Komputika: Jurnal Sistem Komputer
Volume 14, Nomor 1, April 2025, hlm. 79 - 85
DOI: 10.34010/kp5h2h96

p-ISSN: 2252-9039
e-ISSN: 2655-3198

Perbandingan Klasifikasi Penyakit Kanker Paru-Paru Menggunakan Decision Tree Dan Random Forest

Dwi Cahya Julia Kartikasari[1], Rabiatul Adawiyah[2*]

Program Studi Statistika, Fakultas Sains dan Teknologi, Universitas PGRI Adi Buana Surabaya
Jl. Dukuh Menanggal XII, Dukuh Menanggal, Kec. Gayungan, Surabaya, Jawa Timur Indonesia 60234

email: rabiatuladawiyah@unipasby.ac.id

(Naskah masuk: *04 Februari 2025*; direvisi: *25 April 2025*; diterima untuk diterbitkan: *tgl. bulan tahun*)

ABSTRAK *– Kanker paru-paru merupakan salah satu penyebab utama kematian di dunia, sehingga deteksi dini pada kanker paru-paru sangat penting. Penelitian ini membandingkan metode klasifikasi Decision Tree dan Random Forest dalam mendeteksi kanker paru-paru menggunakan 1.010 sampel dengan 7 atribut. Penanganan missing value di terapkan menggunakan metode imputasi modus. Penelitian ini dilakukan beberapa tahapan, diantaranya tahapan pengumpulan data, klasifikasi awal tanpa seleksi fitur, analisis feature importance, serta klasifikasi ulang setelah pemilihan fitur. Sehingga diperoleh Analisis feature importance menunjukkan bahwa Coughing, Chronic Disease, Smoking, dan Shortness of Breath adalah fitur paling berpengaruh. Dari analisis tersebut dapat diketahui Hasil penelitian menunjukkan bahwa Decision Tree tanpa seleksi fitur memberikan akurasi tertinggi sebesar 64,85%, sedangkan Random Forest hanya mencapai 52,62%, keduanya mengalami penurunan akurasi menjadi 55,94% dan 52,47% setelah seleksi fitur. Hasil ini berbeda dari penelitian sebelumnya oleh Idris, J. dkk yang menggunakan dataset dengan sembilan atribut dan 30.000 sampel, di mana metode Random Forest memperoleh akurasi tertinggi sebesar 97,48% dan metode Decision Tree memperoleh akurasi sebesar 95,16%. Hal ini mengindikasikan bahwa Decision Tree lebih efektif dalam menangkap pola data tanpa seleksi fitur, sedangkan Random Forest kurang optimal untuk dataset kecil. Kontribusi utama dari penelitian ini adalah memberikan pemahaman mengenai efektivitas metode klasifikasi dan peran pemilihan fitur dalam sistem deteksi dini kanker paru-paru berbasis machine learning.*

Kata Kunci *– Machine Learning; Klasifikasi; Feature Importance; Entropi; Gain.*

Comparison Of Lung Cancer Classification Using Decision Tree And Random Forest

ABSTRACT *– Lung cancer is a leading cause of deaths in the world, making early detection critically important. This study compares the performance of Decision Tree and Random Forest classification methods in detecting lung cancer using 1,010 samples with 7 attributes. Missing value handling is applies using mode imputation. This research was carried out in several stages, included data collection, initial classification without feature selection, feature importance analysis, and reclassification after selecting key features. So that the feature importance analysis identified Coughing, Chronic Disease, Smoking, and Shortness of Breath as the most influential attributes. From the analysis. It can be seen that results show that Decision Tree without feature selection achieved the highest accuracy of 64.85%, while Random Forest reached only 52.62%, both of which experienced accuracy dropped to 55.94% and 52.47% After feature selection. These findings contrast with a previous study by Idris, J. et al., which used a dataset with nine attributes and 30,000 samples, where Random Forest achieved the highest accuracy of 97.48% and Decision Tree 95.16%. This indicates that Decision Tree is more effective in capturing patterns in smaller datasets without feature selection, while Random Forest may be less optimal in such contexts. The main contribution of this study is to provide insights into the effectiveness of classification methods and the role of feature selection in early lung cancer detection using machine learning.*

Keywords *– Machine Learning; Classification; Feature Importance; Entropy; Gain.*

