



PREDIKSI INTERAKSI *DRUG TARGET* PADA GEN KANKER MENGGUNAKAN METODE LASSO-XGBOOST

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

MUH. FADHIL AL-HAAQ GINOGA



**PROGRAM STUDI ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024**



PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul “Prediksi Interaksi Drug Target pada Gen Kanker Menggunakan Metode LASSO-XGBOOST” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Maret 2024

Muh. Fadhil Al-Haaq Ginoga
G6501202046

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
- a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
- b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.



RINGKASAN

MUH. FADHIL AL-HAAQ GINOGA. Prediksi Interaksi *Drug Target* Pada Gen Kanker Menggunakan Metode LASSO-XGBOOST. Dibimbing oleh WISNU ANANTA KUSUMA dan MUSHTHOFA.

Pengobatan kanker saat ini umumnya dilakukan dengan kemoterapi menggunakan senyawa kimia tunggal (obat) dan dapat menyebabkan berbagai efek samping bagi pasien. Alternatif pengobatan dapat memanfaatkan senyawa herbal yang diketahui memiliki efek samping lebih sedikit. Namun, pencarian senyawa herbal yang berpotensi sebagai kemoprefentif sulit dilakukan karena jumlah data yang sangat besar. Berbagai penelitian telah mengembangkan pendekatan komputasi atau *in silico* melalui Analisis *drug target interaction* (DTI) untuk menyaring interaksi senyawa sebagai obat (*drug*) dan protein terkait penyakit (*target*) yang berpotensi berhasil. Analisis DTI memanfaatkan pendekatan *machine learning* dengan menjadikan informasi senyawa dan protein sebagai masukan untuk diprediksi interaksinya. Salah satu pendekatan dalam Analisis DTI yaitu *chemogenomics* yang menggunakan informasi obat dan protein bersama-sama dalam suatu *space* untuk diprediksi. Pendekatan *chemogenomics* memiliki dua kategori, yaitu *feature based* dan *similarity based*. *Feature based* dilakukan dengan membentuk *feature vector* yang berisi informasi deskriptor senyawa dan deskriptor protein target. *Feature vector* selanjutnya dimasukkan dalam pemodelan *machine learning* untuk mendapatkan prediksi interaksinya sedangkan *similarity based* dilakukan dengan menghitung kemiripan antara senyawa obat dan target protein dengan membentuk suatu *similarity matrices*.

Pada penelitian ini Analisis DTI dilakukan dengan pendekatan *feature based* untuk memprediksi senyawa herbal yang berpotensi memiliki interaksi terhadap protein terkait kanker dengan membangun model yang dapat mempelajari pola interaksi pada data senyawa obat dan protein terkait kanker. Tahapan pada penelitian ini dimulai dengan melakukan pengumpulan dan praproses data, data protein terkait kanker diperoleh dari daftar *cancer gene census*, dari daftar tersebut dilakukan penelusuran pada *database genomics of drug sensitivity in cancer* (GDSC), DrugCentral, dan DrugBank untuk menghasilkan daftar senyawa obat yang berinteraksi dengan protein terkait kanker tersebut. Selain itu, senyawa herbal dihasilkan dari *database* HerbalDB dan KNAPSAcK. Pada penelitian ini juga menggunakan *dataset* Yamanishi *et al.* (*golden dataset*) sebagai *benchmarking* dengan penelitian serupa. Ekstraksi fitur dilakukan pada seluruh data, pada data senyawa herbal dan obat menggunakan PubChem *fingerprint* lalu untuk protein menggunakan *composition transition and distribution* (CTD), *pseudo amino acid composition* (PseAAC), *pseudo position specific scoring matrix* (PsePSSM), dan *dipeptide composition*. Namun, umumnya fitur yang terbentuk dari ekstraksi fitur tersebut sangat besar dan dapat mengurangi efisiensi model sehingga diterapkan pendekatan seleksi fitur *least absolute shrinkage and selection operator* (LASSO) untuk mereduksi dimensi fitur. Permasalahan lainnya yaitu *dataset* DTI yang tidak seimbang, dengan komposisi kelas yang memiliki interaksi (minoritas) dan belum diketahui interaksinya (majoritas) sehingga dilakukan penyeimbangan data dengan *synthetic minority oversampling technique* (SMOTE) dan *random undersampling*. Selanjutnya data yang telah seimbang digunakan sebagai masukan pada model yang



