Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 3 Iss. 1 April 2023, pp: 15-19
ISSN(P) 2797-2313 | ISSN(E): 2775-8575

Implementation of Decision Tree Algorithm and Support Vector Machine for Lung Cancer Classification

Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru

Dhini Septhya[1], **Kharisma Rahayu**[2], **Salsabila Rabbani**[3], **Vindi Fitria**[4], **Rahmaddeni**[5*], **Yuda Irawan**[6], **Regiolina Hayami**[7]

[1,2,3,4,5]Program Studi Teknik Informatika, STMIK Amik Riau, Pekanbaru
[6]Program Studi Teknik Informatika, Universitas Hang Tuah, Pekanbaru
[7]Program Studi Teknik Informatika, Universitas Muhammadiyah Riau, Pekanbaru

E-Mail: [1]2010031802093@sar.ac.id, [2]2010031802129@sar.ac.id, [3]2010031802104@sar.ac.id, [4]2010031802085@sar.ac.id, [5]rahmaddeni@sar.ac.id, [6]yudairawan89@gmail.com, [7]regiolinahayami@umri.ac.id

Received Jan 19th 2023; Revised Feb 27th 2023; Accepted March 20th 2023
Corresponding Author: Rahmaddeni

Abstract

The lung cancer is one of the many causes of death in the world with a percentage of 11.6%, with a mortality rate of up to 18.4%. Lung cancer is one of the deadliest diseases because it is difficult to detect this cancer before it turns into a serious disease and currently there is no effective screening method for early detection of lung cancer. In this study, a classification technique was carried out which is a method of grouping data that has the same characteristics into several groups. The classification technique studied compares 2 algorithms, Decision Tree algorithm and the Support Vector Machine (SVM) algorithm to find out which algorithm gives the best results. In this study, feature selection will be carried out using forward selection which aims to increase the accuracy value. Based on the research that has been done, the results of the SVM algorithm using feature selection have a superior accuracy value of 62.3% using an 80:20 data splitting.

Keyword: Cancer, Classification, Decision Tree, Lung, SVM

Abstrak

Kanker paru merupakan satu dari banyaknya penyebab kematian di dunia dengan persentase 11.6%, dengan tingkat kematian hingga 18,4%. Kanker paru merupakan salah satu penyakit yang mematikan karena kanker ini sulit dideteksi sebelum berubah menjadi penyakit yang serius dan saat ini belum ada metode skrining yang efektif untuk deteksi dini kanker paru. Pada penelitian ini dilakukan teknik klasifikasi yang merupakan suatu metode pengelompokkan data yang memiliki karakter yang sama ke dalam beberapa kelompok. Teknik klasifikasi yang diteliti membandingkan 2 algoritma yaitu, algoritma Decision Tree dan Support Vector Machine (SVM) untuk mengetahui algoritma yang memberikan hasil terbaik. Dalam penelitian ini akan dilakukan seleksi fitur menggunakan forward selection yang bertujuan untuk menaikkan nilai akurasi. Berdasarkan penelitian yang telah dilakukan dapatkan hasil dari algoritma SVM menggunakan feature selection mempunyai nilai akurasi yang lebih unggul yaitu 62,3% menggunakan *splitting data* 80:20.

Kata Kunci: Decision Tree, Kanker, Klasifikasi, Paru, SVM

1. PENDAHULUAN

Kanker merupakan penyakit yang bisa membunuh manusia di berbagai negara. Menurut Organisasi Kesehatan Dunia, kanker menyebabkan kematian dari 9,6 juta manusia di seluruh dunia [1]. Kanker paru adalah penyebab utama kematian akibat keganasannya di seluruh dunia, terhitung mencapai 13% dari semua diagnosis kanker[2]. Dikutip dari Organisasi Kesehatan Dunia (WHO), kanker paru menjadi salah satu penyebab banyaknya kematian di antara kematian yang diakibatkan kanker lainnya, baik pada laki-laki maupun perempuan dari segala usia. Menurut data Globocan (2018), persentase tingkat kasus kanker paru yang terjadi di seluruh dunia sebesar 11,6% dan kematian menyentuh angka 18,4%, sedangkan di Indonesia angka kejadian kanker paru menyentuh angka 8,6% atau 30.023 kasus dengan tingkat kematian mencapai angka 12,6% atau

