangi waktu yang dibutuhkan namun mengurangi performa klasifikasi.

Ensemble SVM adalah gabungan dari beberapa *classifier* SVM untuk mengatasi masalah pada SVM (Kim dkk. 2002). Huang dkk. (2017) yang meneliti pengaruh SVM dan SVM ensemble pada prediksi kanker payudara membuktikan bahwa SVM ensemble bekerja sedikit lebih baik dari single SVM. SVM ensemble yang dilakukan pada penelitian ini merujuk pada penelitian Jian dkk. (2016) dengan menggabungkan model SVM dari hasil *oversampling* dan *undersampling* yang digabungkan dalam metode DCS.

3.3. *Cluster-Based Undersampling*

Undersampling adalah pendekatan data pada proses klasifikasi yang mereduksi data mayoritas untuk meningkatkan akurasi prediksi. Ilustrasi metode *undersampling* dapat dilihat pada Gambar 3. Pada gambar tersebut data negatif direpresentasikan dengan tanda (-) dan data positif direpresentasikan dengan tanda (+). Proses *undersampling* mengurangi data negatif sehingga setara dengan data positif. *Cluster-based undersampling* (CBU) merupakan metode undersampling berbasis *clustering*. Sebelum dilakukan *undersampling*, data mayoritas dikelompokkan terlebih dahulu

3.4. *Synthetic Minority Oversampling Technique*

Oversampling adalah sebuah pendekatan untuk memperbanyak data. Secara umum *oversampling*, tidak seperti *undersampling,* dapat menyebabkan *overfitting* di mana model yang terbentuk terlalu baik sehingga ketika diuji dengan data yang sedikit berbeda tidak akan sebaik model yang didapat. Oleh karena itu pendekatan *oversampling* biasa jarang digunakan untuk menyeimbangkan data (Blagus dan Lusa 2013). *Synthetic Minority Oversampling Technique* (SMOTE) menggunakan data minoritas dan membuat data sintetis dari data tersebut (Chawla dkk. 2002). SMOTE telah digunakan pada banyak penelitian lain dan memiliki performa yang baik dalam kasus data tidak seimbang (Wang dkk. 2006; Maciejewski dkk. 2011 & Ramentol dkk. 2012).

3.5. Different Contribution Sampling

Different Contribution Sampling (DCS) diusulkan oleh Jian dkk. pada 2016 untuk mengatasi masalah klasifikasi pada *imbalanced data*. tanpa menghilangkan bagian data yang penting. Pada penelitian Jian dkk. (2016), metode DCS dapat mengimbangi bahkan melebihi performa metode *under-sampling*, *syntetic minority over-sampling technique* (SMOTE) (Chawla dkk. 2002), dan *random over sampling.* Metode DCS meningkatkan akurasi prediksi kelas minoritas dengan hanya mengorbankan sedikit akurasi kelas mayoritas. Selain itu, dibandingkan metode lain, rata-rata nilai *recall* dengan metode DCS mencapai 0.6749 yang merupakan nilai *recall* paling tinggi dibanding metode lain pada penelitian Jian dkk. (2016).

Implementasi DCS dimulai dengan klasifikasi *support vector (SV)* dan *non support vector (NSV)* menggunakan teknik *biased* SVM (BSVM). Dari hasil BSVM tersebut dilakukan identifikasi dari kelas minoritas yang merupakan *support vector*, kemudian diimplementasikan teknik oversampling SMOTE yang akan memperbanyak data tersebut. Selanjutnya untuk data *non support vector* untuk kelas mayoritas dapat direduksi dengan *random under-sampling* (RUS) beberapa kali sehingga mendapat beberapa dataset negatif. Terakhir, data yang telah digabung dapat dilatih dengan SVM *ensemble* untuk memprediksi label dari *dataset.*