dibangun menggunakan *extreme gradient boosting* (XGBoost). XGBoost dipilih karena cukup unggul dalam melakukan prediksi dengan pendekatan *boosting* untuk meminimumkan bias. Tidak hanya itu, XGBoost juga memiliki parameter regularisasi sebagai nilai pinalti yang dapat menangani permasalahan *variance* yang dapat berakibat *overfitting*. Kombinasi pendekatan LASSO, SMOTE dan XGBoost dibangun menjadi suatu alur proses model *machine learning* pada tahap kedua penelitian ini, kombinasi tersebut selanjutnya disebut sebagai LASSO-XGBoost. Untuk mengevaluasi kinerja model dan menentukan fitur yang paling relevan, dilakukan perbandingan pada LASSO-XGBoost menggunakan *golden dataset* dan hasilnya menunjukkan bahwa kombinasi fitur PubChem *fingerprint* dan PsePSSM memberikan performa terbaik. Tahapan ketiga dilakukan pemodelan DTI senyawa obat dan protein kanker pada LASSO-XGBoost dengan kombinasi fitur PubChem *fingerprint* dan PsePSSM menghasilkan skor F1 0,861; AUROC 0,927; *recall* 0,85; *precision* 0,866; dan *accuracy* 0,897. Tahapan ketiga dilakukan prediksi interaksi senyawa herbal dan protein terkait kanker, model menghasilkan 5152 interaksi pada 396 senyawa herbal dan 62 protein kanker. Beberapa senyawa herbal berhasil diprediksi seperti andrographolide, *ursolic acid* dan *oleanolic acid* memiliki potensi sebagai kemoprefentif pada kanker dan dapat ditemukan pada berbagai tanaman di Indonesia.

Kata kunci: Interaksi obat target, Kanker, LASSO, Senyawa herbal, XGBoost.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



SUMMARY

MUH. FADHIL AL-HAAQ GINOGA. Drug Target Interaction Prediction On Cancer Gene Using LASSO-XGBOOST Method. Supervised by WISNU ANANTA KUSUMA and MUSHTHOFA.

Cancer treatment currently is commonly conducted through chemotherapy using drugs containing single chemical compounds, which can cause various side effects for patients. Alternative treatments may utilize herbal compounds known to have fewer side effects. However, the search for herbal compounds with potential chemopreventive properties is challenging due to the vast amount of data. Various studies have developed computational or *in silico* approaches through drug target interaction (DTI) analysis to screen compound interactions as drugs and disease-related proteins that have the potential to be successful. DTI analysis utilizes machine learning approaches by incorporating compound and protein information as inputs to predict their interactions. One approach in DTI analysis is chemogenomics, which integrates drug and protein information in a space to make predictions. Chemogenomics has two categories that is feature-based and similarity-based. Feature-based involves forming a feature vector containing compound and target protein descriptor information, which is then used in machine learning modeling to predict interactions, while similarity-based involves calculating the similarity between drug compounds and target proteins by forming similarity matrices.

In this study, DTI analysis is conducted using a feature-based approach to predict herbal compounds that potentially interact with cancer-related proteins by building a model that can learn interaction patterns from single chemical compound data and cancer-related proteins. The stages of this research begin with data collection and preprocessing, where cancer-related protein data is obtained from the cancer gene census list. From this list, a search is conducted in the database genomics of drug sensitivity in cancer (GDSC), DrugCentral, and DrugBank databases to produce a list of single chemical compounds that interact with these cancer-related proteins. Additionally, herbal compounds are obtained from the HerbalDB and KNAPSAcK databases. This study also utilizes the Yamanishi et al. dataset (golden dataset) as a benchmark for similar research. Feature extraction is performed on all data, using PubChem fingerprint for herbal and single chemical compound data, and composition transition and distribution (CTD), pseudo amino acid composition (PseAAC), pseudo position specific scoring matrix (PsePSSM), and dipeptide composition for proteins. However, the features formed from this extraction are generally large and can reduce model efficiency, hence the application of the least absolute shrinkage and selection operator (LASSO) feature selection approach to reduce feature dimensions.

