Jurnal Ilmiah Teknik Informatika dan Komunikasi
Vol. 3 No. 2 Juli 2023
E-ISSN: 2827-7945 / P-ISSN: 2827-8127, Hal 27-41
DOI: https://doi.org/10.55606/juitik.v3i2.472

Penerapan Data Mining Dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest

Arifin Yusuf Permana , Hari Noer Fazri , M.Fakhrizal Nur Athoilah
Mohammad Robi ,Ricky Firmansyah
ARS University
Alamat: Jalan Sekolah Internasional no 1 40282 Kiaracondong Jawa Barat

Korespondensi Penulis : arifinyusufp447@gmail.com

Abstract. *Lung cancer is one of the one of the leading causes of death in the world. From this data there are several categories of people who are positive and negative for lung cancer, Here the researcher will display information on the exact number of people who contracted lung cancer from the data, and in this study using the Random Forest algorithm because Random Forest This research uses the Random Forest algorithm because Random Forest has a data set selection process. Has a data set selection process. to improve the performance of classification model. With feature selection, Random Forest can certainly work efficiently on big data with complex parameters, which will greatly facilitate the classification of positive and negative lung cancer patients. Observations will be a reference for analyzing the prognosis of lung disease. Observation will be a reference for analyzing the prognosis of lung disease here how the application of data data mining techniques on the prediction analysis of lung cancer analysis and how performance of the random forest algorithm in predicting lung cancer.by applying data mining techniques and has been tested using a survey dataset of lung cancer survey dataset and using software called Rapidminer toanalyze and predict positive patients with lung cancer It was concluded that the It is concluded that the Random Forest algorithm that has obtained the greatest accuracy obtained accuracy results worth 90.61% with an AUC value of 0.941.*

Keywords *Confusion Matrix, Data Mining, Lung Cancer, Optimize Selection, Random Forest.*

Abstrak. Kanker paru-paru adalah salah satu penyebab utama kematian di dunia. dari data tersebut ada beberapa kategori orang yang positif serta negative terjangkit kanker paru-paru, disini peneliti akan menampilkan informasi jumlah pasti dari orang yang terjangkit penyakit kanker paru dari data tersebut, serta dalam penelitian ini menggunakan algoritma Random Forest karena Random Forest memiliki proses pemilihan kumpulan data. untuk meningkatkan kinerja model klasifikasi. Dengan pemilihan fitur, Random Forest tentunya dapat bekerja secara efisien pada big data dengan parameter yang kompleks, yang akan sangat memudahkan pengklasifikasian penderita kanker paru positif dan negatif. Observasi akan menjadi acuan untuk analisis prognosis penyakit paru-paru disini bagaimana penerapan teknik data mining pada analisis prediksi analisis kanker paru serta bagaimana kinerja algoritma random forest dalam memprediksi kanker paru.Dengan menerapkan Teknik data mining dan telah diuji menggunakan dataset survey kanker paru serta memakai perangkat lunak yang bernama Rapidminer untuk menganalisis dan memprediksi penderita positif terjangkit kanker paru dan negatif tidak terjangkit kanker paru berhasil.Disimpulkan bahwa algoritma Random Forest yang telah mendapatkan akurasi terbesar diperoleh hasil akurasi senilai 90.61% dengan nilai AUC 0.941.

Kata kunci: Data Mining, Evolutionary, Kanker Paru, Optimize Selection, Random Forest.

Received April 30, 2023; Revised Mei 02, 2023; Accepted Juni 09, 2023
* Arifin Yusuf Permana, arifinyusufp447@gmail.com

LATAR BELAKANG

Kanker paru-paru adalah salah satu penyebab utama kematian di dunia. Kematian akibat kanker adalah yang kedua setelah penyakit kardiovaskular di Amerika Serikat dan Inggris. 2 Kanker Menurut WHO, kanker trakea, bronkus, dan paru-paru memiliki angka kematian spesifik penyebab (cause specific death date/CSD) sebesar 13,2 per 100.000 orang dan nilai PMR sebesar 2,3% (Halim, 2020).

Kanker disebabkan oleh perkembangan abnormal dari sel-sel jaringan manusia yang disebabkan oleh kelainan genetik. Pergeseran ini dapat dikaitkan dengan tiga variabel utama: faktor genetik, faktor gaya hidup, dan faktor karsinogenik atau zat yang menimbulkan sel kanker. Faktor karsinogenik dihasilkan oleh stres oksidatif, yang disebabkan oleh radikal bebas dan spesies oksigen reaktif (ROS) yang dihasilkan oleh aktivitas metabolisme dalam tubuh, serta paparan polutan dari luar tubuh. Tubuh membutuhkan antioksidan yang tepat untuk menangkal kerusakan oksidatif. Daun kelor (Moringa Oleifera) mengandung antioksidan dan bahan kimia bioaktif dengan konsentrasi tinggi, yang dapat membantu mencegah stres oksidatif dan kanker (Kusmardika, 2020).