26.095 kasus kematian akibat dari kanker paru [3]. Berdasarkan penelitian sebelumnya, ditemukan bahwa kanker paru banyak ditemukan terjadi pada pria, sedangkan kanker paru termasuk kasus keempat pada wanita, kanker paru terjadi ketika sel-sel paru tumbuh dengan cepat dan tidak terkendali. Ketika sel-sel dari paru-paru berkembang di luar kendali menjadi penyebab kanker paru terjadi. Awal mula kanker paru dari tumor ganas yang terdapat di bagian epitel bronkus (*bronchogenic carcinoma*). Hingga kini metode skrining yang efektif untuk deteksi dini kanker paru belum ditemukan. Faktor risiko lain untuk kanker paru-paru termasuk terpapar radiasi, paparan dari bahan kimia yang dapat menyebabkan kanker, dan pasien yang memiliki riwayat mengidap kanker ataupun keluarga dari pasien. Sulitnya mendeteksi kanker paru menjadikan kanker paru berubah menjadi penyakit yang serius dan menyebabkan kanker paru-paru menjadi penyakit yang sangat mematikan [4]. Biasanya kanker paru banyak ditemukan pada orang yang cenderung perokok dan mempunyai gaya hidup yang tidak sehat, menjadikan kanker paru menjadi jenis kanker dengan posisi tertinggi ketiga di Indonesia [5].

Klasifikasi yaitu proses mengelompokkan objek-objek dengan karakteristik yang mirip dalam beberapa kelas. Pada umumnya pengklasifikasian dokumen diwakili oleh kalimat-kalimat penting dengan menentukan ciri-ciri atau karakteristik [6]. Salah satu metode dalam klasifikasi yaitu Decision Tree dan Support Vector Machine (SVM).

Decision tree adalah algoritma populer dan sangat efektif dengan melakukan pengklasifikasian dan prediksi. Algoritma decision tree dapat merepresentasikan ketentuan dari banyaknya fakta ke dalam bentuk pohon keputusan. pohon keputusan adalah struktur yang membagi sejumlah besar data menjadi sejumlah kecil data. Atribut kelas berfungsi sebagai representasi untuk simpul daun pohon keputusan. Node yang tidak ada termasuk node internal yang dihasilkan oleh kondisi uji atribut pada beberapa record dengan berbagai karakteristik dan node akhir yang terdiri dari akar. [7]. Algoritma Support Vector Machine (SVM) berfungsi untuk memperoleh hasil prediksi pengujian, yang mana hasil dari prediksi untuk pengujian diperoleh dari kelompok yang berupa *feature vector* [8]. Support Vector Machine (SVM) adalah teknik seleksi yang menghasilkan hasil dengan tingkat akurasi klasifikasi tertinggi dengan membandingkan sekumpulan parameter standar dengan nilai diskrit yang disebut sebagai himpunan kandidat. [9]. Keunggulan dari Support Vector Machine (SVM) selain dari populer juga sangat cocok untuk klasifikasi dikarenakan tidak bergantung pada jumlah atribut dan dapat menyelesaikan masalah dari dimensi. Secara komputasi, Support Vector Machine (SVM) dapat melakukan pelatihan secara cepat dan juga teknik learning-nya dapat menghadapi kesulitan dalam keragu-raguan [10].

