LOGIC: Jurnal Penelitian Informatika
Vol. 1, Issue 1, pp. 1–7 (2023)
doi: http://doi.org/10.25124/logic.v1i1.6489

RESEARCH ARTICLE

Model Klasifikasi Berbasis Ekspresi *Gen Non-Small Cell Lung Carcinoma* (NSCLC) pada Wanita Bukan Perokok Menggunakan Metode *Ensemble*

Kholishoh Nur Azizah, Fhira Nhita* and Isman Kurniawan

Fakultas Informatika, Universitas Telkom, Bandung, 40257, Jawa Barat, Indonesia
*Corresponding author: fhiranhita@telkomuniversity.ac.id
Received on 31 July 2023; accepted on 03 September 2023

Abstrak

Kanker paru-paru adalah penyebab utama kematian terkait kanker di seluruh dunia dan membawa dampak sosial ekonomi yang signifikan bagi pasien, keluarga, dan masyarakat secara keseluruhan. Dalam diagnosis kanker, klasifikasi berbagai jenis tumor sangat penting. Prediksi akurat dari berbagai jenis tumor memungkinkan untuk pengobatan yang lebih baik dan meminimalkan toksisitas pada pasien. Untuk menganalisis masalah klasifikasi kanker menggunakan data ekspresi gen, untuk pemilihan fitur dan model prediksi. Penelitian ini bertujuan untuk memprediksi NSCLC dengan menerapkan metode ensemble pada data *microarray*. Penulis menggunakan tiga metode *ensemble* untuk memprediksi NSCLC, yaitu *Random Forest, Adaptive Boosting (AB)*, dan *Extreme Gradient Boosting* (XG). Seleksi fitur dilakukan menggunakan *variance threshold* dan *parameter chi-square* kemudian dilanjutkan dengan membangun model prediksi menggunakan *ensemble*. Hasil validasi model terbaik berdasarkan yang terdiagnosis kanker yaitu model AB dengan 10 fitur, XG dengan 10 fitur, dan XG dengan 20 fitur yang menghasilkan nilai *accuracy*, *recall*, dan *f1-score* yang sama berturut-turut yaitu 0.93, 1.00, dan 0.93.

Key words: *Chi-Square, Ensemble, Microarray, Variance Threshold*

Pendahuluan

Kanker paru-paru adalah penyebab utama kematian terkait kanker di seluruh dunia dan membawa dampak sosial ekonomi yang signifikan bagi pasien, keluarga, dan masyarakat secara keseluruhan [1]. N*on-Small Cell Lung Carcinoma* (NSCLC) menyumbang sebagian besar tumor paru-paru [1]. Di antara *Non-Small Cell Lung Carcinoma*, *adenocarcinoma* adalah tipe histologis utama *lung carcinoma* di Taiwan (52,5%) [1]. Merokok merupakan faktor resiko utama untuk kanker paru-paru, meskipun ada faktor lain, seperti paparan lingkungan (misalnya, bahan kimia, agen fisik, dan radiasi), riwayat klinis penyakit paru-paru (misalnya, bronchitis kronis, emfisema, pneumonia, dan tuberculosis), riwayat tumor familial, atau diet juga dapat dikaitkan dengan perkembangan kanker paru-paru [1]. Di negara Barat 70% sampai 90% dari kanker paru-paru disebabkan oleh merokok, sedangkan di Taiwan hanya 7% dari kasus kanker paru-paru wanita yang berhubungan dengan merokok [1]. Banyak gen (misalnya, TP53, EGFR, KRAS, PIK3CA, dan EML4-ALK) telah dilaporkan berhubungan dengan kanker paru-paru pada tidak pernah merokok, meskipun mekanisme molekuler NSCLC pada wanita tidak merokok masih belum jelas [1]. Pada tahun 2012, kanker paru-paru menyumbang 1,6 juta kematian dan 1,8 juta kasus baru [2]. Ini adalah jenis pembunuh kanker yang paling umum pada pria dan wanita AS, dan menyebabkan lebih banyak kematian daripada gabungan kanker kolorektal, payudara, dan prostat [2].

Menurut ringkasan dari 10 tahun terakhir untuk diagnosis NSCLC kelenjar getah bening mediastinum menggunakan 18F-FDG PET/CT, sensitivitas median hanya 62% yang berarti sebagian besar metastasis dinilai negative palsu [1]. Untuk meningkatkan sensitivitas diagnosis NSCLC kelenjar getah bening mediastinum, diperlukan strategi klasifikasi yang lebih canggih dan algoritma pembelajaran mesin [1]. Dalam diagnosis kanker, klasifikasi berbagai jenis tumor sangat penting [2]. Prediksi akurat dari berbagai jenis tumor memungkinkan untuk pengobatan yang lebih baik dan meminimalkan toksisitas pada pasien [2]. Metode tradisional untuk mengatasi situasi ini terutama didasarkan pada karakteristik morfologi jaringan tumor [2]. Metode konvensional ini dilaporkan memiliki beberapa keterbatasan 7 diagnosis [2]. Untuk menganalisis masalah klasifikasi kanker menggunakan data ekspresi gen, pendekatan yang lebih sistematis telah dikembangkan [2].