Faktor risiko kanker paru salah satunya adalah merokok. Sebagian besar kematian akibat kanker paru dikaitkan oleh merokok. Paparan asap rokok yang berlebihan, baik perokok aktif maupun pasif, polusi udara, dan paparan lingkungan. Kanker paru-paru berhubungan dengan tempat kerja. Batuk darah, suara serak, batuk, nyeri dada, abses paru, sesak napas adalah gejala umum kanker paru. Kemoterapi, terapi radiasi, dan pembedahan dapat digunakan untuk mengobati kanker paru-paru. Karena hampir semua saluran udara terletak di dada, kanker paru dapat diobati dengan operasi toraks. Operasi toraks adalah operasi yang dilakukan pada bagian dada (Ramadhaniaha, 2020).

Disini peneliti menggunakan data sekunder dengan metode kuantitatif yang diperoleh secara online dari website yang menyediakan kumpulan data yang disebut kaggle, dari data tersebut ada beberapa kategori orang yang positif serta negative terjangkit kanker paru-paru, disini peneliti akan menampilkan informasi jumlah pasti dari orang yang terjangkit penyakit kanker paru dari data tersebut, serta dalam penelitian ini menggunakan algoritma Random Forest karena Random Forest memiliki proses pemilihan kumpulan data. untuk meningkatkan kinerja model klasifikasi. Dengan pemilihan fitur, Random Forest tentunya dapat bekerja secara efisien pada big data dengan parameter yang kompleks, yang akan sangat memudahkan pengklasifikasian penderita kanker paru positif dan negatif.

Sesuai menggunakan rumusan kasus yang sudah dibahas diatas, maka tujuan penelitian ini yaitu adalah pertama tujuan dari penelitian ini adalah untuk menguji dan mengevaluasi kemampuan teknik data mining dalam memprediksi kanker paru.

Dalam hal ini, penelitian bertujuan untuk mengeksplorasi penerapan teknik data mining pada analisis prediksi kanker paru dengan tujuan meningkatkan kemampuan diagnosis dan pengobatan kanker paru. Tujuan kedua dari penelitian ini adalah untuk mengevaluasi kinerja algoritma Random Forest dalam memprediksi kanker paru. Dalam hal ini, penelitian bertujuan untuk menguji efektivitas dan efisiensi algoritma Random Forest dalam memprediksi kanker paru, serta mengevaluasi kekuatan dan kelemahan dari algoritma tersebut. Tujuan utama dari penelitian ini adalah untuk meningkatkan kemampuan diagnosis dan pengobatan kanker paru dengan menggunakan algoritma Random Forest sebagai alat bantu analisis prediksi (Khasanah, 2019).

Fokus pada prediksi kanker paru: Jurnal ini berfokus pada masalah prediksi kanker paru dengan menggunakan algoritma Random Forest. Oleh karena itu, jurnal ini tidak membahas masalah prediksi penyakit lainnya atau penggunaan algoritma yang berbeda. Keterbatasan algoritma Random Forest: Jurnal ini hanya membahas kelebihan algoritma Random Forest dalam memprediksi kanker paru. Oleh karena itu, jurnal ini tidak membahas keterbatasan atau kelemahan algoritma tersebut. Keterbatasan algoritma Random Forest: Jurnal ini hanya membahas kelebihan algoritma Random Forest dalam memprediksi kanker paru. Oleh karena itu, jurnal ini tidak membahas keterbatasan atau kelemahan algoritma tersebut.

KAJIAN TEORITIS

Data mining merupakan proses menggunakan suatu teknik statistik, matematika, dan kecerdasan buatan untuk mengekstrak dan mengidentifikasi informasi dan pengetahuan (atau pola) dari kumpulan data yang sangat besar. Sederhananya, data mining adalah proses penggalian informasi dari data yang ada. Penggalian informasi pada penelitian ini berupa klasifikasi.

Dengan demikian, data mining dapat dinamakan penambangan pengetahuan atau penemuan pengetahuan. Terlepas dari ketidaksesuaian makna dan terminologi ini, data mining telah muncul sebagai metode yang lebih disukai oleh masyarakat. Banyak istilah lain untuk data mining termasuk ekstraksi pengetahuan, analisis pola, arkeologi data, pengumpulan informasi, dan penemuan pola. Kumpulan fakta yang telah dicatat sebagai data, atau sesuatu

yang tidak berharga. Entitas yang tidak berarti dan tidak dikenali. Namun, penambangan adalah tindakan menambang. Sebagai hasilnya, data mining dapat digambarkan sebagai proses penambangan data yang menghasilkan output berbasis pengetahuan. Pengetahuan adalah bentuk (output). Tujuan dari proses klasifikasi adalah untuk mengatur item dan variabel data dengan mengklasifikasikan objek sesuai dengan atribut data yang akan digunakan (Haristu, 2019).