1. PENDAHULUAN

Kanker adalah penyakit mematikan yang menyerang berbagai negara. Menurut Organisasi Kesehatan Dunia, kanker menyebabkan 9,6 juta kematian di seluruh dunia [12]. *National Cancer Institute* mendefinisikan kanker sebagai penyakit genetik akibat perubahan gen yang mengendalikan fungsi sel, terutama dalam pertumbuhan dan pembelahan. Sel kanker secara cepat menyerang jaringan tubuh, selain itu juga bisa membentuk tumor, dan menyebabkan gangguan pada fungsi tubuh [1]. Penyakit ini dapat tumbuh di berbagai organ, seperti payudara, paru-paru, prostat, dan ginjal [7].

Selain itu, Menurut *Global Cancer Observatory* WHO, kanker paru-paru merupakan penyebab utama kematian akibat kanker di berbagai usia, baik pria maupun wanita. Terjadinya kanker paru-paru ditandai oleh pertumbuhan sel tidak terkendali di jaringan paru-paru, yang dapat menyebar ke organ lain melalui metastasis [9]. Faktor risiko kanker paru-paru meliputi paparan asap rokok, usia, faktor genetik, gas radon, dan polusi udara. Sulitnya deteksi dini menjadikan kanker paru-paru sangat mematikan [8]. Oleh karena itu, proses klasifikasi pada penelitian ini dilakukan untuk dapat membantu deteksi dini mengenali pola kanker paru-paru.

Klasifikasi adalah proses penggolongan objek berdasarkan karakteristik tertentu [3]. Algoritma seperti Decision Tree dan *Random Forest* sering digunakan untuk klasifikasi. *Decision Tree* adalah metode populer yang membagi data menjadi kelompok lebih kecil melalui pohon keputusan. Sementara itu, *Random Forest* adalah algoritma berbasis *ensemble* yang menggabungkan beberapa *Decision Tree* dengan pemilihan fitur acak, sehingga lebih akurat, tahan terhadap *outliers*, dan efisien dalam penyimpanan data [11]. Kedua metode ini banyak digunakan dalam klasifikasi data medis, termasuk deteksi kanker paru-paru.

Penelitian oleh Idris, J. dkk [5] pada dataset prediksi terkena penyakit paru-paru dengan sembilan atribut dan 30.000 sampel data menunjukkan bahwa *Random Forest* memiliki akurasi tertinggi dalam klasifikasi kanker paru-paru dengan nilai 97,48%, dibandingkan *Decision Tree* sebesar 95,16%. Metode lain, seperti SVM, K-NN, *Naïve Bayes*, dan *Logistic Regression*, memiliki akurasi lebih rendah. Berdasarkan penelitian terdahulu, penelitian ini dilakukan untuk mengkaji lebih lanjut mengenai efektivitas algoritma *Decision Tree* dan *Random Forest* dalam mengklasifikasikan data, khususnya data kanker paru-paru, sehingga dapat membantu mengembangkan sistem pendeteksian dini yang lebih akurat dan efektif.

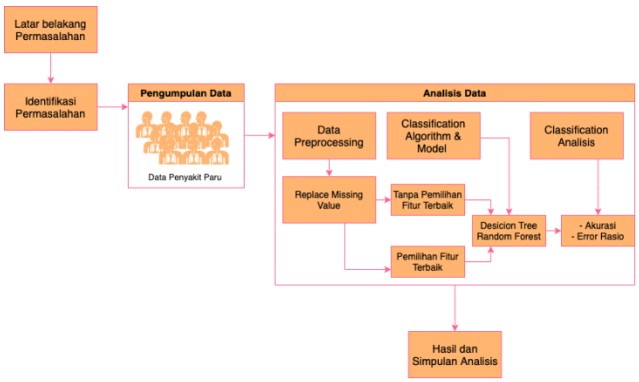
Penelitian ini membandingkan klasifikasi metode *Decision Tree* dan *Random Forest* sebelum dan sesudah pemilihan fitur untuk mengetahui perbedaan dalam tingkat akurasi. Selain itu, penelitian ini menganalisis pengaruh pemilihan fitur terhadap tingkat akurasi klasifikasi dan mengidentifikasi fitur-fitur yang memiliki pengaruh tertinggi terhadap klasifikasi kanker paru-paru pada kedua metode tersebut. Batasan dalam penelitian ini meliputi penggunaan atribut dataset kanker paru-paru, yaitu jenis kelamin, batuk, merokok, penyakit kronis, alkohol, sesak napas, dan kanker paru-paru, dengan teknik imputasi modus untuk menangani missing value dan memastikan kualitas data sebelum analisis lebih lanjut.