4. HASIL DAN PEMBAHASAN

4.1. Praproses data

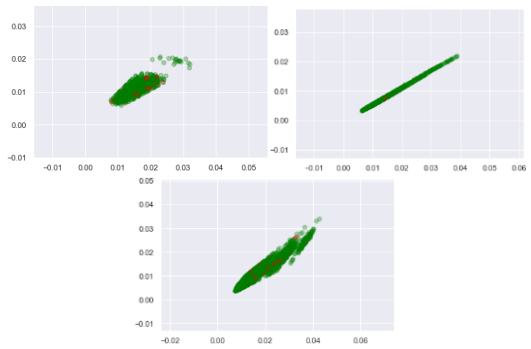
Pada tahap praproses data, fitur data yang sebelumnya 1282 berkurang setelah dilakukan pembersihan data dengan menghapus fitur data yang memiliki nilai yang sama pada setiap baris data dan dipisah menjadi data uji dan data latih. Hasil praproses data dapat dilihat pada Tabel 4. Selanjutnya dilakukan *tuning* pada data latih untuk mendapat parameter yang digunakan pada proses selanjutnya.

Tabel 4. Hasil praproses data

		Data latih		
	Jumlah fitur	**Jumlah data**	**Jumlah data negatif**	**Jumlah data positif**
NR	805	1 053	983	70
GPCR	944	15 888	15 429	468
IC	951	32 130	31 012	1 118
		Data uji		
	Jumlah fitur	**Jumlah data**	**Jumlah data negatif**	**Jumlah data positif**
NR	805	351	331	20
GPCR	944	5 297	5 130	167
IC	951	10 710	10 352	358

4.1. Pembuatan Model Prediksi SVM

Data latih yang digunakan pada skenario ini tidak melalui proses *sampling*, distribusinya dapat dilihat pada Gambar 3. Warna hijau menggambarkan data negatif sedangkan warna merah menggambarkan data positif.



Gambar 3. Distribusi data NR (kiri atas), GPCR (kanan atas) dan IC (bawah)

Tabel 5. Hasil evaluasi SVM

Evaluasi (%)	NR	GPCR	IC	rataan
AUC	50	49.9	51	50.3
Akurasi	94.3	97	97	96.1
Precision	0	0	38	12.3
Recall	0	0	2	0.76
F-measure	0	0	4	1.3
G-means	-	-	60.7	20.2

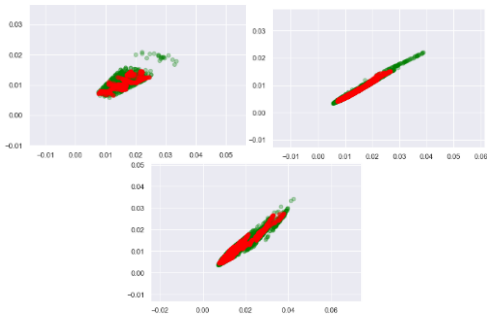
Pada gambar tersebut terlihat data didominasi data negatif. Hasil evaluasi pada Tabel 5 memperlihatkan nilai akurasi yang sangat tinggi pada seluruh dataset yang mencapai rataan 96.1%. Nilai tersebut adalah bias karena kenyataannya akurasi yang tinggi tersebut adalah dari prediksi benar dari data mayoritas yang negatif sedangkan seluruh data positif pada NR salah diklasifikasikan. Hal ini dapat dilihat pada *confusion matrix* pada Tabel 6, begitu pula pada dataset GPCR. Pada dataset IC, walaupun nilai G-means yang mencapai 60.7%, masih sangat tidak berimbang dengan prediksi benar dari data positif berjumlah delapan dari 350 data positif.

Tabel 6. *Confusion Matrix* menggunakan SVM

Dataset		*Predicted Negative*	*Predicted Positive*
NR	*Actual Negative*	331	0
	Actual Positive	20	0
GPCR	*Actual Negative*	5 122	8
	Actual Positive	167	0
IC	*Actual Negative*	10 339	13
	Actual Positive	350	8