Klasifikasi merupakan teknik data mining yang memetakan data ke kelompok yang telah di terapkan. Untuk memprediksi nilai kelas objek yang tidak diketahui, klasifikasi adalah proses mengidentifikasi sekumpulan pola yang cukup mewakili kelas data. Kita perlu memvalidasi data pelatihan untuk mendapatkan model. Pada saat yang sama, tingkat akurasi dan model yang dibuat dievaluasi menggunakan data uji. Peramalan nama atau nilai objek data dapat dilakukan dengan menggunakan klasifikasi (Idris, 2019).

Random Forest merupakan evolusi dari metode pohon keputusan menggunakan beberapa pohon keputusan, di mana setiap pohon keputusan telah dilatih menggunakan sampel individu dan setiap atribut didekomposisi menjadi pohon yang dipilih dari atribut subset acak. Random Forest memiliki sejumlah keunggulan, dapat meningkatkan akurasi hasil jika data hilang dan menolak outlier, serta efisien untuk penyimpanan data. Selanjutnya, Random Forest memiliki proses seleksi fitur yang mampu memilih fitur terbaik untuk meningkatkan performa model klasifikasi. Dengan pemilihan fitur, Random Forest tentunya dapat bekerja secara efisien pada big data dengan parameter yang kompleks (Mulyahati, 2020).

Algoritma random forest ini dapat digunakan untuk banyak dimensi dengan skala berbeda dan kinerja yang kuat. Penggunaan algoritma ini memprediksi pasien berisiko tinggi, memprediksi kegagalan suku cadang dalam produksi, memprediksi kegagalan pembayaran pinjaman, dan lainnya (Prakoso, 2019).

Jenis Algoritma Evolusi (EA) yang paling umum adalah algoritma genetika. Dengan evolusi teknologi informasi yang cepat, algoritma genetik berevolusi. Karena kemampuannya untuk mengatasi berbagai macam masalah yang kompleks. Operator pengoptimalan berbobot (Evolving) adalah operator yang memiliki sub-proses. Subproses operator Optimalkan Bobot (evolusi) harus selalu mengembalikan vektor daya. Operator Optimalkan Bobot (Evolusi) menghitung bobot atribut tertentu dalam kalimat contoh menggunakan evolutionary atau genetika, semakin tinggi sebuah bobot atribut, semakin krusial atribut tersebut dipertimbangkan (Saputra, 2019).

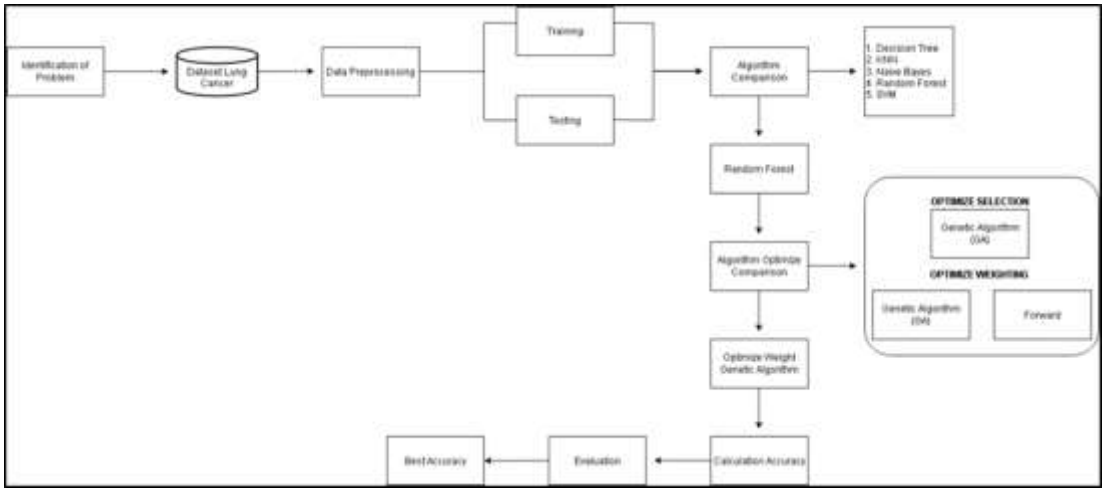
Algoritma Genetika adalah alat pencarian ilmu komputer yang digunakan untuk mengidentifikasi perkiraan jawaban atas masalah pengoptimalan dan pencarian. mengatasi masalah pengoptimalan dan pencarian. Algoritme genetika adalah bagian dari algoritme evolusioner yang menggunakan prinsip-prinsip biologi evolusioner seperti keturunan, mutasi, seleksi alam, dan rekombinasi (atau persilangan). Algoritma genetika (AG) adalah algoritma pencarian yang didasarkan pada seleksi alam dan mekanisme genetika. Algoritma Genetika adalah salah satu algoritma yang sangat cocok untuk digunakan dalam menangani masalah optimasi yang kompleks yang sulit dipecahkan dengan menggunakan pendekatan konvensional (Jonathan, 2019).