Sebagai solusi, penelitian ini menggabungkan metode *Decision Tree* dan *Random Forest* untuk klasifikasi kanker paru-paru berdasarkan atribut yang telah ditentukan. Akurasi kedua metode akan dibandingkan untuk menentukan efektivitas. Penelitian ini diharapkan dapat membantu dalam memahami perbedaan metode klasifikasi serta peran pemilihan fitur dalam meningkatkan akurasi klasifikasi, sehingga dapat menjadi referensi bagi mahasiswa dalam mengembangkan sistem klasifikasi khususnya mendeteksi dini pola data dalam bidang kesehatan.

2. METODE DAN BAHAN

Tahapan Penelitian

Tahapan penelitian yang dilakukan tercantum pada Gambar 1 :



Gambar 1. Diagram Blok Tahapan Penelitian

Tahapan penelitian ini dimulai dengan latar belakang persamalahan, selanjutnya identifikasi masalah penyakit paru-paru untuk menentukan fokus dan tujuan yang jelas. kemudian, dilakukan pengumpulan dataset terkait penyakit paru-paru serta penanganan data yang mengandung *missing value* guna memastikan kualitas data yang optimal. Setelah itu, dilakukan proses klasifikasi menggunakan metode *Decision Tree* dan *Random Forest* tanpa pemilihan fitur terbaik. Kemudian, klasifikasi kembali dilakukan dengan kedua metode

tersebut setelah melalui tahap pemilihan fitur terbaik. Hasil dari setiap model dilakukan perbandingkan akurasi. Pada tahap akhir, penelitian ini ditarik kesimpulan berdasarkan hasil klasifikasi yang telah dilakukan. Pada gambar 1 merupakan gambaran tahapan penelitian yang dilakukan.

Pengumpulan Data

Pada Gambar 2 terdapat data penelitian yang digunakan :

	GENDER	COUGHING	SMOKING	CHRONIC_DISEASE	ALCOHOL_CONSUMING	SHORTNESS_OF_BREATH	LUNG_CANCER
0	NaN	0.0	1.0	0.0	0.0	0.0	0.0
1	NaN	1.0	1.0	1.0	1.0	1.0	0.0
2	NaN	1.0	0.0	1.0	1.0	0.0	1.0
3	NaN	0.0	0.0	0.0	1.0	1.0	1.0
4	NaN	1.0	1.0	1.0	1.0	1.0	0.0
5	1.0	NaN	1.0	0.0	1.0	1.0	1.0
6	1.0	NaN	1.0	0.0	0.0	0.0	1.0
7	1.0	NaN	0.0	1.0	1.0	1.0	0.0
8	0.0	NaN	1.0	1.0	0.0	1.0	0.0
9	1.0	NaN	0.0	1.0	0.0	0.0	0.0

Gambar 2. Data Awal Penelitian

Penelitian ini menggunakan data yang diperoleh dari situs kaggle, yang terdiri dari 1.006 baris dan 7 atribut yakni Gender, Coughing, Smoking, Chronic Disease, Alcohol Consuming, Shortness of Breath. Atribut Lung Cancer memiliki dua kategori, di mana nilai 1 menunjukkan bahwa pasien menderita penyakit paru-paru, sedangkan nilai 0 menunjukkan bahwa pasien tidak memiliki penyakit paru-paru.