Penggunaan pembelajaran mesin pada data ekspresi gen sangat diperlukan untuk mendeteksi adanya kanker paru-paru. Beberapa penelitian telah menggunakan metode tersebut, antara lain pada tahun 2015, Chen Yen dkk. melakukan kemoterapi adjuvant (ACT) untuk Non-Small Cell Lung Carcinoma dengan hasil akurasi klasifikasi adalah 65,71% [3]. Tahun 2021, Margarita Kirienko dkk. melakukan penelitian menggunakan 18FFDG PET/CT yang menghasilkan area di bawah kurva (AUC) sebesar 0,87 [4]. Tahun 2009, Peng Guan dkk. melakukan penelitian menggunakan Support Vector Machine (SVM)

yang menghasilkan akurasi metode yang dimodifikasi meningkat dari 98,86% menjadi 100% pada set pelatihan dan dari 98,51% menjadi 99,06% pada set pengujian serta standar deviasi dari metode yang dimodifikasi meningkat dari 98,86% menjadi 100% pada set pelatihan dan dari 98,51% menjadi 99,06% pada set pengujian [7]. Tahun 2022, Jose Marcio Luna dkk. melakukan penelitian menggunakan pendekatan pembelajaran mesin mengasilkan 11,4% pasien mengalami esophagitis radiasi derajat 3 [5].

Tahun 2021, Nguyen Quoc Khanh Le dkk. melakukan penelitian menggunakan genetic algoritm plus XGBoost classifier menghasilkan akurasi 0,836 dan 0,86 [6]. Salah satu metode pembelajaran mesin yang biasa digunakan dalam tugas prediksi adalah metode ensemble [7]. Ensemble merupakan algoritma efektif yang menggabungkan semua algoritma pembelajaran untuk meningkatkan akurasi [7]. Keuntungan teknik algoritma ini yaitu dapat mengurangi masalah ukuran sampel yang kecil secara rata-rata dan menggabungkan dari model untuk mencegah overfitting dari data latih [7]. Oleh karena itu, metode ensemble menjanjikan untuk meningkatkan akurasi prediksi pada NSCLC [7]. Penelitian ini bertujuan untuk memprediksi NSCLC dengan menerapkan metode *ensemble* pada data *microarray*. Penulis menggunakan tiga metode ensemble untuk memprediksi NSCLC, yaitu *Random Forest, Adaptive Boosting*, dan *Extreme Gradient Boosting*.

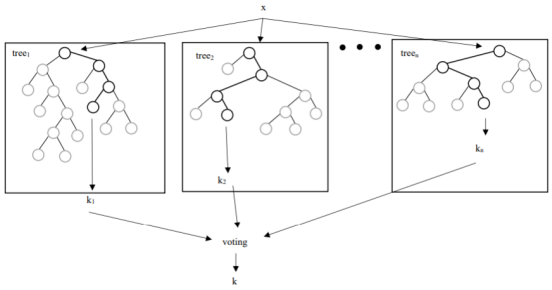
Tinjauan Pustaka

Penggunaan pembelajaran mesin pada data ekspresi gen sangat diperlukan untuk mendeteksi adanya kanker paru-paru. Beberapa penelitian telah menggunakan metode tersebut, antara lain pada tahun 2015, Chen Yen dkk. melakukan kemoterapi *adjuvant* (ACT) untuk *Non-Small Cell Lung Carcinoma* dengan hasil akurasi klasifikasi adalah 65,71% [3]. Tahun 2021, Margarita Kirienko dkk. melakukan penelitian menggunakan 18FFDG PET/CT yang menghasilkan area di bawah kurva (AUC) sebesar 0,87 [4]. Tahun 2009, Peng Guan dkk. melakukan penelitian menggunakan *Support Vector Machine* (SVM) yang menghasilkan akurasi metode yang dimodifikasi meningkat dari 98,86% menjadi 100% pada set pelatihan dan dari 98,51% menjadi 99,06% pada set pengujian serta standar deviasi dari metode yang dimodifikasi meningkat dari 98,86% menjadi 100% pada set pelatihan dan dari 98,51% menjadi 99,06% pada set pengujian [7].