Sebagai salah satu algoritma dari metode tersebut, algoritma genetika atau evolutionary (GA) dapat memberikan solusi yang andal agar model random forest yang diusulkan menjadi lebih baik dan optimal. Algoritma genetika atau evolutionary saat ini dapat digunakan untuk menemukan bobot terbaik temukan bobot terbaik dalam model RF. Namun GA adalah model yang digunakan oleh banyak ahli sebelumnya sebagai model yang dapat memberikan akurasi yang lebih besar dalam kumpulan data kanker paru-paru.

METODE PENELITIAN

1. Identifikasi Masalah

Observasi akan menjadi acuan untuk analisis prognosis penyakit paru-paru disini bagaimana penerapan teknik data mining pada analisis prediksi analisis kanker paru serta bagaimana kinerja algoritma random forest dalam memprediksi kanker paru, disini peneliti akan menampilkan informasi jumlah pasti dari orang yang terjangkit penyakit kanker paru dari data tersebut. Hasil pengolahan data kanker paru menjadi informasi dan pengetahuan yang diharapkan, dari mana deteksi yang lebih baik atau pengetahuan potensial dapat dibuat atau lebih akurat dalam membaca data, waktu yang akurat dan tepat waktu dari data ini untuk dapat melakukan analisis prediktif kanker paru dan menemukan peluang baru dan membuat rencana strategis serta untuk menganalisis dan memprediksi kanker paru, selain itu dapat digunakan sebagai alat untuk menginformasikan pengambilan keputusan. Dengan membuat klasifikasi dan meningkatkan akurasi dengan bantuan optimasi dan algoritma genetika atau evolutionary atau evolutionary dengan alur penelitian yang terdapat pada (gambar 1).



Gambar 1. Metode Penelitian

2. Pengumpulan Data

Dataset yang digunakan adalah kanker paru yang diambil dari website Kaggle, efektivitas sistem prediksi kanker membantu orang mengambil keputusan terbaik berdasarkan pengetahuan mereka tentang risiko kanker dengan biaya serendah mungkin. Informasi yang dikumpulkan dari situs web Sistem Prognostik Kanker Paru online. Dataset kanker paru ini memiliki jumlah memiliki jumlah atribut enam belas dengan jumlah contoh data sebanyak dua ratus delapan puluh empat.

3. Preprocessing

Validasi split adalah pendekatan validasi yang memisahkan data secara acak terdiri dari beberapa yaitu data uji dan data pelatihan. Pelatihan tes ini berdasarkan split ratio yang telah ditentukan akan dilakukan menggunakan split validation, dan data pelatihan split ratio yang tersisa kemudian akan digunakan sebagai data uji. Data tersebut dapat digunakan untuk pembelajaran yang disebut data latih, sedangkan data yang akan digunakan untuk memverifikasi keakuratan atau kebenaran hasil pembelajaran disebut sebagai data uji. Nilai akurasi dan AUC untuk dataset yang digunakan dihitung dengan menggunakan rasio perbandingan 0.5 hingga 0.9.

Teknik statistik yang disebut cross-validation (CV) dapat digunakan untuk menilai seberapa baik kinerja model atau algoritme ketika data dibagi menjadi dua subset: data pelatihan dan data validasi/evaluasi. Dengan menggunakan subset pelatihan dan subset validasi, model atau algoritme dilatih. Teknik validasi X-Fold, V-Fold, dan Cross validation digunakan untuk menentukan nilai akurasi dan AUC untuk dataset yang dipertimbangkan.

4. Komparasi Algoritma

Ada banyak algoritma yang digunakan dalam berbagai bidang. Dalam penelitian ini perbandingan algoritma kami menggunakan lima algoritma diantaranya Random Forest,Super Vector Machine(SVM), Decision Tree,, K-Nearest Neighbors (KNN).

5. Algoritma Random Forest

Algoritma random forest terpilih karena menghasilkan algoritma terbaik dari lima algoritma lainnya. Dengan memakai teknik split validation dengan rasio 0,5 hingga 0,9.

6. Komparasi Optimasi

Pada tahap komparasi disini peneliti menggunakan dua tipe optimasi yaitu optimize feature dan optimize weight. Untuk optimasi feature peneliti memakai evolusionary dan untuk optimasi weight peneliti memakai 2 algoritma yaitu menggunakan evolutionary dan forward.

7. Optimize Selection

Optimize selection dan weight evolutionary mendapatkan nilai akurasi terbaik maka dari itu dipilihlah diantara beberapa optimasi lainnya.