4.2. Pembuatan model menggunakan SVM dan *oversampling*

Pada skenario ini data positif diperbanyak dengan metode SMOTE, distribusinya dapat dilihat pada Gambar 4. Warna hijau menggambarkan data negatif sedangkan warna merah menggambarkan data positif. Pada gambar tersebut terlihat warna merah atau data positif mulai tersebar merata dengan data negatif. Hasil evaluasi pada Tabel 7 memperlihatkan nilai akurasi yang sangat tinggi pada seluruh dataset yang mencapai rataan 81%, nilai ini memang tidak terlalu tinggi dibanding akurasi model SVM tetapi model SMOTE-SVM ini dapat memprediksi data positif lebih baik seperti yang terlihat pada *confusion matrix* pada Tabel 8. Prediksi data negatif pada model ini menurun sebanyak 15-25%, namun ini sebanding dengan peningkatan prediksi data positif yang mencapai 44-55%.



Gambar 4 Distribusi data dengan *oversampling* pada NR (atas), GPCR (tengah) dan IC (bawah)

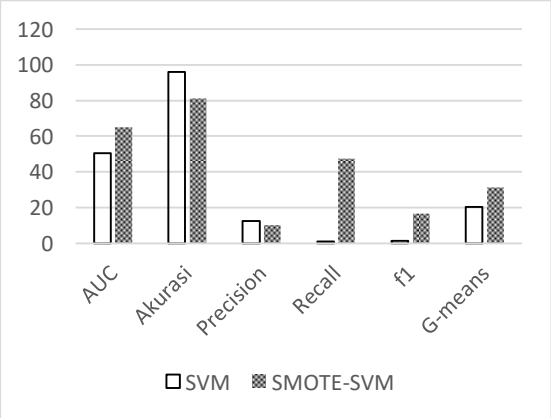
Tabel 7. Hasil evaluasi model SMOTE-SVM

Evaluasi (%)	NR	GPCR	IC	Rataan
AUC	66	64.6	64	64.9
Akurasi	76	83	84	81
Precision	13	8	9	10
Recall	55	45	42	47.3
f1	21	14	15	16.7
G-means	35	28.8	30	31.2

Tabel 8. *Confusion Matrix* model dengan SMOTE-SVM

Dataset		*Predicted Negative*	*Predicted Positive*
NR	*Actual Negative*	256	75
	Actual Positive	9	11
GPCR	*Actual Negative*	4 321	809
	Actual Positive	92	75
IC	*Actual Negative*	8 825	1 527
	Actual Positive	206	152

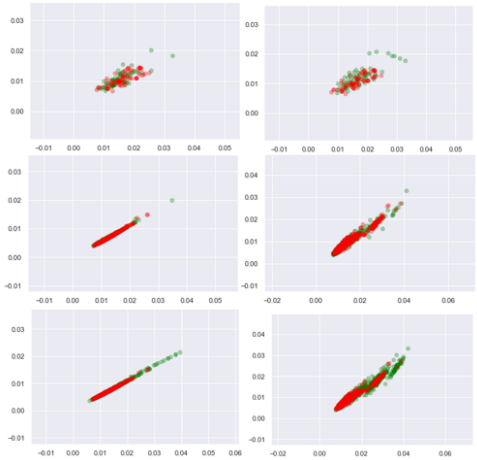
Penambahan data sintetik dengan SMOTE ini terbukti meningkatkan hasil evaluasi model dan menurunkan bias pada data. Pada Gambar 5 dapat dilihat SMOTE-SVM walaupun akurasinya lebih rendah dari model SVM, nilai evaluasi lainnya lebih tinggi. Nilai AUC, presisi, *recall*, *f1,* dan g-means SMOTE-SVM mencapai 64.9%, 10%, 47.3%, 16.7% dan 31.2% ketika model SVM hanya mencapai 50.3%, 12.3%, 0.76%, 1.3% dan 20.2%. Dari hasil evaluasi tersebut, model SMOTE-SVM memiliki performa yang lebih baik dari model SVM.