Tahun 2022, Jose Marcio Luna dkk. melakukan penelitian menggunakan pendekatan pembelajaran mesin mengasilkan 11,4% pasien mengalami esophagitis radiasi derajat 3 [5]. Tahun 2021, Nguyen Quoc Khanh Le dkk. melakukan penelitian menggunakan *genetic algoritm plus XGBoost classifier* menghasilkan akurasi 0,836 dan 0,86 [6]. Tahun 2018 oleh Arunkumar dan Ramakrishnan menggunakan *algoritma fuzzy rough quick reduct* menghasilkan akurasi pengklasifikasi 97,22%, 99,45% dan 99,6% masing-masing pada set data ekspresi gen kanker paru-paru dan ovarium [8]. Tahun 2014, Samuel Hawkins dkk. melakukan penelitian menggunakan CT scan menghasilkan akurasi 77,5% [9]. Tahun 2013, Rainer J Klemetnt dkk. melakukan penelitian menggunakan *Support Vector Machine* (SVM) menghasilkan sensitivitas 67,0% ± 0,5% dan spesifitas 78,7% ± 0,3% [13].

Random Forest

Random forest adalah metode klasifikasi dan regresi berdasarkan agregasi sejumlah besar pohon keputusan [10]. *Random forest classifier* adalah pengklasifikasi *ensemble* yang menggunakan satu set CART untuk membuat prediksi [11]. *Random forest* juga merupakan kumpulan ratusan hingga ribuan pohon, dimana setiap pohon ditanam menggunakan sampel *bootstrap* dari data asli [12]. Pohon dibuat dengan menggambar *subset* sampel pelatihan melalui penggantian (pendekatan *bagging*) [11]. Sampel yang sama dapat dipilih beberapa kali, sementara yang lain tidak dapat dipilih sama sekali [11]. Sekitar dua pertiga dari sampel digunakan untuk melatih pohon dengan sepertiga sisanya digunakan dalam persilangan internal [11]. Teknik validasi untuk memperkirakan seberapa baik kinerja model *random forest* yang dihasilkan [11]. Berikut adalah persamaan untuk membuat model klasifikasi *random forest*: 1



Gambar 1. Arsitektur *Random Forest*

Table 1. Nilai Parameter Untuk *Hyperparameter Tuning*

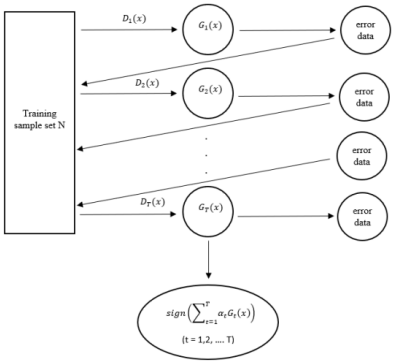
Method	Parameters	Ranges
Random Forest	N estimators	[200, 300, 400, 500]
	Min_samples _leaf	[2, 3, 4, 5]
	Min_samples _split	[4, 6, 8, 10]
	Criterion	[‘gini’, ‘entropy’]
AdaBoost	N estimators	[150, 200, 250, 500]
	Learning _rate	[0.1, 1.0]
	Algorithm	[‘SAMME’, ‘SAMME.R’]
XGBoost	N estimators	[200, 300, 400, 500]
	Learning _rate	[0.1]
	Max_depth	[6, 7, 8]

Adaptive Boosting

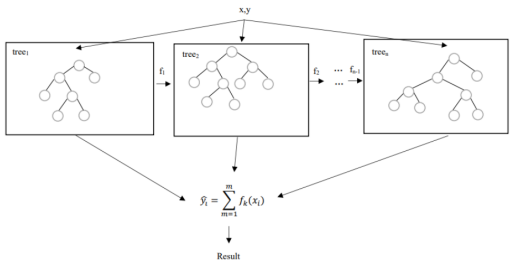
Adaptive Boosting (*AdaBoost*) adalah salah satu algoritma klasifikasi boosting yang dapat meningkatkan sekelompok *classifier* ‘lemah’ menjadi *classifier* ‘kuat’ [13]. Ide utama *AdaBoost* adalah untuk membangun suksesi pelajar yang lemah dengan menggunakan set pelatihan yang berbeda yang berasal dari resampling data asli [14]. Algoritma ini biasanya menggunakan algoritma klasifikasi dasar yang kemampuan klasifikasinya lebih baik daripada *random guest* untuk melatih pengklasifikasi dasar dari sampel pelatihan awal [13]. Kemudian sesuaikan semua sampel pelatihan yang memiliki bobot yang sama [15]. Sampel kesalahan pengklasifikasi sebelumnya akan digunakan untuk melatih pengklasifikasi berikutnya, yaitu probabilitas memilih sampel kesalahan yang benar untuk masuk ke pengklasifikasi lemah berikutnya, dan kemungkinan memilih sampel berpasangan akan berkurang

Table 2. *Confussion Matrix*

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



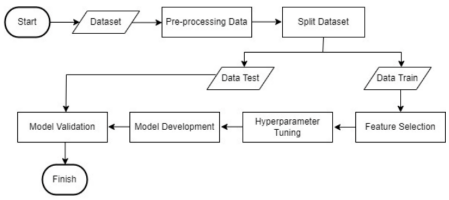
Gambar 2. S*Flowchart Adaboost*



Gambar 3. Arsitektur *Xgboost*

Table 3. Nilai Akurasi Pada Tahap Seleksi Fitur

Model	Jumlah Fitur	Akurasi
Random Forest 1 (RF1)	30	0.989
Random Forest 2 (RF2)	40	0.989
Random Forest 3 (RF3)	50	0.989
AdaBoost 1 (AB1)	10	0.968
AdaBoost 2 (AB2)	20	0.968
AdaBoost 3 (AB3)	120	0.957
XGBoost 1 (XG1)	10	0.989
XGBoost 2 (XG2)	20	0.979
XGBoost 3 (XG3)	30	0.979



Gambar 4. Alur Pemodelan Program.