8. Evaluasi

Confusion Matrix adalah sebuah tolak ukur yang digunakan oleh para peneliti untuk mengevaluasi efektivitas teknik machine learning, khususnya supervised learning(Chicco, 2021). Confusion matrix adalah representasi dari data yang dihasilkan oleh estimasi algoritma ML dan skenario nyata. Kita dapat menghitung presisi, akurasi, recall, dan spesifisitas dengan menggunakan confusion matrix. Dalam tulisan ini, saya membahas bagaimana cara mengevaluasi output dari algoritma ML. Tulisan ini didasarkan pada tulisan tersebut dan termasuk sebuah studi kasus. Saya kemudian menulis ulang dalam bahasa saya sendiri untuk meningkatkan pemahaman.

Peneliti menjelaskan Akurasi, Presisi, Recall, dan Spesifisitas dengan menggunakan skenario kasus prediksi kanker paru-paru Drop Out (DO). Kinerja sistem dinilai dengan menggunakan Precision dan Recall. Presisi adalah kesesuaian antara informasi yang telah diambil dengan informasi yang dibutuhkan. Recall mengukur seberapa baik sebuah sistem menemukan informasi. Tingkat kesamaan antara nilai yang diperoleh dengan nilai yang

sebenarnya dikenal sebagai akurasi. Confusion matrik dapat digunakan untuk menghitung presisi, recall, dan akurasi.

Tabel 1. Confusion Matrix

	PREDICTION	
ACTUAL	True	False
True (Positive)	TP	FN
False (Negative)	FP	TN

Sebagai representasi dari hasil proses klasifikasi, ada 4 (empat) istilah yang digunakan dalam pengukuran kinerja confusion matrix. Keempat istilah tersebut adalah True Positive (TP), True Negative (TN), False Positive (FP), dan False Negative (FN). Jumlah data negatif yang terdeteksi secara akurat adalah angka True Negative (TN), sedangkan false positive (FP) adalah contoh di mana data negatif secara keliru diidentifikasi sebagai positif. True Positive (TP), di sisi lain, mengacu pada data positif yang terdeteksi secara akurat. Data False Negative (FN) adalah data positif yang secara keliru diidentifikasi sebagai negatif (kebalikan dari data True Positive) (Solichin, 2019). Keterangan dan penjelasan tentang confusion matrix pada tabel dua akan ditulis dibawah ini.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

$$Presisi = \frac{TP}{FP+TP} * 100\%$$

$$Recall = \frac{TP}{FN+TP} * 100\%$$

Gambar 2. Rumus

a. **True Positive** (TP) Yang berarti predikti tersebut positif dan benar.
b. **True Negative** (TN) Yang berarti prediksi tersebut negatif dan benar.
c. **False Positive** (FP) Yang berarti prediksi tersebut positif dan hal tersebut salah.
d. **False Negative** (FN) Yang berarti prediksi tersebut Negatif dan hal tersebut salah.

Kemudian dibawah merupakan gambar sebuah rumus untuk menghitung nilai Accuracy, Precision, dan Recall. Kemudian membagi dataset menjadi dua bagian yaitu data latih dan juga data uji menggunakan metode cross Validation X Fold data latih digunakan untuk model, dan data uji digunakan untuk mendapatkan hasil akurasi. Setelah memisahkan data, langkah selanjutnya yaitu menerapkan dataset kedalam beberapa algoritma untuk membandingkan hasil

akurasi dari tiap algoritma yang dipakai. Kemudian jika sudah dilakukan komparasi algoritma, akan menghasilkan model terbaik dari beberapa algoritma yang telah dicoba sebelumnya. Nilai akurasi dari model yang terbaik bisa mendapatkan hasil yang lebih maksimal dengan menggunakan algoritma optimasi. Dalam penelitian ini menggunakan fitur optimasi Algoritma Genetika atau evolutionary yang diklasifikasi menggunakan Random Forest.

9. Metode Yang Diusulkan

Metode yang diusulkan untuk penerapan data mining dalam analisis prediksi kanker paru menggunakan algoritma Random Forest

a. Pengumpulan Data

Data yang digunakan dalam analisis prediksi kanker paru harus terdiri dari informasi yang relevan seperti gejala, hasil tes, faktor risiko, dan informasi medis lainnya. Data ini dapat diperoleh dari sumber seperti pusat medis, lembaga penelitian, atau basis data terbuka seperti UCI Machine Learning Repository.

b. Persiapan Data

Setelah data terkumpul, langkah selanjutnya adalah mempersiapkan data untuk analisis. Proses ini mencakup menghilangkan data yang tidak relevan, mengisi nilai yang hilang, melakukan normalisasi data, dan mengkonversi data menjadi format yang dapat diproses oleh algoritma Random Forest.

c. Pembuatan Model

Setelah data dipersiapkan, model dapat dibangun menggunakan algoritma Random Forest. Algoritma ini menghasilkan banyak pohon keputusan yang masing-masing memberikan prediksi. Prediksi dari setiap pohon digabungkan untuk menghasilkan prediksi akhir.