Gambar 5. Perbandingan model SVM dan SMOTE-SVM

4.3. Pembuatan model menggunakan SVM dan *undersampling*

Pada pemodelan ini digunakan teknik *undersampling* untuk mereduksi data negatif. Metode *undersampling* yang digunakan adalah Random Undersampling dan Cluster Centroid Undersampling. Distribusi data dapat dilihat pada Gambar 10. Pada gambar tersebut terlihat warna hijau atau data negatif yang pada distribusi awal (Gambar 6) hampir menutup data positif, berkurang sehingga data positif terlihat lebih jelas.



Gambar 6. Distribusi data dengan undersampling NR (atas), GPCR (tengah) dan IC (bawah) (RUS, CCU)

Distribusi data RUS dan CCU di sini tidak terlalu berbeda, namun pada Gambar 6 distribusi data CCU terlihat data negatif masih terlihat distribusinya lebih jelas dibanding data RUS. Hasil evaluasi pada Tabel 9 memperlihatkan bahwa hasil evaluasi SVM-RUS dan SVM-CCU tidak terlalu berbeda. Pada *confusion matrix* yang dihasilkan (Tabel 10) dapat dilihat data positif yang diprediksi benar meningkat sebanyak 10-65% namun disisi lain prediksi data negatif menurun hingga 30-56%. Model dengan *undersampling* mendapat nilai akurasi, AUC, presisi, *recall*, *F-measure* dan g-means: 60.2%, 63.2%, 6.5%, 66.5%, 12.0% dan 25.0%. Berbeda dari pembuatan model dengan *oversampling,* model ini memiliki nilai akurasi, g-means dan F-measure yang lebih kecil karena walaupun jumlah prediksi data positif meningkat, prediksi data negatif menurun. Nilai *recall* pada model *undersampling* meningkat karena prediksi positif lebih baik dari *oversampling*.

Tabel 9. Hasil evaluasi model dengan RUS-SVM dan CCU-SVM

Evaluasi RUS-SVM (%)						
Dataset	**AUC**	**Akurasi**	***Precision***	***Recall***	**f1**	**G-means**
NR	61.5	58	9	65	15	28.7
GPCR	66.3	63	6	69	11	23.9
IC	61.9	62	5	62	10	22.8
Rataan RUS	61	61	6.7	65.3	12	25.1
Evaluasi dan CCU-SVM (%)						
Dataset	**AUC**	**Akurasi**	***Precision***	***Recall***	**f1**	**G-means**
NR	59.7	60	8	60	14	28.1
GPCR	67.3	62	6	73	11	24
IC	62.4	56	5	70	10	22.3
Rataan CCU	59.3	59.3	6.3	67.7	12	24.8
Rataan Under-sampling	60.2	60.2	6.5	66.5	12	25

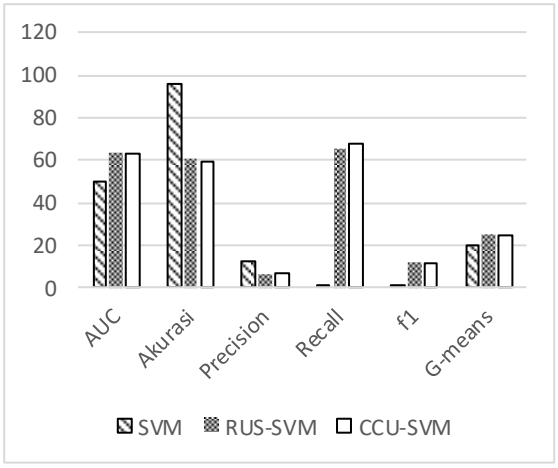
Tabel 10. *Confusion Matrix* model RUS-SVM dan CCU-SVM

		RUS		**CCU**	
Data-set		***Pred. Neg.***	***Pred. Pos.***	***Pred. Neg.***	***Pred. Pos.***
NR	*Actual Neg.*	192	139	197	134
	Actual Pos.	7	13	8	12
GPCR	*Actual Neg.*	3 240	1 890	3 160	1 970
	Actual Pos.	51	116	45	122
IC	*Actual Neg.*	6 395	3 975	5 702	4 650
	Actual Pos.	136	222	108	250