[15]. Pendekatan ini memungkinkan *AdaBoost* untuk fokus pada sampel yang telah salah diklasifikasikan [15]. Berikut adalah *flowchart* dari *AdaBoost*: 2

Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) adalah sistem penambah pohon gradien yang dioptimalkan untuk membuat pohon keputusan dalam bentuk sekuensial [20]. *XGBoost* juga merupakan salah satu aplikasi dari algoritma *Gradient Boosting* (GB) yang berbasis pohon keputusan sebagai *classifier* [16]. *XGBoost* memiliki kemampuan untuk menghitung perhitungan yang relevan relatif lebih cepat di semua lingkungan komputasi [17]. Performa model *XGBoost* diakui secara luas di beberapa tantangan penambangan data dan pembelajaran mesin [22]. *Algoritma XGBoost* dimulai dengan pendekatan berbasis *decision tree* dimana representasi grafis dari solusi yang mungkin untuk keputusan dihitung tergantung pada kondisi tertentu [17]. Kemudian, *algoritma meta ensemble* yang menggabungkan prediksi dari berbagai *decision tree* berdasarkan teknik voting mayoritas disebut '*bagging*' [17]. Pendekatan bagging ini selanjutnya berkembang untuk membangun agregasi dari *decision tree* dengan memilih fitur secara acak [17]. Performa model ditingkatkan dengan mengurangi kesalahan dari membangun model sekuensial [17]. Berikut persamaan pada *algoritma XGBoost*: 3

Metodologi Penelitian

Sistem yang Dibangun

Sistem yang dibangun pada penelitian ini adalah model prediksi NSCLC dengan *variance threshold* dan parameter *chi-square* untuk seleksi fitur dan *ensemble* untuk membangun model prediksi. Proses ini dimulai dengan dataset kemudian data di *preprocessing* dan displit menjadi dua yaitu *data test* dan data train. Pada data train proses selanjutnya yaitu *feature selection*, diproses ini data akan mengalami dua tahap seleksi fitur yaitu *variance threshold dan chi-square*. Setelah itu akan dilakukan proses *hyperparameter tuning*, dilanjutkan dengan model *development* dan terakhir model *validation*. Untuk data test akan langsung masuk pada proses model *validation*. Alur pemodelan program pada penelitian tugas akhir ini dapat dilihat pada diagram dibawah ini: 4

Data Set

Data yang digunakan dalam penelitian ini berasal dari dataset GEO dengan kode GSE adalah GSE19804 [18]. Dataset berupa 54.677 ekspresi gen yang terdiri dari 120 sampel dengan dua kelas, yaitu 60 sampel *lung cancer* dan 60 sampel normal [18]. Data dibagi menjadi dua yaitu data uji dan data latih dengan perbandingan 75:25.

Seleksi Fitur

Seleksi fitur yang digunakan pada penelitian ini yaitu *variance threshold* dan parameter *chi-square*. *Variance threshold* adalah pendekatan dasar sederhana untuk pemilihan fitur [19]. *Variance threshold* menghapus semua fitur yang variansnya tidak memenuhi ambang batas tertentu [19]. Secara *default*, *variance threshold* menghapus semua fitur varian nol yaitu fitur yang memiliki nilai yang sama di semua sampel [19]. Penelitian ini menggunakan nilai *variance threshold* 0.1 yang diimplementasikan menggunakan *library sklearn*. Persamaan yang digunakan untuk mencari nilai *varians* dari masing-masing fitur adalah:

$$\sigma^2 = \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad (1)$$

dimana:

n = jumlah data
xi = fitur
x = rata-rata dari suatu fitur

Setelah dilakukan *threshold*, jumlah fitur menjadi 497 fitur dan selanjutnya akan dilakukan pemilihan fitur menggunakan parameter *chi-square* dan iterasi sebanyak 30 kali. *Chi-square* adalah salah satu jenis uji komparatif non parametris yang dilakukan oleh dua variabel, dimana skala data kedua variabel adalah nominal [20]. Uji *chi-square* digunakan untuk melihat ketergantungan antara variabel bebas dan variabel tergantung berskala nominal atau ordinal [20]. Prosedur uji *Chi-Square* membuat tabulasi satu atau variabel ke dalam kategori-kategori dan menghitung angka statistik *chi-square* [20]. Untuk satu variabel disebut sebagai uji keselarasan atau *goodness of fit tes*t yang berfungsi untuk membandingkan frekuensi yang diamati dengan frekuensi yang diharapkan [20]. Jika terdiri dari 2 variabel disebut sebagai uji independensi yang berfungsi untuk hubungan dua variabel [20]. Parameter *chi-square* yang digunakan pada penelitian ini adalah implementasi *library sklean*. Hasil dari pemilihan fitur kemudian dimasukkan ke dalam model random *forest, adaboost,* dan *xgboost*. Hasilnya berupa akurasi berdasarkan fiturnya. Nilai akurasi terbaik digunakan untuk membangun model akhir. Berikut adalah rumus dari *chi-square*.