Decision Tree

Decision Tree adalah suatu struktur berbasis proses yang bersifat sekuensial. Proses dimulai dari akar, kemudian dilanjutkan dengan evaluasi terhadap suatu fitur dan pemilihan salah satu dari dua cabang yang tersedia. Langkah ini terus berlanjut hingga mencapai cabang terakhir, atau yang disebut sebagai daun, yang umumnya merepresentasikan target akhir yang dicari [2]. Pemilihan atribut sebagai akar didasarkan pada atribut dengan nilai gain tertinggi di antara semua atribut yang tersedia [10]. Nilai Gain dapat diperoleh ketika nilai semua Entropi sudah didapatkan, kemudian dihitung. Nilai gain dihitung menggunakan rumus yang ditunjukkan sebagai berikut [4].

$$Gain(S,A) = Entropy(S)\sum_{i=1}^{n}\frac{|S_i|}{|S|} * Entropy(S_i) \quad (1)$$

Entropi digunakan untuk mengukur seberapa informatif suatu atribut dalam menghasilkan sebuah keputusan [14]. Adapun rumus Entropi adalah sebagai berikut:

$$Entropy(S) = \sum_{i=1}^{n} -(pi) * log_2 pi \quad (2)$$

Random Forest

Random Forest adalah kumpulan dari beberapa pohon keputusan (*Decision Tree*) yang dibangun dengan sampel yang dipilih secara acak, namun dengan aturan pembagian simpul yang berbeda [15]. Model ini bekerja dengan menggunakan subset fitur pada setiap pohon dan mencari ambang batas optimal untuk memisahkan data. Hasilnya adalah sekumpulan pohon yang dilatih dengan pendekatan yang lebih lemah, di mana masing-masing pohon menghasilkan prediksi yang berbeda.

Kanker Paru-Paru

Kanker paru-paru adalah jenis kanker yang berkembang di organ paru-paru akibat perubahan yang tidak normal pada sel. Penyakit ini dapat dipicu oleh berbagai faktor, termasuk kebiasaan merokok, paparan polusi udara, alergi terhadap debu, faktor genetik, gaya hidup, dan faktor lainnya.

3. HASIL DAN PEMBAHASAN

Penanganan *Missing Value*

Pada data yang digunakan terdapat *missing value* secara acak dengan porsi 5 data di semua atribut. Tabel 1 menunjukkan hasil penanganan *Missing Value* setiap atribut.

Tabel 1. Penanganan *Missing Value* Tiap Atribut

Atribut	Missing Value	
	Sebelum Penanganan	Setelah Penanganan
Gender	5	0
Coughing	5	0
Smoking	5	0
Chronic Disease	5	0
Alcohol Consuming	5	0
Shortness Of Breath	5	0
Lung Cancer	5	0

Jenis *missing value* pada data ini adalah MCAR (*Missing Completely at Random*). *Missing value* yang ditangani dengan menggunakan metode modus, yakni mengisi *missing value* dengan kategori yang sering muncul menggunakan fitur *replace* dengan bantuan microsoft excel.

Klasifikasi Sebelum Pemilihan Fitur

Pada tahap ini, dilakukan proses klasifikasi kanker paru-paru menggunakan model *Random Forest* dan *Decision Tree* tanpa melalui tahap pemilihan fitur. Seluruh fitur yang tersedia dalam dataset digunakan untuk membangun model guna mengevaluasi performanya secara keseluruhan. Tujuan dari analisis ini adalah untuk melihat bagaimana kedua metode bekerja dengan semua

fitur yang ada serta membandingkan tingkat akurasinya sebelum dilakukan optimasi melalui seleksi fitur.

Klasifikasi *Decision Tree*

Pada proses ini, model *Decision Tree* dibangun dengan menggunakan seluruh fitur yang tersedia dalam dataset, yaitu *Gender, Coughing, Chronic Disease, Alcohol Consuming ,Smoking, dan Shortness of Breath*. Tujuan dari penggunaan semua fitur ini adalah untuk melihat bagaimana setiap atribut memengaruhi proses klasifikasi dan pengambilan keputusan oleh model. Dalam algoritma *Decision Tree*, setiap fitur akan dievaluasi berdasarkan nilai entropi dan information gain untuk menentukan fitur yang paling berpengaruh dalam klasifikasi.

Tabel 2. Nilai Enteropi *Decision Tree*

Atribut	kategori	Entropi
Gender	M	0,9961
	F	0,9991
Coughing	Yes	0,9991
	No	0,9967
Smoking	Yes	0,9964
	No	0,9992
Chronic Disease	Yes	0,9972
	No	0,9987
Alcohol Consuming	Yes	0,9996
	No	0,9949
Shortness of Breath	Yes	0,9955
	No	0,9995

Tabel 2 menunjukkan nilai entropi dari masing-masing atribut. Selanjutnya dilakukan pencarian nilai gain untuk semua setiap atribut.