4.4. Pembuatan model menggunakan DCS

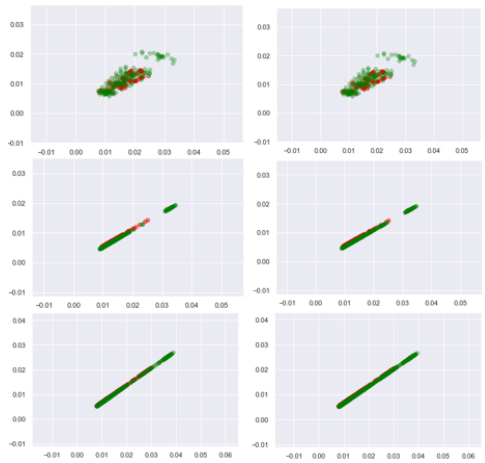
Distribusi data yang digunakan sebagai data latih pada metode DCS dapat dilihat pada Gambar 8. Distribusi data tersebut terlihat warna hijau atau data negatif mendominasi, namun sebenarnya data sudah dilakukan *sampling* dengan SMOTE dan *undersampling* sehingga rasio data positif dan negatif seimbang. Distribusi data dengan

undersampling CCU dan RUS tidak terlihat begitu berbeda begitu pula dengan evaluasi yang dilakukan.



Gambar 7. Perbandingan model SVM dengan RUS-SVM dan CCU-SVM

Hasil evaluasi pembuatan model dengan DCS-RUS dan DCS-CCU dapat dilihat pada Tabel 11. Nilai rataan evaluasi akurasi, AUC, presisi, *recall, f1* dan g-means yang didapat dengan metode DCS secara berturut-turut 66.8%, 65.7%, 8.3%, 64%, 14%, dan 27.9%. Walaupun CCU dan RUS secara umum tidak terlalu berpengaruh, pada data GPCR dan IC, jumlah data positif yang benar diprediksi meningkat pada model DCS-CCU sedangkan pada data NR, DCS-RUS memiliki performa yang lebih baik.



Gambar 8. Distribusi data dengan DCS NR (atas), GPCR (tengah) dan IC (bawah) (RUS, CCU)

Perbandingan SVM dengan DCS dapat dilihat pada Gambar 9. Pada grafik ini terlihat walaupun akurasi SVM jauh lebih baik, DCS mampu meningkatkan nilai evaluasi lain. Berdasarkan *confusion matrix* (Tabel 12), DCS juga terlihat mampu menurunkan kesalahan prediksi data negatif pada metode *undersampling* SVM hingga 55% untuk data NR dan 9% untuk data GPCR. Selain itu DCS dapat meningkatkan nilai prediksi data positif hingga 65% dibandingkan metode *oversampling*.

Tabel 11. Hasil evaluasi model dengan DCS-RUS dan DCS-CCU

Evaluasi DCS-RUS (%)						
Dataset	**AUC**	**Akurasi**	***Precision***	***Recall***	**f1**	**G-means**
NR	66.8	77	13	55	22	36
GPCR	70.2	66	7	74	12	25.7
IC	61.2	59	5	63	9	22.3
Rataan RUS	67.3	66.1	8.3	64	13.7	28
Evaluasi dan DCS-CCU (%)						
Dataset	**AUC**	**Akurasi**	***Precision***	***Recall***	**f1**	**G-means**
NR	63.4	80	13	45	20	35.4
GPCR	71.4	65	7	78	12	25.8
IC	61.3	54	5	69	9	21.9
Rataan CCU	66.3	65.4	8.3	64	13.7	27.7
Rataan DCS	66.8	65.7	8.3	64	13.7	27.9

Tabel 12. *Confusion Matrix* model DCS-RUS dan DCS-CCU

Data-set		**RUS**		**CCU**	
		Pred. Neg.	***Pred. Pos.***	***Pred. Neg.***	***Pred. Pos.***
NR	*Actual Neg.*	260	71	271	60
	Actual Pos.	9	11	11	9
GPCR	*Actual Neg.*	3 395	1 735	3 337	1 793
	Actual Pos.	43	124	37	130
IC	*Actual Neg.*	6 129	4 223	5 545	4 807
	Actual Pos.	132	226	111	247