Another issue is the imbalanced DTI dataset in terms of classes, where interactions exist (minority) and interactions are unknown (majority), so data balancing is performed using synthetic minority oversampling technique (SMOTE) and random undersampling. The balanced data is then used as input for the model constructed using extreme gradient boosting (XGBoost). XGBoost is chosen because it excels in prediction with a boosting approach to minimize bias. Additionally, XGBoost has regularization parameters as penalty values to handle



variance issues that may lead to overfitting. The combination of LASSO, SMOTE, and XGBoost approaches is constructed into a machine learning model workflow in the second stage of this research, referred to as LASSO-XGBoost. To evaluate model performance and determine the most relevant features, a comparison is made using the golden dataset, indicating that the combination of PubChem fingerprint and PsePSSM features provides the best performance. In the third stage, DTI modeling of single chemical compounds and cancer-related proteins using LASSO-XGBoost with the combination of PubChem fingerprint and PsePSSM features yields F1 score of 0,861; AUROC of 0,927; recall of 0,85; precision of 0,866; and accuracy of 0,897. In the third stage, interaction prediction between herbal compounds and cancer-related proteins is performed, resulting in 5152 interactions involving 396 herbal compounds and 62 cancer related proteins. Some herbal compounds successfully predicted, such as andrographolide, ursolic acid, and oleanolic acid, have the potential as chemopreventive agents in cancer and can be found in various plants in Indonesia.

Keywords: Cancer, Drug target interaction, Herbal compounds, LASSO, XGBoost.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah,
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2024¹
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.



PREDIKSI INTERAKSI *DRUG TARGET* PADA GEN KANKER MENGGUNAKAN METODE LASSO-XGBOOST

MUH. FADHIL AL-HAAQ GINOGA

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister pada
Program Studi Ilmu Komputer

**PROGRAM STUDI ILMU KOMPUTER
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
INSTITUT PERTANIAN BOGOR
BOGOR
2024**



IPB University

©Hak cipta milik IPB University

Tim Penguji pada Ujian Tesis:
Dr. Aziz Kustiyo, S.Si., M.Kom.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



Judul Thesis : Prediksi Interaksi *Drug Target* Pada Gen Kanker Menggunakan Metode LASSO-XGBOOST
Nama : Muh. Fadhil Al-Haaq Ginoga
NIM : G6501202046

Disetujui oleh



Pembimbing 1:
Dr. Eng. Wisnu Ananta Kusuma, S.T., M.T.

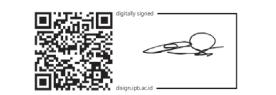


Pembimbing 2:
Dr. Mushthofa, S.Kom., M.Sc.

Diketahui oleh



Ketua Program Studi:
Prof. Dr. Imas Sukaesih Sitanggang, S.Si, M.Kom
NIP 197501301998022001



Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam:

Dr. Berry Juliandi, S.Si., M.Si.
NIP 19780723 200701 1 001



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

PRAKATA

Puji dan syukur penulis panjatkan kepada Allah subhanaahu wa ta'ala atas segala karunia-Nya sehingga karya ilmiah ini berhasil diselesaikan. Tema yang dipilih dalam penelitian yang dilaksanakan sejak bulan Januari 2022 sampai bulan Februari 2024 ini ialah bioinformatika, dengan judul “Prediksi Interaksi Drug Target Pada Gen Kanker Menggunakan Metode LASSO-XGBOOST”.

Ucapan terima kasih disampaikan kepada orangtua penulis bapak Dr. Verri Ginoga, S.E., M.M. dan ibu Musdiana, A.Md yang telah merawat dan membimbing penulis sedari kecil hingga dapat menyelesaikan penelitian ini. Ucapan terima kasih juga ditujukan kepada para dosen, selaku dosen pembimbing Dr. Eng. Wisnu Ananta Kusuma, S.T., M.T. dan Dr. Mushthofa, S.Kom., M.Sc., dosen moderator Prof. Dr. Imas Sukaesih Sitanggang S.Si., M.Kom, dan dosen penguji luar komisi Dr. Aziz Kustiyo, S.Si., M.Kom. Di samping itu, penghargaan penulis sampaikan kepada Kementerian Pendidikan dan Pusat Penelitian Biofarmaka Tropika LPPM IPB yang telah membantu selama kegiatan penelitian.

Tidak lupa pula penulis haturkan terima kasih kepada Sitti Filzha Fitrya Ginoga, Nila Wildanul Husna, Eka Pratama Ginoga dan keluarga yang tidak dapat penulis sebutkan satu per satu karena selalu memberikan semangat dalam menyelesaikan penelitian ini.

Semoga karya ilmiah ini bermanfaat bagi pihak yang membutuhkan dan bagi kemajuan ilmu pengetahuan.