Tabel 3. Nilai Gain *Decision Tree*

Atribut	Gain
Alcohol Consuming	0,0006
Shortness Of Breath	0,0004
Smoking	0,0002
Gender	0,0002
Coughing	0,0001
Chronic Disease	0,000006

Tabel 3 menunjukkan *Alcohol Consuming* memiliki nilai gain tertinggi, sehingga fitur ini akan dipilih sebagai akar pada pohon keputusan untuk pemisahan data menjadi beberapa cabang atau ranting. Dari jumlah data sebanyak 1010, data dibagi menjadi dua pembagian dengan persentase 80% untuk data *training* dan 20% untuk data *testing*. Dari pembagian ini, 808 sebagai data *training* dan 202 data *testing*. Hasil evaluasi menunjukkan bahwa model memiliki akurasi sebesar 64,85% pada data testing yang ditunjukkan oleh gambar 3.

```
Tabel Evaluasi Decision Tree:
      Metrik     Nilai
0    Akurasi  0.648515
1  Precision  0.649730
2     Recall  0.649862
3   F1 Score  0.648506
```

Gambar 3. Nilai Akurasi Klasifikasi *Decision Tree*

Artinya, model mampu mengklasifikasikan data dengan benar sebesar 64,85% dari total data uji. Akurasi ini tergolong sedang, yang menunjukkan bahwa performa model cukup baik.

Klasifikasi *Random Forest*

Proses klasifikasi dengan *random forest* mencakup pembagian data menjadi *training* dan *testing*, analisis *feature importance*, serta perbandingan akurasi model sebelum dan sesudah pemilihan fitur. Data dibagi menjadi dua bagian utama dengan rasio 80:20, di mana 808 data dengan 6 fitur digunakan untuk melatih model, dan 202 data dengan 6 fitur digunakan untuk menguji kinerja model.

```
Tabel Evaluasi Random Forest:
      Metrik     Nilai
0    Akurasi  0.524752
1  Precision  0.262376
2     Recall  0.500000
3   F1 Score  0.344156
```

Gambar 4. Nilai Akurasi Random Forest

Terlihat pada Gambar 4 bahwa Model *random forest* mencapai akurasi terbaik sebesar 52,47%, menunjukkan bahwa kemampuan prediksi model masih kurang optimal. Hal ini kemungkinan disebabkan oleh kualitas data atau fitur yang digunakan. *Feature importance* digunakan untuk mengidentifikasi fitur-fitur yang memiliki pengaruh terbesar dalam menentukan hasil prediksi. Dengan mengetahui fitur yang paling berpengaruh, model dapat dioptimalkan melalui pemilihan fitur terbaik.

Tabel 4. *Feature Importance Random Forest*

Fitur	***Important***
Chronic Disease	0,263
Smoking	0,219
Shortness of Breath	0,200
Alcohol Consuming	0,156
Gender	0,092
Coughing	0,067

Berdasarkan Tabel 4, fitur *chronic disease* memiliki

pengaruh terbesar dalam proses klasifikasi kanker paru-paru, dengan nilai kepentingan (*Important*) sebesar 0,263. Hal ini menunjukkan bahwa riwayat penyakit kronis merupakan indikator utama dalam menentukan kemungkinan seseorang menderita kanker paru-paru. Selanjutnya, fitur *smoking* memiliki nilai kepentingan 0,219, yang mengindikasikan bahwa bahwa kebiasaan merokok merupakan faktor risiko yang signifikan. Fitur *shortness of breath* atau sesak napas menempati posisi ketiga dengan nilai 0,200, menunjukkan bahwa gejala ini juga memiliki peran penting dalam mendeteksi kanker paru-paru. *Alcohol Consuming* memiliki nilai kepentingan 0,156, diikuti oleh Gender dengan nilai 0,092, menunjukkan bahwa riwayat konsumsi alkohol dan jenis kelamin memiliki kontribusi yang relatif lebih rendah dalam klasifikasi. Sementara itu, fitur *coughing* memiliki pengaruh paling rendah dalam klasifikasi, dengan nilai 0,067, menunjukkan bahwa gejala batuk memiliki hubungan yang lebih lemah dibandingkan faktor-faktor lainnya dalam mendeteksi kanker paru-paru. Sehingga dipilih 4 fitur dengan nilai *important* tertinggi, fitur tersebut adalah *chronic disease, smoking, shortness of breath, dan alcohol consuming*.