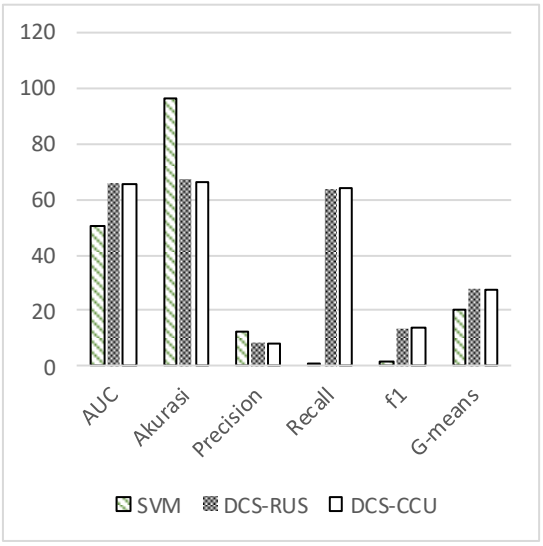
4.5. Evaluasi

Hasil evaluasi dari eksplorasi optimasi data tidak seimbang interaksi senyawa-protein dengan SVM dan *sampling* dapat dilihat pada Tabel 11. Nilai akurasi yang tinggi pada model SVM tidak dapat membuktikan bahwa model tersebut merupakan model yang representatif begitu pula dengan presisi model SVM yang mendapat nilai tertinggi dibanding lainnya; yaitu 96.1% dan 12.3%. Hal ini karena prediksi yang dihasilkan banyak yang benar namun untuk data negatif, sedangkan data positif yang benar diprediksi hanya sedikit sekali.

Pada model *oversampling* (SMOTE) prediksi data positif yang benar masih lebih sedikit dibanding model *undersampling* (RUS, CCU) dan DCS namun prediksi data negatif yang salah juga sangat sedikit dibanding model *undersampling* dan DCS. Model *undersampling* memiliki nilai *recall* tertinggi yaitu 66.5%. Oleh karena itu model *undersampling* dapat memprediksi data positif lebih baik dibanding model *oversampling* namun tidak dapat memprediksi data negatif lebih baik. Model DCS mendapat nilai AUC paling tinggi yaitu 65.7% walaupun akurasinya lebih kecil dari model SVM dan *oversampling*.

Tabel 1 Rataan hasil evaluasi setiap skenario pembuatan model

Metode	Evaluasi (%)					
	AUC	**Akurasi**	***Precision***	***Recall***	**f1**	**G-means**
SVM	50.3	**96.1**	**12.3**	0.76	1.3	20.2
SMOTE	64.9	81	10	47.3	**16.7**	**31.2**
RUS	61	61	6.7	65.3	12	25.1
CCU	59.3	59.3	6.3	**67.7**	12	24.8
Rataan	60.2	60.2	6.5	66.5	12	25
DCS-RUS	**67.3**	66.1	8.3	64	13.7	28
DCS-CCU	66.3	65.4	8.3	64	13.7	27.7
Rataan	65.7	66.8	8.3	64	13.7	27.9



Gambar 9. Perbandingan model SVM dengan DCS-RUS dan DCS-CCU

5. KESIMPULAN DAN SARAN

Penelitian yang telah dilakukan dapat membuktikan bahwa model SVM bias karena tidak dapat melakukan prediksi data positif dengan baik pada kasus data tidak seimbang. Selain itu model dengan hanya *sampling* pada satu sisi (*oversampling* data positif/ *undersampling* data negatif) juga masih bias karena pada model *oversampling* tidak bisa mendapat prediksi data positif sebaik model *undersampling,* sebaliknya model *undersampling* tidak bisa melakukan prediksi data negatif sebaik model *oversampling.* Model *oversampling* menghasilkan nilai f1 yang terbaik dari model lainnya, model *undersampling* mendapat nilai *recall* terbaik dan model SVM mendapat nilai akurasi dan presisi tertinggi. DCS yang menggabungkan dua sisi proses *sampling* dengan *ensemble* SVM dapat meningkatkan nilai AUC dari metode SVM, *oversampling* dan *undersampling*. Nilai AUC, pada kasus pemodelan seperti penelitian ini lebih difavoritkan karena menilai model lebih menyeluruh jika dibandingkan dengan metriks lainnya.