d. ***Pelatihan Model***

Setelah model dibuat, langkah selanjutnya adalah melatih model menggunakan data yang telah dipersiapkan. Data dibagi menjadi dua set yaitu set pelatihan dan set validasi. Model dilatih pada set pelatihan dan dievaluasi pada set validasi untuk menentukan performanya.

e. Evaluasi Model

Setelah model dilatih, performanya dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Model yang memiliki performa terbaik dipilih untuk pengujian lebih lanjut.

f. Pengujian Model

Setelah model terpilih, model dapat diuji menggunakan data yang belum pernah dilihat sebelumnya. Pengujian ini akan memberikan gambaran tentang seberapa baik model dapat digunakan untuk memprediksi kanker paru.

g. Evaluasi Hasil

Setelah pengujian model selesai, hasil dapat dievaluasi dan dianalisis. Ini dapat dilakukan dengan memeriksa kinerja model dan menganalisis fitur-fitur yang paling berkontribusi dalam membuat prediksi.

Dalam keseluruhan, metode ini dapat digunakan untuk mengembangkan model prediksi kanker paru yang akurat dengan menggunakan algoritma Random Forest dan analisis data mining. Dengan menggunakan teknik ini, dokter dan peneliti dapat meningkatkan pemahaman mereka tentang kanker paru dan mengembangkan strategi untuk mencegah, mengidentifikasi, dan mengobati penyakit ini.

HASIL DAN PEMBAHASAN

1. Pengujian Sistem dan Pembahasan

Pada hasil 5 algoritma yang dibandingkan algoritma random forest yang menjadi hasil terbaik diantara algoritma lainnya. Pada tabel dibawah hasil sebelum menggunakan optimasi.

Tabel 2. Perbadingan 5 algoritma

Algoritma	**Validasi**	**Akurasi**	**AUC**
Decision Tree	Cross Validation	89.66%	0.808
KNN	Cross Validation	88.35%	0.828
Naïve Bayes	Cross Validation	89.31%	0.917
Random Forest	Cross Validation	90.61%	0.941
SVM	Cross Validation	90.26%	0.939

Setelah melakukan survey hasil akurasi dan nilai AUC dengan menggunakan algoritma random forest dengan split validation dari 0.5 hingga 0.9 terlihat pada tabel 4 berikut.

Tabel 3. Dataset lung cancer sebelum optimasi

Algoritma	Random Forest	
Validation	Akurasi	AUC
Split 0,5	91.56%	0.926
Split 0,6	90.32%	0.829

Split 0,7	91.40%	0.905
Split 0,8	91.94%	0.963
Split 0,9	93.55%	0.991
Cross Validation	90.61%	0.941

Peneliti menggunakan optimize feature evolutionary dan optimize weight evolutionary sehingga mendapatkan akurasi dan nilai yang lebih besar seperti pada tabel 5 dan tabel 6 berikut.

Tabel 4. Optimize Feature Evolutionary

Algoritma	Random Forest	
Validation	Akurasi	AUC
Split 0,5	94.16%	0.952
Split 0,6	93.55%	0.956
Split 0,7	97.85%	0.970
Split 0,8	98.39%	0.993
Split 0,9	100.00%	1.000
Cross Validation	94.17%	0.942

Terlihat pada tabel 3 bahwa nilai akurasi dan AUC mengalami peningkatan sebelum menggunakan optimize feature evolutionary dan mengalami peningkatan kembali Ketika memakai optimize weighting evolutionary seperti terlihat pada tabel 5.

Tabel 5. Optimize weight evolutionary

Algoritma	Random Forest	
Validation	Akurasi	AUC
Split 0,5	94.81%	0.934
Split 0,6	95.16%	0.886
Split 0,7	95.70%	0.965
Split 0,8	96.77%	0.993

Split 0,9	100.00%	1.000
Cross Validation	93.20%	0.946

Berikut merupakan hasil dari confusion matrix yang dapat dilihat dari tabel 7 dibawah ini.

Tabel 6. Confusion Matrix

	true YES	true NO	class precision
pred. yes	262	13	95.27%
pred.no	8	26	76.47%
class recall	97.04%	66.67%	

1. Accuracy

Disini peneliti mendapatkan hasil akurasi senilai 93.20%, jika dihitung menggunakan rumus maka hasil serupa didapat

$$Accuracy = \frac{262 + 26}{262 + 13 + 18 + 26} 100\% = \frac{288}{309} 100\% = 93.20\%$$

2. Precision

Class precision yang didapat senilai 95.27% dan 76.47% jika dihitung menggunakan rumus manual maka didapat hasil yang sama sebagai berikut.

$$Precision\ class\ 1\ = \frac{262}{13 + 262} 100\% = 0.952727\ x\ 100 = 95\%$$

$$Precision\ class\ 2 = \frac{26}{8 + 26} 100\% = 0.7647x100 = 76\%$$

3. Recall

Class recall yang didapat senilai 97.04% dan 66.67% dan jika dihitung menggunakan rumus manual maka didapat hasil yang sama sebagai berikut.