Klasifikasi Setelah Pemilihan Fitur

Setelah didapatkan *feature importance*, hanya fitur-fitur dengan kontribusi tertinggi yang dipilih untuk meningkatkan efektivitas model klasifikasi. Pemilihan fitur ini bertujuan untuk mengurangi dimensi data, meningkatkan efisiensi komputasi, serta mengoptimalkan akurasi prediksi.

Klasifikasi *Decision Tree*

Pada proses ini, model klasifikasi *decision tree* dibangun dengan menggunakan 4 fitur tertinggi yang memiliki pengaruh paling signifikan dalam klasifikasi kanker paru-paru. Pemilihan fitur ini dilakukan berdasarkan perhitungan nilai entropi dan gain.

Tabel 5. Nilai Enteropi *Decision Tree*

Atribut	Kategori	Entropi
Chronic Disease	Yes	0,9971
	No	0,9986
Smoking	Yes	0,9963
	No	0,9991
Shortness Of Breath	Yes	0,9954
	No	0,9994
Alcohol Consuming	Yes	0,9995
	No	0,9949

Tabel 5 menunjukkan nilai entropi dari masing-masing atribut. Selanjutnya dilakukan pencarian nilai *gain* untuk semua setiap atribut.

Tabel 6. Nilai Gain *Decision Tree*

Atribut	Gain
Chronic Disease	0,00006
Smoking	0,00023
Shortness Of Breath	0,00049
Alcohol Consuming	0,00061

Tabel 6 menunjukkan *alcohol consuming* memiliki nilai gain tertinggi, sehingga fitur ini akan dipilih sebagai akar pada pohon keputusan untuk pemisahan data menjadi beberapa cabang atau ranting. Dari jumlah data sebanyak 1010, data dibagi menjadi dua dengan persentase 80% untuk data training dan 20% untuk data testing. Hasil evaluasi pada Gambar 5, menunjukkan bahwa model memiliki akurasi sebesar 51.48%.

```
Tabel Evaluasi Decision Tree:
      Metrik     Nilai
0    Akurasi  0.514851
1  Precision  0.513199
2     Recall  0.513168
3   F1 Score  0.513133
```

Gambar 5. Nilai Akurasi Klasifikasi *Decision Tree* Setelah Pemilihan Fitur

Artinya, model mampu mengklasifikasikan data dengan benar sebesar 51.48% dari total data *testing*. Akurasi ini tergolong sedang, yang menunjukkan bahwa performa model masih perlu ditingkatkan

Klasifikasi *Random Forest*

Setelah dilakukan pemilihan fitur berdasarkan *feature importance*, model klasifikasi *random forest* digunakan kembali namun hanya menggunakan fitur yang memiliki pengaruh terbesar dalam data kanker paru-paru. Langkah ini bertujuan untuk meningkatkan akurasi model dengan menghilangkan fitur yang kurang relevan, sehingga analisis menjadi lebih efisien dan interpretasi hasil lebih optimal.

```
Tabel Evaluasi Random Forest:
      Metrik     Nilai
0    Akurasi  0.509901
1  Precision  0.479887
2     Recall  0.491254
3   F1 Score  0.415997
```

Gambar 6. Nilai Akurasi Klasifikasi *Random Forest* Setelah Pemilihan Fitur

Setelah seleksi fitur berdasarkan nilai importance tertinggi, data tetap dibagi dengan rasio 80:20 yang menghasilkan akurasi model sedikit mengalami penurunan menjadi 50,99% seperti terlihat pada Gambar 6, hal ini menunjukkan bahwa pengurangan fitur tidak secara signifikan meningkatkan performa

model pada model random forest.

Perbandingan Klasifikas*i Random Forest* dan *Decision Tree*

Setelah melakukan klasifikasi menggunakan random forest dan decision tree, dilakukan perbandingan akurasi masing-masing model dalam mendeteksi kanker paru-paru. Perbandingan ini mencakup akurasi model sebelum dan sesudah pemilihan fitur.