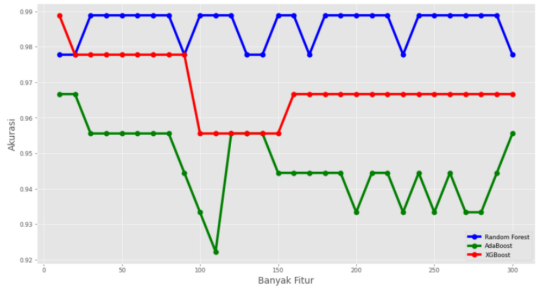
$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)}{E_i} \quad (2)$$

dimana:

X2 = nilai *chi-square*
Oi = frekuensi yang diperoleh atau diamati
Ei = frekuensi yang diharapkan

Model Prediksi

Pembangunan model dilakukan dengan dua macam yaitu model *hyperparameter* tuning dan model tanpa hyperparameter tuning. Model-model tersebut menggunakan metode *random forest, adaboost, dan xgboost*. Pada modell tuning menggunakan *grid search cross validation (Grid Search CV)* dengan cv sebanyak 10 kali untuk mencari parameter terbaik. Metode *random forest* menggunakan parameter n_estimataors, min_samples_leaf, min_samples_split, dan criterion. *Adaboost* menggunakan parameter *n_estimators, learning_rate*, dan *algorithm*. Sedangkan *xgboost* menggunakan parameter *n_estimators, learning_rate,* dan *max_depth*. Detail lebih rinci untuk parameter yang digunakan bisa dilihat pada tabel 1. Setelah selesai dilakukan pembangunan model dengan hyperparameter tuning, selanjutnya lakukan perbandingan antara model *hyperparameter tuning* dengan model tanpa *hyperparameter tuning*. Setelah pembangunan model selesai maka dilanjutkan dengan validasi model.



Gambar 5. Hasil Akurasi Untuk Tahap Seleksi Fitur.

Validasi Model

Untuk mengevaluasi kinerja model prediksi menggunakan parameter *confussion matrix*. *Confussion matrix* adalah suatu metode yang digunakan untuk melakukan perhitungan kinerja sistem dievaluasi menggunakan data dalam *matrix* [26]. Pada tahap ini peneliti menggunakan *confusion matrix* untuk menghitung nilai dari *accuracy, precision, recall* dan *f1-score*. tabel *confusion matrix* dapat dilihat pada tabel.Berikut persamaan dari *accuracy, precision, recall*, dan *f1-score*.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 - Score = \frac{2x\ (\text{ Precision x Recall })}{Precision + Recall} \quad (6)$$

Hasil dan Pembahasan

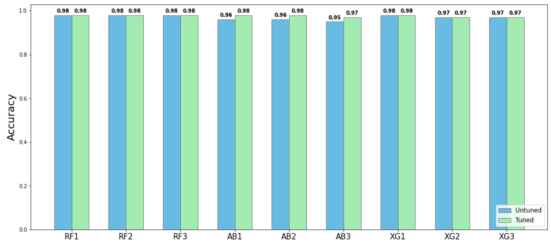
Seleksi Fitur

Penelitian ini menggunakan beberapa algoritma pada model yang dibangun, algoritma tersebut adalah random *forest, adaboost,* dan *xgboost*. Model ini kemudian akan dibandingkan pada model pengujian penggunaan hyperparameter tuning. Sebelum dapat membangun model tersebut, dilakukan sebuah pengujian pada model dengan tujuan untuk mendapatkan jumlah fitur optimal pada masing-masing model. Hasil model pada masing-masing algoritma yang dapat dilihat pada Gambar 3. Hasil pengujian pada Gambar 3 dilakukan dengan menggunakan parameter *chi-square* pada masing- masing model dan model ini dibangun tanpa menggunakan *hyperparameter* tuning. Pada gambar 5 tersebut rata- rata hasil akurasi tertinggi didapat pada model *random forest* dan rata-rata hasil terendah didapat pada model *adaboost*. Grafik tersebut menunjukkan nilai akurasi terhadap jumlah fitur dengan rentang fitur 0-300.