Dalam penelitian sebelumnya, menggunakan algoritma Support Vector Machine (SVM) sebagai diagnosis kanker memiliki tingkat akurasi sebesar 56,69% [19]. Selanjutnya [21] yang memprediksi kanker payudara menggunakan algoritma decision tree dan Support Vector Machine (SVM) yang memiliki tingkat akurasi masing-masing algoritma yaitu sebesar 87,12% dan 91,92%. Penelitian [20] yang dilakukan menggunakan algoritma Decision Tree dalam memprediksi kanker pankreas didapatkan akurasi 72%.

Penelitian ini diharapkan dapat digunakan untuk melihat hasil perbandingan dalam klasifikasi kanker paru-paru menggunakan algoritma Decision Tree dan Support Vector Machine (SVM) dengan menambahkan Wrapper methods (*forward selection*) dari metode feature selection dan diharapkan dapat memberikan informasi tentang performa yang paling baik sehingga menghasilkan klasifikasi yang lebih akurat dalam pendiagnosaan kanker paru-paru.

2. METODE PENELITIAN

Dalam penelitian sudah melewati beberapa proses. Distribusi data latih dan data uji dibagi menjadi 3 perbandingan yaitu, 80:20, 70:30 dan, 60:40. Selanjutnya tahapan pembuatan model algoritma decision tree dan Support Vector Machine (SVM), setelah itu menjalankan model dengan data latih. Kemudian data uji memprediksi hasil, dari hasil data uji sebenarnya akan dibandingkan dengan hasil prediksi sehingga mendapatkan tingkat akurasi dari hasil prediksi. Atribut yang digunakan pada penelitian ini terdiri dari No, gender, age, smoking, yellow_fingers, anxiety, peer_pressure, chronic disease, fatigue, allergy, wheezing, alcohol consuming, coughing, shortness of breath, swallowing difficulty, chest pain dan lung_cancer. Flowchat metodologi dapat dilihat pada gambar 1.

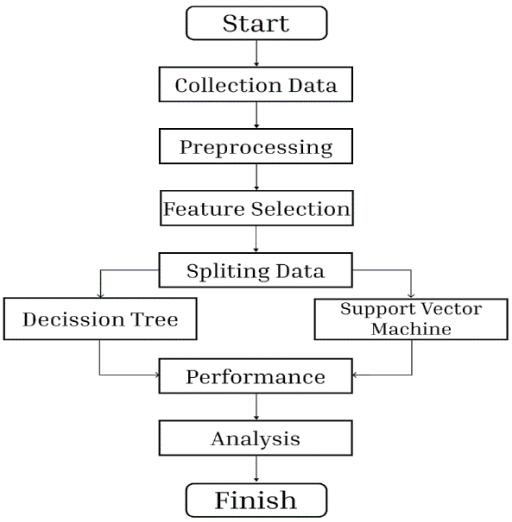
2.1 Collection Data (Pengumpulan Data)

Pada penelitian ini, data utama diperoleh dengan meninjau data kanker paru pada laman Kaggle.com. Data diproses agar mendapatkan daftar orang yang berpotensi mengidap penyakit kanker paru. Data bisa dilihat pada tabel 1.

Tabel 1. Dataset Penyakit Kanker Paru

No	gender	age	smoking	yellow_fingers	anxiety	peer_pressure	chronic disease	…	lung_cancer
1	0	69	1	2	2	1	1	…	1
2	0	74	2	1	1	1	2	…	1
3	1	59	1	1	1	2	1	…	2

No	gender	age	smoking	yellow_fingers	anxiety	peer_pressure	chronic disease	…	lung_cancer
4	0	63	2	2	2	1	1	…	2
…	…	…	…	…	…	…	…	…	…
1417	0	63	2	2	2	2	2	…	2
1418	1	67	1	2	2	1	1	…	1
1419	0	44	2	2	1	2	2	…	2



Gambar 1. Alur Penelitian

2.2 Pre-processing Data

Preprocessing data adalah salah satu proses data mentah untuk operasi pada pemrosesan lainnya [17]. Tahapan preprocessing data menghapus data yang bernilai null atau kosong dan merubah data menjadi lebih terstruktur dengan menggunakan implementasi dari cleaning data dan transformasi data. Langkah ini berfungsi agar data yang akan diproses menjadi lebih terstruktur dan memperlancar jalannya pemodelan.