Tabel 7. Perbandingan Akurasi Model

Metode	Sebelum Pemilihan Fitur		Setelah Pemilihan Fitur	
	Decision Tree	*Random Forest*	*Decision Tree*	*Random Forest*
Akurasi	64,85%	52,47%	51,48%	50,99%
Precision	64,97%	26,23%	51,31%	47,98%
Recall	64,99%	50,00%	51,31%	49,12%
F1 Score	64,85%	34,41%	51,31%	41,59%

Tabel 7 menunjukkan perubahan kinerja metode *Decision Tree* dan *Random Forest* sebelum dan setelah pemilihan fitur. Pada model *Decision Tree*, terjadi penurunan yang signifikan pada semua metrik evaluasi setelah pemilihan fitur. Akurasi menurun dari 64,85% menjadi 51,48%, *Precision* dari 64,97% menjadi 51,31%, *Recall* dari 64,99% menjadi 51,31%, dan *F1 Score* dari 64,85% menjadi 51,31%. Hal ini mengindikasikan bahwa pengurangan fitur mengurangi kemampuan model *Decision Tree* untuk menangkap informasi penting, sehingga menurunkan kinerjanya secara keseluruhan. Pada model *Random Forest*, juga terjadi penurunan kinerja setelah pemilihan fitur, meskipun dengan pola yang berbeda. Akurasi menurun dari 52,47% menjadi 50,99%, namun *Precision* justru meningkat secara signifikan dari 26,23% menjadi 47,98%, sementara *Recall* sedikit menurun dari 50,00% menjadi 49,12%. *F1 Score* model *Random Forest* meningkat dari 34,41% menjadi 41,59% setelah pemilihan fitur, yang menunjukkan keseimbangan yang lebih baik antara *Precision* dan *Recall*. Hasil ini mengindikasikan bahwa pemilihan fitur memiliki dampak yang lebih besar pada model *Decision Tree* dibandingkan *Random Forest*. Meskipun kedua model mengalami penurunan akurasi, namun pemilihan fitur justru meningkatkan keseimbangan antara *Precision* dan *Recall* pada model *Random Forest*, yang tercermin dari peningkatan nilai *F1 Score*. Hal ini menunjukkan bahwa pemilihan fitur membantu model *Random Forest* untuk mengurangi jumlah *false positive*, meskipun akan menurunkan nilai akurasi.

4. KESIMPULAN

Data kanker paru-paru memiliki 7 atribut dan 1010 baris, serta terdapat *missing value* sebanyak 5 data pada masing-masing atribut. Jenis *missing value* pada data ini adalah MCAR (*Missing Completely at Random*) dan ditangani dengan metode modus. *Feature importance* dengan model *random forest* menghasilkan 4 fitur dengan nilai important tertinggi, fitur tersebut adalah *chronic disease, smoking, shortness of breath, dan alcohol consuming*. Model Klasifikasi menunjukkan *decision tree* tanpa pemilihan fitur memberikan performa terbaik dengan akurasi 64,85%, *precision* 64,97%, *recall* 64,99%, dan *F1-Score* 64,85%, sedangkan *Random Forest* hanya mencapai akurasi 52,47%, *precision* 26,23%, *recall* 50,00%, dan *F1-Score* 34,41%. Pemilihan 4 fitur terpenting dengan Random Forest dan diklasifikasikan menggunakan *decision tree* menurunkan akurasi menjadi 51,48%(precision 51,31%, recall 51,31%, *F1-Score 51,31%*). Hasil ini menunjukkan bahwa *Decision Tree* lebih efektif dalam menangkap pola data yang penting tanpa seleksi fitur dibandingkan dengan *Random Forest* dan metode *Random Forest* cenderung kurang optimal apabila menggunakan dataset yang relatif kecil.

Sehingga dari penelitian yang diperoleh dapat ditarik simpulan, untuk penelitian selanjutnya direkomendasikan menggunakan sampel data yang lebih besar dan beragam agar model *ensemble* seperti *Random Forest* serta *Decission Tree* dapat memaksimalkan keunggulannya dalam mengenali pola data yang kompleks sehingga dapat meningkatkan akurasi dan ketahanan model dalam mendeteksi kanker paru-paru.