Model *random forest* mengalami kenaikan dan penurunan secara stabil, model *adaboost* mengalami penurunan secara drastis pada jumlah fitur 110 dan kenaikan drastis pada jumlah fitur 120 dan 300, sedangkan *xgboost* mengalami penurunan pada jumlah fitur 20 dan 100, serta mengalami kenaikan pada jumlah fitur 110 dan sisanya stabil. Berdasarkan hasil pengujian tersebut diambil 3 jumlah fitur yang memiliki nilai akurasi tertinggi, sehingga masing-masing model dengan algoritma yang berbeda memiliki 3 jenis model dengan jumlah fitur yang berbeda. Hasil model yang diambil untuk dilakukan pengujian pada tahap berikutnya dapat dilihat pada tabel 3. Berdasarkan tabel 3 model dengan nilai akurasi terbaik didapat dengan nilai 0.989 yaitu model *random forest* 1, 2, 3, dan xgboost 1 dengan masing-masing jumlah fitur

Table 4. Hasil Parameter Terbaik *Hyperparameter Tuning*

Method	Parameters	Best Parameter Value		
		RF1	**RF2**	**RF3**
Random Forest	N estimators	200	200	200
	Min_samples_leaf	2	2	2
	Min_samples_split	4	4	4
	Criterion	'gini'	'gini'	'gini'
		AB1	**AB2**	**AB3**
AdaBoost	N estimators	150	150	150
	Learning_rate	0.1	0.1	0.1
	Algorithm	'SAMME'	'SAMME'	'SAMME.R'
		XG1	**XG2**	**XG3**
Xgboost	N estimators	200	200	200
	Learning_rate	0.1	0.1	0.1
	Max_depth	6	6	6



Gambar 6. Perbandingan Hasil Akurasi Model Dengan *Hyperparameter Tuning* Dan Tanpa *Hyperparameter Tuning*.

30, 40, 50, dan 10. Hal ini dapat disimpulkan bahwa untuk hasil akurasi yang tinggi tidak terlalu berpengaruh terhadap jumlah fitur yang banyak.

Hyperparameter Tuning

Pada pengujian ini, model yang digunakan adalah 9 model dengan jumlah fitur yang optimal yang didapatkan pada tahap pengujian sebelumnya. Selanjutnya, pada tahap ini akan dilakukan pembangunan model dengan mengaplikasikan hyperparameter tuning pada 9 model, dan kemudian model yang dibangun dengan *hyperparameter tuning* ini akan dibandingkan performanya pada model yang tanpa menggunakan *hyperparameter tuning*. Digunakan beberapa konfigurasi pada pengaplikaisan *hyperparameter tuning* yang dapat dilihat pada tabel 6. Pada model *Random Forest* memiliki hasil yang sama untuk semua parameter model. Pada model *AdaBoost* peneliti menemukan bahwa parameter algoritma untuk model AB1 dan AB2 adalah 'SAMME' dan model AB3 adalah 'SAMME.R', sedangkan parameter n_estimator, dan *learning rate* memiliki hasil yang sama. *XGBoost* menghasilkan parameter yang sama untuk semua model.

Dari pengujian yang dilakukan didapatkan hasil yang dapat dilihat pada Gambar 6. Pada gambar tersebut dapat dilihat bahwa proses *hyperparameter tuning* pada model RF1, RF2, RF3, XG1, XG2, dan XG3 tidak memiliki perubahan nilai pada model tanpa *hyperparameter tuning* dan dengan hyperparameter tuning. Sedangkan pada model dengan algoritma *adaboost* baik model AB1, AB2, dan AB3 terdapat kenaikan nilai akurasi sebesar 0.02% setelah diaplikasikan *hyperparameter tuning* pada model yang dibangun. Hal ini bisa terjadi dikarenakan beberapa faktor yaitu pemilihan parameter yang kurang tepat dan banyaknya parameter yang digunakan sehingga *gridsearchcv* yang digunakan pada *hyperparameter tuning* kesulitan karena dimensi yang terlalu tinggi.

Validasi Model

Untuk mendapatkan hasil yang sesuai pada model yang dibangun, dilakukan validasi model yang mencakup nilai T*rue Positive, False Positive, False Negative, True Negative, Accuracy, Precision, Recall,* dan *F1-Score* terhadap data latih dan data uji. Hasil validasi model dapat dilihat pada tabel 5 dan tabel 6 Berdasarkan tabel 5 hasil validasi model pada data latih nilai *accuracy* untuk model dengan algoritma *adaboost* dan *xgboost* memiliki nilai sebesar 1.00 sedangkan untuk model dengan algoritma *random forest* memiliki nilai *accuracy* sebesar 0.99. Pada nilai *precision* yang terkena kanker memiliki nilai yang sama yaitu sebesar 1.00 untuk model *random forest*, *adaboost,* dan *xgboost*, sedangkan untuk yang normal model *random forest* memiliki nilai sebesar 0.98, model *adaboost* dan *xgboost* memiliki nilai sebesar 1.00.