Hasil dari penelitian ini diharapkan akan mengurangi masalah yang disebabkan oleh data tidak seimbang pada interaksi senyawa-protein. Metode dari penelitian ini dapat selanjutnya dapat digunakan sebagai model prediksi untuk mendukung proses *drug repositioning* pada bidang farmasi.

Model DCS mampu mendapat hasil prediksi data cukup baik dengan prediksi data positif hampir sebaik model *undersampling* yang bisa dilihat dari nilai *recall* DCS yang lebih baik dari model *oversampling* dan mendekati nilai *recall* model *undersampling*. Model DCS juga mampu mendapat hasil prediksi data negatif lebih baik dari model *undersampling* yang bisa dilihat dari nilai presisi model DCS lebih baik dari model *undersampling*. Model DCS dengan CCU bekerja sebanding dengan RUS dan bisa menjadi alternatif yang cukup baik.

Pada penelitian selanjutnya mengenai metode DCS untuk meningkatkan nilai evaluasi dapat digunakan metode SVM yang dimodifikasi seperti *penalized* SVM sehingga dapat diujikan pada data tidak seimbang dari metode algoritmanya. Selain itu, metode *oversampling* dan *undersampling* juga bisa diubah menjadi modifikasi lainnya yang belum digunakan. Penelitian selanjutnya juga dapat melihat perbedaan metode dengan fitur yang berbeda.

UCAPAN TERIMA KASIH

Terima kasih penulis ucapkan atas kesempatan yang diberikan dari Kementrian Riset, Teknologi, dan Pendidikan melalui program hibah Penelitian Tesis Magister (PTM) dengan nomor 4408/IT3.L1/ PN/2019 pada tanggal 4 April 2019 sehingga hasil penelitian ini dapat dipublikasikan. Semoga tulisan ini dapat bermanfaat untuk penelitian lainnya.

DAFTAR PUSTAKA

BLAGUS, R., LUSA, L., 2012. Evaluation of smote for high-dimensional class-imbalanced microarray data. 11th International Conference on Machine Learning and Applications. 2, pp.89–94. doi:10.1109/ ICMLA.2012.183.

CAO, D.S., XIAO, N., XU, Q.S., CHEN, A.F., 2014. Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. Bioinformatics, 31(2), pp.279-281.

CHANG, CHIH-CHUNG; LIN, CHIH-JEN, 2011. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2.3, p.27.

CHAWLA, N.V., BOWYER, K.W., HALL, L.O., KEGELMEYER, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357.

CHAWLA, N.V., 2009. Data Mining for Imbalanced Datasets an Overview. Di dalam: Maimon, O., Rokach, L., editor. Data Mining and Knowledge Discovery Handbook. Boston: Springer.

CHOI, J.M., 2010. *A selective sampling method for imbalanced data learning on support vector machines.* Graduate Theses. Ioawa State University, USA.

CHONG, C.R., SULLIVAN, JR. DJ., 2007. New uses for old drugs.Nature,448(7154), p.645.

ELHASSAN, T., ALJURF, M., 2016. Classification of Imbalance Data using Tomek Link (T-Link) Combined with Random Under-sampling (RUS) as a Data Reduction Method. S1, p.111.

EZZAT, A., WU, M., LI, X.L., KWOH, C.K., 2016. Drug-target interaction prediction via class imbalance-aware ensemble learning. BMC bioinformatics, 17(19), p.509.

FITRIAWAN, A., 2013. Sistem klasifikasi khasiat formula jamu dengan metode support vector machine. Skripsi. Institut Pertanian Bogor, Indonesia.