2.3 Feature Selection

Feature selection (seleksi fitur) bertujuan untuk memilih feature yang berpengaruh dan mengesampingkan feature yang tidak berpengaruh dalam suatu kegiatan pemodelan atau penganalisaan data. Menggunakan teknik seleksi fitur mengurangi jumlah fitur digunakan untuk belajar dan memilih fitur diskriminasi tinggi dalam proses seleksi fitur. Selain itu, seleksi fitur membantu meningkatkan akurasi dengan memilih fitur yang optimal [22].

2.4 Splitting Data

Setelah tahap preprocessing, langkah selanjutnya adalah membagi data menjadi data latih dan data uji. Dengan pemisahan data 60:40, 70:30, 80:20, Data yang belum pernah digunakan dalam suatu penelitian, tetapi juga berguna untuk mengevaluasi keberhasilan atau kegagalan suatu penelitian, disebut data pengujian, sedangkan data pelatihan adalah data yang digunakan untuk melakukan penelitian [12]

2.5 Decision Tree

Decision tree adalah algoritma dari klasifikasi yang terfavorit karena bagi orang gampang untuk menginterpretasikannya [13]. Algoritma ini merupakan algoritma yang terdapat dalam metode klasifikasi dari data mining. Decision tree adalah konsep flowchart dengan struktur pohon (tree) dimana setiap node (node internal) mewakili atribut dan cabang menggambarkan hasil pengujian atau nilai input atribut, sedangkan daun mewakili kelas atau distribusi kelas [14]. ID3, CART, dan C4.5 merupakan algoritma yang dapat digunakan dalam penyusunan decision tree. Untuk memilih atribut sebagai akar, pada nilai *gain* tertinggi dari aribut-atribut yang ada. Untuk menghitung gain digunakan rumus seperti tertera dalam persamaan 1 berikut.

$$Gain \quad (S, A) = Entropy\ (s) - \sum_{i=1}^{n} \frac{|si|}{|s|} * Enteropy \tag{1}$$

Keterangan:

S = Himpunan Kasus
A = Atribut

n = Jumlah partisi atribut A
|Si| = Jumlah kasus pada partisi ke-i
|S| = Jumlah kasus dalam S

Sementara itu, perhitungan nilai entropi dapat dilihat pada persamaan 2.

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

Keterangan:
S = Himpunan kasus
A = Fitur
n = Jumlah partisi S
Pi = Proporsi dari Si terhadap S

2.6 Support Vector Machine

Support Vector Machine (SVM) adalah metode dari supervised learning untuk menganalisis data dan mengidentifikasi pola pada pengelompokan data[11]. Support Vector Machine (SVM) mengkonversi sebuah teks menjadi vektor sebelum digunakan pada klasifikasi. Ide dasar dari support vector machine ini yaitu mendapatkan area keputusan optimal (Hyperlane) untuk setiap titik data [15]. Proses klasifikasi mempunyai dua langkah pengerjaan, yaitu proses pengujian dan proses pelatihan. Proses pelatihan digunakan sebagai membuat model untuk suatu pengujian set [16].

2.7 Analisis

Pada tahap ini, metode dievaluasi dengan mengukur kinerja dari Support Vector Machine (SVM) dan algoritma Decision Tree. Metode tersebut dievaluasi dengan melakukan perbandingan dari tingkat akurasi algoritma memakai splitting data yang sama, yaitu 60:40, 70:30 dan 80:20.

3. HASIL DAN PEMBAHASAN

Pengujian yang telah dilakukan menggunakan algoritma pohon keputusan dan algoritma Support Vector Machine (SVM) yang diuji menghasilkan prediksi yang berbeda pada kasus kanker paru. Dalam pengujian ini menggunakan 1419 dataset dengan 11 feature sebelum dilakukannya preprocessing data. Tahapan preprocessing, dilakukan transformasi data agar menjadi bilangan binary.