Kebalikan dari nilai *precision*, nilai *recall* untuk yang normal yaitu sebesar 1.00, sedangkan untuk yang terkena kanker model *random forest* memiliki nilai sebesar 0.98, model *adaboost* dan *xgboost* memiliki nilai sebesar 1.00. Nilai *f1- score* yang terkena kanker dan yang normal untuk model *random forest* memiliki nilai sebesar 0.99 dan untuk model *adaboost* dan *xgboost* memiliki nilai sebesar 1.00. Selanjutnya hasil validasi model pada data uji yang dapat dilihat pada tabel 6, untuk nilai *precision* yang terkena kanker pada model random *forest, adaboost,* dan *xgboost* memiliki nilai yang sama yaitu sebesar 0.87. Untuk nilai *accuracy, recall*, dan *f1-score* berdasarkan yang terkena kanker nilai terbaik didapatkan oleh model AB1, XG1, dan XG2 dengan nilai yang sama berturut-turut yaitu sebesar 0.93, 1.00, dan 0.93 karena pada saat *hyperparameter tuning* menggunakan parameter yang berbeda dan jumlah fitur yang digunakan juga berbeda.

Kesimpulan

Pada penelitian, penulis telah membangun model untuk memprediksi nsclc pada wanita bukan perokok dengan menggunakan metode *ensemble*. Pemilihan fitur berhasil dilakukan dengan menggunakan *variance threshold* dan *parameter chi-square*. Pembangunan model dilakukan dengan dua macam yaitu model *hyperparameter tuning* dan

Table 5. Hasil Validasi Model Pada Data Latih

Model	T P	F P	F N	T N	Accuracy	Precision		Recall		F1-score	
						Normal	Kanker	Normal	Kanker	Normal	Kanker
RF1	44	1	0	45	0.99	0.98	1.00	1.00	0.98	0.99	0.99
RF2	44	1	0	45	0.99	0.98	1.00	1.00	0.98	0.99	0.99
RF3	44	1	0	45	0.99	0.98	1.00	1.00	0.98	0.99	0.99
AB1	45	0	0	45	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AB2	45	0	0	45	1.00	1.00	1.00	1.00	1.00	1.00	1.00
AB3	45	0	0	45	1.00	1.00	1.00	1.00	1.00	1.00	1.00
XG1	45	0	0	45	1.00	1.00	1.00	1.00	1.00	1.00	1.00
XG2	45	0	0	45	1.00	1.00	1.00	1.00	1.00	1.00	1.00
XG3	45	0	0	45	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 6. Hasil Validasi Model Pada Data Uji

Model	T P	F P	F N	T N	Accuracy	Precision		Recall		F1-score	
						Normal	Kanker	Normal	Kanker	Normal	Kanker
RF1	13	2	2	13	0.87	0.87	0.87	0.87	0.87	0.87	0.87
RF2	13	2	2	13	0.87	0.87	0.87	0.87	0.87	0.87	0.87
RF3	13	2	2	13	0.87	0.87	0.87	0.87	0.87	0.87	0.87
AB1	15	0	2	13	0.93	1.00	0.87	0.88	1.00	0.94	0.93
AB2	14	1	2	13	0.90	0.93	0.87	0.88	0.93	0.90	0.90
AB3	13	2	2	13	0.86	0.87	0.87	0.87	0.87	0.87	0.87
XG1	15	0	2	13	0.93	1.00	0.87	0.88	1.00	0.94	0.93
XG2	15	0	2	13	0.93	1.00	0.87	0.88	1.00	0.94	0.93
XG3	13	2	2	13	0.86	0.87	0.87	0.87	0.87	0.87	0.87

model tanpa *hyperparameter tuning*. Proses *hyperparameter tuning* yang menghasilkan kenaikan akurasi yaitu model *it*, sedangkan model random *forest* dan *xgboost* menghasilkan nilai akurasi yang sama dengan model tanpa *hyperparameter tuning*. Evaluasi model pada data uji penulis menemukan tiga model terbaik berdasarkan yang terdiagnosis kanker yaitu model AB1 dengan 10 fitur, XG1 dengan 10 fitur, dan XG2 dengan 20 fitur dan nilai untuk *accuracy, recall*, dan *f1-score* berturut-turut adalah 0.93, 1.00, dan 0.93. Untuk penelitian selanjutnya, disarankan memilih algoritma lainnya, hal ini dikarenakan metode *hyperparameter tuning* yang dibangun menggunakan algoritma *random forest* dan *xgboost* tidak memiliki peningkatan sama sekali.