UCAPAN TERIMA KASIH

Akhir dari penelitian ini terimakasih pada segala pihak yang terlibat.

DAFTAR PUSTAKA

[1] Aqila, A., & Faisal, M. (2023). Lung Cancer EDA Classification Using the Decision Trees Method in Python. *Informatics and Software Engineering, 1*(1), 8-13.

[2] Depari, D. H., Widiastiwi, Y., & Santoni, M. M. (2022). Perbandingan Model Decision Tree, Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung. *Informatik: Jurnal Ilmu Komputer, 18*(3), 239-248.

[3] Desiani, A., Maiyanti, S. I., Andriani, Y., Suprihatin, B., Amran, A., Marselina, N. C., & Salsabila, A. (2023). Perbandingan Klasifikasi Penyakit Kanker Paru-Paru menggunakan Support Vector Machine dan K-Nearest Neighbor. *Jurnal PROCESSOR, 18*(1).

[4] Hafizan, H., & Putri, A. N. (2020). Penerapan Metode Klasifikasi Decision Tree Pada Status Gizi Balita Di Kabupaten Simalungun. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen), 1*(2), 68-72.

[5] Idris, J. F., Ramadhani, R., & Mutoffar, M. M. (2024). Klasifikasi Penyakit Kanker Paru Menggunakan Perbandingan Algoritma Machine Learning. *Jurnal Media Akademik (JMA), 2*(2).

[6] Purba, W., Wardani, S., Lumbantoruan, D. F., Celia, F., Silalahi, I., & Edison, T. L. (2023). Optimization Of Lung Cancer Classification Method Using Eda-Based Machine Learning. *6*(2), 43–50.

[7] Putra, H. W. N. S., Atina, V., & Maulindar, J. (2023). Penerapan Algoritme Decision Tree Pada Klasifikasi Penyakit Kanker Paru-Paru. *Jutisi: Jurnal Ilmiah Teknik Informatika dan Sistem Informasi, 12*(3).

[8] Rifai, A., & Prabowo, Y. (2022). Diagnosis Kanker Paru-Paru dengan Sistem Fuzzy. *Krea-TIF: Jurnal Teknik Informatika, 10*(1), 19-28.

[9] Rofiani, R., Oktaviani, L., Vernanda, D., & Hendriawan, T. (2024). Penerapan Metode Klasifikasi Decision Tree dalam Prediksi Kanker Paru-Paru Menggunakan Algoritma C4.5. *Jurnal Tekno Kompak, 18*(1), 126-139.

[10] Rosandy, T. (2016). Perbandingan Metode Naive Bayes Classifier Dengan Metode Decision Tree (C4.5) Untuk Menganalisa Kelancaran Pembiayaan (Study Kasus: KSPPS/BMT Al-Fadhila). *Jurnal Teknologi Informasi Magister, 2*(01), 52-62.

[11] Sari, L., Romadloni, A., & Listyaningrum, R. (2023). Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest. *Infotekmesin, 14*(1), 155-162.

[12] Septhya, D., Rahayu, K., Rabbani, S., Fitria, V., Rahmaddeni, R., Irawan, Y., & Hayami, R. (2023). Implementation of Decision Tree Algorithm and Support Vector Machine for Lung Cancer Classification. *MALCOM: Indonesian Journal of Machine Learning and Computer Science, 3*(1), 15-19.

[13] Tarigan, L. R. A., & Dahlan, D. (2024). Optimalisasi Fitur Dengan Forward Selection Pada Estimasi Tingkat Penyakit Paru-Paru Menggunakan Algoritma Klasifikasi Random Forest. *JATI (Jurnal Mahasiswa Teknik Informatika), 8*(5), 10341-10348.

[14] Kamagi, D. H., & Hansun, S. (2014). Implementasi Data Mining dengan Algoritma C4. 5 untuk Memprediksi Tingkat Kelulusan Mahasiswa. *Ultimatics: Jurnal Teknik Informatika*, *6*(1), 15-20.

[15] Sinambela, D. P., Naparin, H., Zulfadhilah, M., & Hidayah, N. (2023). Implementasi Algoritma Decision Tree dan Random Forest dalam Prediksi Perdarahan Pascasalin. *Jurnal Informasi dan Teknologi*, 58-64.