Setelah dilakukannya preprocessing data. Tahap selanjutnya yaitu memodelkan algoritma SVM, SVM + *Forward Selection*, Decision Tree dan, Decision Tree + *Forward Selection* menggunakan rasio 60:40, 70:30 dan, 80:20. Berikut hasil implementasi model algoritma yang telah dibangun pada tabel 2.

Tabel 2. Perbandingan Tingkat Akurasi Algoritma

Splitting Data	*Decision Tree*	*Decision Tree + Forward Selection*	*SVM*	*SVM + Forward Selection*
60:40:00	53,3%	52,3%	54,6%	58,5%
70:30:00	55,4%	50,9%	56,1%	60,1%
80:20:00	56,7%	53,2%	54,2%	62,3%

Tabel 2 menunjukkan hasil perbandingan tingkat akurasi algoritma Decision Tree dan Support Vector Machine awal dan ditambah penggunaan forward selection. Diperoleh pengujian algoritma paling tinggi pada Decision Tree awal dengan rasio 80:20 bernilai 56.7, ditambah Forward Selection 53.2 dan Algoritma Support Vector Machine awal dengan rasio 80:20 bernilai 54.2, ditambah Forward Selection 63.2.

4. KESIMPULAN

Berdasarkan hasil dari pengolahan serta analisis dengan membandingkan dua algoritma, yaitu Decision Tree dengan Support Vector Machine didapatkan kesimpulan bahwa tingkat akurasi terbaik terdapat pada algoritma Support Vector Machine (SVM) menggunakan Forward Selection *splitting data* 80:20 dengan tingkat akurasi sebesar 62,3%. Penggunaan Forward Selection berpengaruh untuk menaikkan tingkat akurasi pada algoritma Support Vector Machine (SVM).

REFERENSI

[1] Naufal, S. A., Adiwijaya, A., & Astuti, W. (2020). Analisis Perbandingan Klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) untuk Deteksi Kanker dengan Data Microarray. *JURIKOM (Jurnal Riset Komputer)*, *7*(1), 162-168.

[2] Reynaldi, A., & Adiningsih, D. (2020). Gambaran Kualitas Hidup Pasien Kanker Paru Stadium Lanjut di RS Paru Dr. HA ROTINSULU Bandung. *Journal of Nursing Care*, *3*(2).

[3] Yunianto, M., Soeparmi, S., Cari, C., Anwar, F., Septianingsih, D. N., Ardyanto, T. D., & Pradana, R. F. Klasifikasi Kanker Paru Paru menggunakan Naïve Bayes dengan Variasi Filter dan Ekstraksi Ciri GLCM. *INDONESIAN JOURNAL OF APPLIED PHYSICS*, *11*(2), 256-268.

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

[5] Priyanti, E. (2021). Penerapan Algoritma Neural Network untuk Klasifikasi Kanker Paru. *Bianglala Informatika*, *9*(1), 56-60.

[6] Widiastuti, N. I., Rainarli, E., & Dewi, K. E. (2017). Peringkasan dan Support Vector Machine pada Klasifikasi Dokumen. *Jurnal Infotel*, *9*(4), 416-421.

[7] Muzakir, A., & Wulandari, R. A. (2016). Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree. *Scientific Journal of Informatics*, *3*(1), 19-26.

[8] Prasetyo, T. M., Amrullah, A., Syahrir, S., & Sari, B. N. (2022). Implementasi Algoritma SVM (Support Vector Machine) Dalam Klasifikasi Penyakit Paru-Paru Berdasarkan Fitur Pola Bentuk. (JurTI) Jurnal Teknologi Informasi, 6(1), 1-6

[9] Suhardjono, S., Wijaya, G., & Hamid, A. (2019). Prediksi Waktu Kelulusan Mahasiswa Menggunakan Svm Berbasis Pso. *Bianglala Informatika*, *7*(2), 97-101.