Daftar Pustaka

1. Wang H, Zhou Z, Li Y, Chen Z, Lu P, Wang W, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images. EJNMMI Research. 2017 jan;7(1). Available from: https://doi.org/10.1186%2Fs13550-017-0260-9.
2. Feature Selection for Microarray Gene Expression Data using Simulated Annealing guided by the Multivariate Joint Entropy — arxiv.org;. [Accessed 04-08-2023]. http://arxiv.org/abs/1302.1733.
3. Chen YC, Chang YC, Ke WC, Chiu HW. Cancer adjuvant chemotherapy strategic classification by artificial neural network with gene expression data: An example for non-small cell lung cancer. Journal of Biomedical Informatics. 2015 aug;56:1-7. Available from: https://doi.org/10.1016%2Fj.jbi.2015.05.006.
4. Kirienko M, Sollini M, Corbetta M, Voulaz E, Gozzi N, Interlenghi M, et al. Radiomics and gene expression profile to characterise the disease and predict outcome in patients with lung cancer. European Journal of Nuclear Medicine and Molecular Imaging. 2021 may;48(11):3643-55. Available from: https://doi.org/10.1007%2Fs00259-021-05371-7.
5. Machine learning highlights the deficiency of conventional dosimetric constraints for prevention of high-grade radiation esophagitis in non-small cell lung cancer treated with chemoradiation — reader.elsevier.com;. [Accessed 04-08-2023]. https://reader.elsevier.com/reader/sd/pii/S2405630820300203?token=FF5A59A64847D17F354292C4969AD2BD3D7A3BC522586F6E9EBB23EB976CE012D9173A854079FEC6EDA2E412655002C0&originRegion=eu-west-1&originCreation=20211216072710.
6. Le NQK, Kha QH, Nguyen VH, Chen YC, Cheng SJ, Chen CY. Machine Learning-Based Radiomics Signatures for EGFR and KRAS Mutations Prediction in Non-Small-Cell Lung Cancer. International Journal of Molecular Sciences. 2021 aug;22(17):9254. Available from: https://doi.org/10.3390%2Fijms22179254.
7. metatags generator. View of Implementation of Ensemble Method in Schizophrenia Identification Based on Microarray Data — http;. [Accessed 04-08-2023]. http://www.jurnal.iaii.or.id/index.php/RESTI/article/view/3788/535.
8. Attribute selection using fuzzy roughset based customized similarity measure for lung cancer microarray gene expression data — reader.elsevier.com;. [Accessed 04-08-2023]. https://reader.elsevier.com/reader/sd/pii/S2314728817300338?token=1B

641F7FB1B6ABC4944AA3773C288461FCB216FFCE372E92EF8BA21A332D3977EED982A97C362E7ED341ACDEF9BDC243&originRegion=eu-west-1&originCreation=20211214042630.

9. Hawkins SH, Korecki JN, Balagurunathan Y, Gu Y, Kumar V, Basu S, et al. Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features. IEEE Access. 2014;2:1418-26. Available from: https://doi.org/10.1109%2Faccess.2014.2373335.
10. Boulesteix AL, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012 oct;2(6):493-507. Available from: https://doi.org/10.1002%2Fwidm.1072.
11. Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing. 2016 apr;114:24-31. Available from: https://doi.org/10.1016%2Fj.isprsjprs.2016.01.011.
12. Chen X, Ishwaran H. Random forests for genomic data analysis. Genomics. 2012 jun;99(6):323-9. Available from: https://doi.org/10.1016%2Fj.ygeno.2012.04.003.
13. Feature Learning Viewpoint of Adaboost and a New Algorithm — ieeexplore.ieee.org;. [Accessed 04-08-2023]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8868178(accessedJul.25,2022).
14. Cao J, Kwong S, Wang R. A noise-detection based AdaBoost algorithm for mislabeled data. Pattern Recognition. 2012 dec;45(12):4451-65. Available from: https://doi.org/10.1016%2Fj.patcog.2012.05.002.
15. Zhang Y, Ni M, Zhang C, Liang S, Fang S, Li R, et al. Research and Application of AdaBoost Algorithm Based on SVM. In: 2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE; 2019. Available from: https://doi.org/10.1109%2Fitaic.2019.8785556.
16. Abdurrahman G, Sintawati M. Implementation of xgboost for classification of parkinson's disease. Journal of Physics: Conference Series. 2020 may;1538(1):012024. Available from: https://doi.org/10.1088%2F1742-6596%2F1538%2F1%2F012024.
17. Bhattacharya S, S SRK, Maddikunta PKR, Kaluri R, Singh S, Gadekallu TR, et al. A Novel PCA-Firefly Based XGBoost Classification Model for Intrusion Detection in Networks Using GPU. Electronics. 2020 jan;9(2):219. Available from: https://doi.org/10.3390%2Felectronics9020219.
18. geo. GEO2R - GEO - NCBI — https;. [Accessed 04-08-2023]. https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE19804(accessedDec.16,2021).
19. Gupta A. Feature Selection Techniques in Machine Learning (Updated 2023) — analyticsvidhya.com;. [Accessed 04-08-2023]. https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/.
20. Deteksi Rumor pada Twitter Menggunakan Metode Multilayer Perceptron — openlibrary.telkomuniversity.ac.id;. [Accessed 04-08-2023]. https://openlibrary.telkomuniversity.ac.id/home/catalog/id/164216/slug/deteksi-rumor-pada-twitter-menggunakan-metode-multilayer-perceptron.html.