Bogor, Maret 2024

Muh. Fadhil Al-Haaq Ginoga

**DAFTAR TABEL****DAFTAR GAMBAR****PENDAHULUAN**

1.1	Latar Belakang	xi
1.2	Rumusan Masalah	xi
1.3	Tujuan	1
1.4	Manfaat	1
1.5	Ruang Lingkup	3

TINJAUAN PUSTAKA

2.1	Kanker	5
2.2	<i>Drug Target Interaction</i>	5
2.3	Seleksi Fitur	6
2.4	<i>Least Absolute Shrinkage and Selection Operator</i>	8
2.5	<i>Ensemble Boosting</i>	9
2.6	<i>Extreme Gradient Boosting</i>	10
2.7	<i>Synthetic Minority Oversampling Technique</i>	10
2.8	<i>Random Undersampling</i>	13
2.9	<i>Golden Dataset</i>	14

III METODE

3.1	Data Penelitian	15
3.2	Tahapan Penelitian	15
3.3	Praproses Data	19
3.4	Pembangkitan label DTI	20
3.5	Pemodelan DTI dengan LASSO-XGBOOST	26
3.6	Evaluasi Model	27
3.7	Klasifikasi Senyawa Herbal	28
3.8	Lingkungan Pengembangan	31

IV HASIL DAN PEMBAHASAN

4.1	Praproses dan Analisis Data	34
4.2	Pemodelan LASSO-XGBoost	34
4.3	Klasifikasi Senyawa Herbal	36

V SIMPULAN DAN SARAN

5.1	Simpulan	42
5.2	Saran	47

DAFTAR PUSTAKA**LAMPIRAN****RIWAYAT HIDUP**



1	Deskripsi beberapa model prediksi DTI yang telah dikembangkan	7
2	Deskripsi <i>database</i> yang digunakan	15
3	Contoh data protein	16
4	Contoh data interaksi senyawa obat dan protein terkait kanker	17
5	Contoh data senyawa	18
6	Jumlah data yang dikumpulkan	18
7	Jumlah data pada <i>golden dataset</i>	19
8	Tabel nilai <i>Hydrophobicity</i> , <i>Hydrophilicity</i> , dan <i>Side Chain Mass</i>	23
9	Nilai <i>physicochemical</i> yang digunakan pada CTD	24
10	Nilai <i>hyperparameter</i> pada model XGBoost	28
11	<i>Confussion Matrix</i>	29
12	Deskripsi <i>library</i> yang digunakan	32
13	Jumlah Fitur Setiap Jenis Ekstraksi Fitur	34
14	Contoh data protein hasil ekstraksi fitur CTD	34
15	Contoh data protein hasil ekstraksi fitur DC	34
16	Contoh data protein hasil ekstraksi fitur PsePSSM	35
17	Contoh data protein hasil ekstraksi fitur PseAAC	35
18	Contoh data senyawa hasil ekstraksi fitur PubChem <i>fingerprint</i>	35
19	Hasil pengujian pada model <i>dataset</i> enzim	37
20	Hasil pengujian pada model <i>dataset</i> NR	37
21	Hasil pengujian pada model <i>dataset</i> GPCR	37
22	Hasil pengujian pada model <i>dataset</i> IC	37
23	Hyperparameter Optimal pada LASSO-XGBOOST	42
24	<i>Confussion matrix</i> prediksi DTI senyawa obat	42
25	Hasil prediksi interaksi senyawa herbal-protein	43
26	Tanaman yang memiliki kandungan senyawa herbal diprediksi	45

DAFTAR GAMBAR

1	Cara kerja XGBoost	11
2	Transformasi senyawa paclitaxel ke bentuk SMILES	18
3	Tahapan penelitian	19
4	Representasi PubChem <i>fingerprint</i>	20
5	Skema <i>sequence order</i> dengan 3 tingkatan (Chou 2001)	21
6	Visualisasi <i>complete bipartite graph</i>	27
7	Alur LASSO-XGBoost	27
8	Contoh kurva AUROC	30
9	Visualisasi kurva AUCPR	31
10	Perbandingan label 0 dan 1 pada <i>dataset</i>	35
11	Perbandingan label 0 dan 1 pada <i>dataset</i> yang diseimbangkan	36
12	PRC <i>dataset</i> enzim	38
13	PRC <i>dataset</i> NR	39

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



14	PRC dataset GPCR	39
15	PRC dataset IC	40
16	Perbandingan ROCAUC pada beberapa model prediksi DTI	41
17	Perbandingan PRC LASSO-XGBoost dengan model lain	41

@Hak cipta milik IPB University

IPB University

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber:
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.