[10] Purwaningsih, E. (2016). Seleksi Mobil Berdasarkan Fitur dengan Komparasi Metode Klasifikasi Neural Network, Support Vector Machine, dan Algoritma C4. 5. *Jurnal Pilar Nusa Mandiri*, *12*(2), 153-160

[11] Muslehatin, W., Ibnu, M., & Mustakim, M. (2017). Penerapan Naïve Bayes Classification untuk Klasifikasi Tingkat Kemungkinan Obesitas Mahasiswa Sistem Informasi UIN Suska Riau. In *Seminar Nasional Teknologi Informasi Komunikasi dan Industri* (pp. 250-256).

[12] Nurkholifah, M., & Umar, Y. (2023). ANALISA PERFORMA ALGORITMA MACHINE LEARNING DALAM PREDIKSI PENYAKIT LIVER. *Jurnal Indonesia: Manajemen Informatika dan Komunikasi*, *4*(1), 164-172.

[13] Pramudiono, I. (2014). Pengantar Data Mining: Menambang Permata Pengetahuan di Gunung Data, 2003. *Ilmukomputer. com diunduh tanggal*, *13*.

[14] Ditendra, E., Suryani, S., Romelah, S., Tanjung, M. H. A., & Sarah, M. (2022). Perbandingan Algoritma Klasifikasi untuk Analisis Sentimen Islam Nusantara di Indonesia: Comparison of Classification Algorithms for Sentiment Analysis of Islam Nusantara in Indonesia. *Malcom: Indonesian Journal of Machine Learning and Computer Science*, *2*(1), 71-77.

[15] Supriyatna, A., & Mustika, W. P. (2018). Komparasi Algoritma Naive bayes dan SVM Untuk Memprediksi Keberhasilan Imunoterapi Pada Penyakit Kutil. *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, *2*(2), 152-161.

[16] Chazar, C., & Erawan, B. (2020). Machine Learning Diagnosis Kanker Payudara Menggunakan Algoritma Support Vector Machine. *INFORMASI (Jurnal Informatika Dan Sistem Informasi)*, *12*(1), 67-80.

[17] Akbar, F., Saputra, H. W., Maulaya, A. K., Hidayat, M. F., & Rahmaddeni, R. (2022). Implementasi Algoritma Decision Tree C4. 5 dan Support Vector Regression untuk Prediksi Penyakit Stroke: Implementation of Decision Tree Algorithm C4. 5 and Support Vector Regression for Stroke Disease Prediction. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *2*(2), 61-67.

[18] Ridwan, M., Suyono, H., & Sarosa, M. (2013). Penerapan Data Mining Untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes Classifier. *Jurnal EECCIS (Electrics, Electronics, Communications, Controls, Informatics, Systems)*, *7*(1), 59-64.

[19] Ab Hamid, T. M. T., Sallehuddin, R., Yunos, Z. M., & Ali, A. (2021). Ensemble based filter feature selection with harmonize particle swarm optimization and support vector machine for optimal cancer classification. *Machine Learning with Applications*, *5*, 100054.

[20] Vigia, E., Ramalhete, L., Chumbinho, B., Custódio, P., Macedo, M., Aguiar, C., ... & Marques, H. P. (2022). Machine Learning Decision Tree Help to Avoid Early Recurrence in Resectable Pancreatic Cancer. *HPB*, *24*, S322-S323.

[21] Tsehay Admassu Assegie, S. S. (2020). A Support Vector Machine and Decision Tree Based Breast Cancer Prediction. *International Journal of Engineering and Advanced Technology (IJEAT), ISSN*, 2249-8958.

[22] Lee, J., Park, D., & Lee, C. (2017). Feature selection algorithm for intrusions detection system using sequential forward search and random forest classifier. *KSII Transactions on Internet and Information Systems (TIIS)*, *11*(10), 5132-5148.