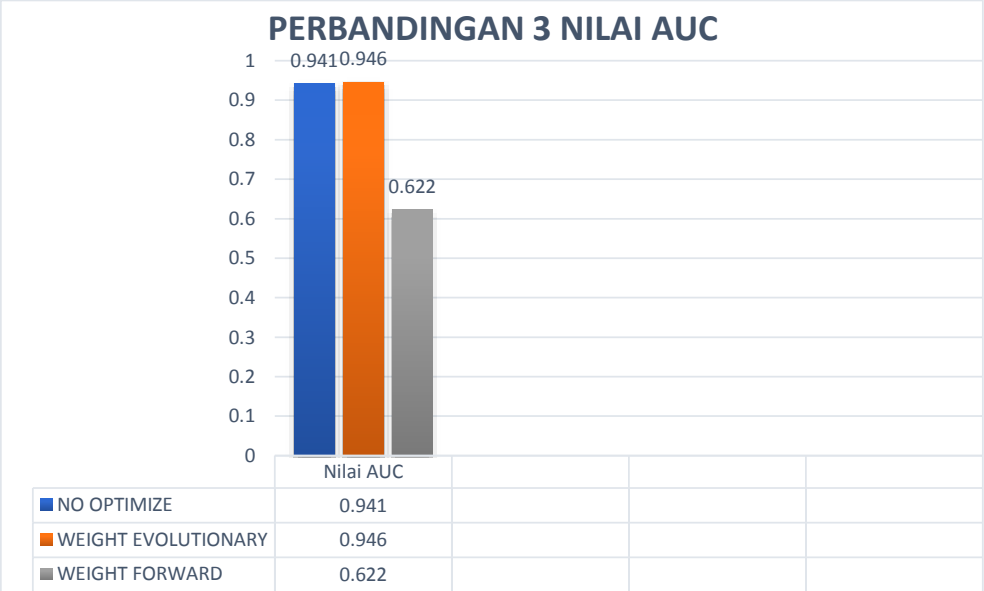
$$Recall\ 1 = \frac{262}{262 + 8} 100 = 0.9704X100 = 97\%$$

$$Recall\ 2 = \frac{26}{13 + 26} 100\% = 0.66667\ x\ 100 = 67\%$$

4. Area Under Curve (AUC)

Statistik yang umum digunakan untuk menilai model pengklasifikasi biner dalam penggalian data adalah area di bawah kurva (AUC). Statistik ini menunjukkan kemungkinan bahwa kejadian positif yang dipilih secara acak akan lebih mungkin menjadi positif daripada kejadian negatif yang dipilih secara acak. AUC merupakan metrik yang disukai karena bersifat agnostik terhadap distribusi kelas dan pengaturan ambang batas, serta memiliki rentang 0,5 (prediksi acak) hingga 1 (prediksi sempurna).

Terlihat pada grafik dibawah bahwa nilai AUC terbaik didapat adalah 0.946 dengan bantuan optimize weighting evolutionary disini peneliti akan membandingkan dengan nilai AUC saat tidak memakai optimize weighting evolutionary dan saat memakai optimize weighting forward.



Grafik 1. Perbandingan nilai AUC

Langkah selanjutnya peneliti akan melakukan uji coba dengan membandingkan dataset kanker paru ini menggunakan algoritma yang sama berupa random forest serta dibandingkan antara optimize weighting evolutionary dan optimize weighting forward berikut hasil yang terlihat pada table dibawah.

Tabel 8. Perbandingan nilai AUC

Validasi	Nilai AUC	Keterangan
Cross Validation	0.941	Tanpa Optimasi
Cross Validation	0.946	Weight Evolutionary
Cross Validation	0.622	Weight Forward

Pada tahap ini terdapat beberapa fitur yang terseleksi dari dataset kanker paru diantaranya:

Feature Weighting + Evolutionary: Age, Anxiety, Allergy, Alcohol Consumming, Chest Pain, Chronic Disease, Coughing, Fatigue, Peer Preasure, Smoking, Shortness Of Breath, Swallowing Difficulty, Wheezing.

Feature Weighting + Forward: Age, Anxiety, Allergy, Alcohol Consumming, Chest Pain, Chronic Disease, Coughing, Fatigue, Peer Preasure, Smoking, Shortness Of Breath, Swallowing Difficulty, Wheezing.

Dengan menerapkan Teknik data mining dan telah diuji menggunakan dataset survey kanker paru serta memakai perangkat lunak yang bernama Rapidminer untuk menganalisis dan memprediksi penderita positif terjangkit kanker paru dan negatif tidak terjangkit kanker paru berhasil, yang ditampilkan dalam tabel berikut.