GU, J., ZHANG, H., CHEN, L., XU, S., YUAN, G., XU, X., 2011. Drug–target network and polypharmacology studies of a Traditional Chinese Medicine for type II diabetes mellitus. Computational Biology and Chemistry, 35(5), pp.293–297.

GUPTA, S.C., SUNG, B., PRASAD, S., WEBB, L.J., & AGGARWAL, B.B., 2013. Cancer drug discovery by repurposing: teaching new tricks to old dogs. Trends in pharmacological sciences, 34(9), p.508

HUANG, M.W., CHEN, C.W., LIN, W.C., KE, S.W., TSAI, C.F., 2017. SVM and SVM Ensembles in BreastCancer Prediction. PLoS ONE, 12(1), p.e0161501.

JIAN, C., JIAN, G., AO, Y., 2016. A New Sampling Method for Classifying Imbalanced Data Based on Support Vector Machine Ensemble. Neurocomputing, 2(6).

KIM, H.C., PANG, S., JE, H.M., KIM, D., BANG, S.Y., 2002. Support Vector Machine Ensemble with Bagging. Di dalam: Lee SW, Verri A, editor. Pattern Recognition with Support Vector Machines: First International Workshop, SVM; 2002 Agustus 10; Niagara Falls, Kanada. Proceedings. doi:2388. 397-407.10.1007/3-540-45665-131.

KURNIA, A., 2017. Prediksi Formula Jamu Berkhasiat Menggunakan Teknik Link Prediction dari Jejaring Bipartite Senyawa Aktif dan Protein. Skripsi. Institut Pertanian Bogor, Indonesia.

KUSUMA, W.A., RAHMI, A.S., HERYANTO, R., 2019. Implementation of hybrid sampling technique for predicting active compound and protein interaction in unbalanced dataset. IOP Conference Series: Earth and Environmental Science, 335(1). doi:10.1088 /1755-1315/335/1/012005.

LI, Z.R., LIN, H.H., HAN, L.Y., JIANG, L., CHEN, X., & CHEN, Y.Z., 2006. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Research, 34(suppl_2), pp.W32-W37.

LIN, K.B., WENG, W., LAI, K., LU, P., 2014. Imbalance data classification algorithm based on SVM and clustering function. Conference: 2014 9th International Conference on Computer Science & Education (ICCSE), pp.544-548. doi:10.1109/ICCSE.2014.6926521.

MACIEJEWSKI, T., & STEFANOWSKI, J., 2011. Local neighbourhood extension of SMOTE for mining imbalanced data. IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp.104-111.

MOUSAVIAN, Z., KHAKABIMAMAGHANI, S., KAVOUSI, K., MASOUDI-NEJAD, A., 2016. Drug–target interaction prediction from PSSM based evolutionary information. J Pharmacol Toxicol Methods, 78, pp.42-51.

PESSETTO, Z.Y., WEIR, S.J., SETHI, G., Broward, M.A., Godwin, A.K., 2013. Drug repurposing for gastrointestinal stromal tumor. Molecular cancer therapeutics, 12(7), pp.1299-309.

RAMENTOL, E., CABALLERO, Y., BELLO, R., HERRERA, F., 2012. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowl Inf Syst, 33(2), pp.245-265.

RODER C., THOMSON M.J., 2015. Auranofin: repurposing an old drug for a golden new age. Drugs in R&D, 15(1), pp.13-20.

WANG, J., XU, M., WANG, H., ZHANG, J., 2006. Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. *8th international Conference on Signal Processing*; 2006 Nov 16-20; Beijing, China. doi:10.1109/ICOSP.2006. 345752.

YAMANISHI, Y., ARAKI, M., GUTTERIDGE, A., HONDA, W., KAMEHISA, M., 2008. Prediction of drug–target interaction networks from the intefration of chemical and genomic spaces. Bioinformatics, 24(8), pp.i232–i240.

ZHANG, Y. P., ZHANG, L. N., & WANG, Y. C., 2010. Cluster-based majority under-sampling approaches for class imbalance learning. In 2010 2nd IEEE International Conference on Information and Financial Engineering pp.400-404.