Tabel 9. Hasil Survey

KANKER PARU	JUMLAH
Positif (YES)	270
Negatif (NO)	39

KESIMPULAN DAN SARAN

Hasil penelitian dengan menerapkan proses data mining dengan menggunakan perangkat lunak Rapidminer dan memakai algoritma beberapa seperti Decision Tree, K-Nearest Neighbor (KNN), Naïve Bayes, Random Forest, Support Vector Machine (SVM), disimpulkan bahwa algoritma Random Forest yang telah mendapatkan akurasi terbesar serta digunakan dan menggambarkan bahwa dataset lung cancer diperoleh hasil akurasi senilai 90.61% dengan nilai AUC 0.941 dan setelah memakai optimasi fitur evolutionary dengan algoritma genetika atau evolutionary hasil akurasi meningkat menjadi 94.17% dengan AUC senilai 0.942 dan kembali meningkat setelah memakai optimasi bobot evolutionary menjadi 93.20% dan nilai AUC sebesar 0.946. dengan demikian algoritma genetika atau evolutionary sangat bagus untuk meningkatkan nilai akurasi dan AUC.

Peneliti pun mencoba dengan memakai optimasi bobot forward pada dataset lung cancer tetapi mendapatkan hasil akurasi dan nilai AUC yang menurun menjadi disini terbukti bahwa optimasi bobot evolutionary terbukti memiliki performa baik dibandingkan fitur seleksi lainnya Dan tujuan penelitian ini yaitu meningkatkan performa dari klasifikasi penyakit kanker paru-paru dan memprediksi jumlah orang yang positif terjangkit dan orang yang tidak terjangkit kanker paru (negative), dengan data yang diperoleh dari situs online di internet.

UCAPAN TERIMA KASIH

Mengucapkan terima kasih yang sebesar-besarnya atas bimbingan dan dukungannya dalam proses penyusunan jurnal ini. Kami juga ingin mengucapkan terima kasih kepada semua pihak yang tidak dapat disebutkan satu per satu yang telah memberikan dukungan dan dorongan dalam proses pembuatan jurnal ini.

DAFTAR PUSTAKA

Halim, V. G., Darwis, Y., Rahmiati, R., Limantara, S., & Isa, M. (2020). Gambaran Tingkat Depresi pada Pasien Kanker Paru di RSUD Ulin Banjarmasin. *Homeostasis*, *3*(2), 309–318.

Haristu, R. A., & Rosa, P. H. P. (2019). Penerapan Metode Random Forest Untuk Prediksi Win Ratio Pemain Player Unknown Battleground. *Media Informasi Analisa Dan Sistem*, *2*, 120–128.

Idris, M. (2019). Implementasi Data Mining Dengan Algoritma Naïve Bayes Untuk Memprediksi Angka Kelahiran. *Pelita Informatika: Informasi Dan Informatika*, *7*(3), 421–428.

Jonathan, C. N. (2019). Implementasi Metode Algoritma Genetika Pada Penentuan Menu Makanan Untuk Membentuk Berat Badan Ideal. *Jurnal Teknologi Informasi Dan Terapan*, *6*(1), 35–40.

Khasanah, N. A., Oktaviyanti, I. K., & Yuliana, I. (2019). Hubungan riwayat merokok dan tempat tinggal dengan gambaran sitopatologi kanker paru. *Homeostasis*, *2*(1), 93–98.

Kusmardika, D. A. (2020). Potensi aktivitas antioksidan daun kelor (Moringa oleifera) dalam mencegahan kanker. *Journal of Health Science and Physiotherapy*, *2*(1), 46–50.

Mulyahati, I. L. (2020). *Implementasi Machine Learning Prediksi Harga Sewa Apartemen Menggunakan Algoritma Random Forest Melalui Framework Website Flask Python (Studi Kasus: Apartemen di DKI Jakarta Pada Website mamikos. com)*.

Prakoso, B. S., Rosiyadi, D., Aridarma, D., Utama, H. S., Fauzi, F., & Qhomar, M. A. N. (2019). Optimalisasi Klasifikasi Berita Menggunakan Feature Information Gain Untuk Algoritma Naive Bayes Terhubung Random Forest. *Jurnal PILAR Nusa Mandiri*, *15*(2), 211–218.

Ramadhaniaha, F., & Syarifb, S. (2020). Studi Tinjauan Pustaka: Risiko Kejadian Kanker Paru pada Penderita Tuberkulosis Paru. *Jurnal Epidemiologi Kesehatan Indonesia Vol*, *4*(1).

Saputra, E. W. (2019). Optimasi fungsi keanggotaan fuzzy Mamdani menggunakan algoritma genetika untuk penentuan penerima beasiswa. *Jurnal SIMADA (Sistem Informasi Dan Manajemen Basis Data)*, *2*(2), 160–175.