Jurnal Komputer dan Informatika
Vol 18 No 1, April 2023: hlm 69 - 79

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

PENERAPAN METODE *MACHINE LEARNING* DALAM KLASIFIKASI DATA TEKS PENDERITA KANKER

Ardiansyah Jaya Winata
Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara
Jln. Letjen S. Parman No. 1, Jakarta, 11440, Indonesia
E-mail: *ardiansyah.535210014@stu.untar.ac.id*

ABSTRAK

Penyakit kanker merupakan kondisi yang ditandai oleh pertumbuhan sel-sel abnormal yang tak terkendali dan mampu menyerang berbagai organ tubuh. Penelitian ini bertujuan untuk membandingkan model klasifikasi penderita kanker dengan memanfaatkan metode *Random Forest*, XGBoost, dan *Support Vector Machine* (SVM), dengan hasil akurasi dilaporkan untuk setiap algoritma. Dataset yang digunakan terdiri dari 7569 baris dengan dua kolom, di mana salah satunya menjadi target klasifikasi, yakni "Target". Sebelum digunakan dalam penelitian, dilakukan proses preprocessing Data untuk meningkatkan kualitas data. Hasil penelitian mengungkap bahwa *Random Forest* dan XGBoost mencapai akurasi tertinggi, yaitu sebesar 100%, dan SVM masing-masing mencapai akurasi 92%. Penelitian ini tidak hanya membandingkan akurasi algoritma, tetapi juga mengeksplorasi jenis kanker yang paling umum terjadi dari ketiga jenis kanker ini. Melalui penelitian ini, diharapkan dapat memberikan wawasan lebih lanjut tentang efektivitas algoritma klasifikasi dalam identifikasi penderita kanker serta menyoroti jenis kanker dengan jumlah kasus tertinggi.

Kata kunci: Kanker, *Preprocessing, Random Forest,* XGBoost, SVM

ABSTRACT

Cancer is a condition characterized by the uncontrolled growth of abnormal cells that can invade various organs of the body. This research aims to compare cancer classification models utilizing Random Forest, XGBoost, and Support Vector Machine (SVM) methods, with accuracy results reported for each algorithm. The dataset used consists of 7569 rows with two columns, one of which is the classification target, "Target". Before being used in the study, data preprocessing was conducted to improve data quality. The results revealed that Random Forest and XGBoost achieved the highest accuracy of 100%, while SVM achieved 92% accuracy respectively. This research not only compares the accuracy of the algorithms, but also explores the most common cancer types of these three cancers. Through this research, it is hoped to provide further insight into the effectiveness of classification algorithms in cancer patient identification as well as highlighting the types of cancer with the highest number of cases.

Keywords: *Cancer, Preprocessing, Random Forest, XGBoost, SVM.*

1 PENDAHULUAN

Kanker adalah sejumlah besar penyakit yang memungkinkan menyerang tubuh pada bagian mana saja. Istilah lain yang sering didengar adalah tumor ganas ataupun neoplasma. Salah satu ciri khas kanker adalah penciptaan cepat sel-sel abnormal yang tumbuh melampaui batas biasanya, dan yang kemudian dapat menyerang bagian tubuh yang berdekatan dan menyebar ke organ lain, proses terakhir disebut sebagai metastasis. Metastasis adalah penyebab utama kematian akibat kanker [1]. Angka penderita kanker selalu meningkat setiap tahun [2]. dalam penelitian ini ada 3 jenis kanker, *Thyroid Cancer, Lung Cancer* ,dan *Colon Cancer*.

Thyroid Cancer atau Kanker tiroid adalah tumor ganas yang terjadi pada sel parenkim tiroid. Keganasan ini mengenai dua jenis sel utama pada parenkim tiroid, yaitu sel folikel tiroid yang dapat berkembang menjadi kanker tiroid berdiferensiasi atau *Differentiated Thyroid Cancer* (DTC) dan sel parafolikular tiroid atau sel C yang dapat berkembang menjadi karsinoma tiroid meduler atau *Medullary Thyroid Carcinoma* (MTC) [3]. Salah satu gangguan yang cukup sering ditemukan pada kelenjar tiroid adalah munculnya nodul pada kelenjar tiroid [4]. Dalam penyembuhan kanker tiroid

ada sebuah metode yang Bernama Tiroidektom, Tiroidektomi adalah operasi pengangkatan kelenjar tiroid merupakan operasi yang bersih dan tergolong operasi besar. Seberapa luas kelenjar yang akan diambil tergantung keadaan klinis dan penggolongan risiko dari kanker tiroid serta perluasan tumor [5]. Selain kanker tiroid, ada juga *Lung Cancer* atau kanker paru.

Kanker paru merupakan suatu keganasan pada paru yang disebabkan oleh perubahan genetika pada sel epitel saluran nafas, sehingga terjadi proliferasi sel yang tidak terkendali. Keganasan ini dapat berasal dari organ paru itu sendiri (primer) maupun yang berasal dari luar paru (metastasis). Menurut data WHO menyebutkan bahwa Penyebab paling umum kematian akibat kanker pada tahun 2020 adalah kanker paru dengan jumlah sebanyak 1,80 juta kematian [6], kanker paru menjadi salah satu penyebab banyaknya kematian di antara kematian yang diakibatkan kanker lainnya, baik pada laki-laki maupun perempuan dari segala usia [7]. Selain kedua kanker yang sudah dijelaskan sebelumnya Adapun *Colon Cancer* atau kanker usus besar.

Kanker usus besar adalah salah satu organ pencernaan yang merupakan lanjutan dari usus halus. Usus besar sering juga disebut sebagai kolon. Fungsi utama dari usus besar adalah untuk melakukan penyerapan makanan yang tidak mampu diserap di usus besar, Juga Menyerap air dan garam sehingga dapat mengatur keseimbangan cairan dalam tubuh [8]. Sebagian besar kasus kanker usus besar diawali dengan pembentukan gumpalan-gumpalan sel berukuran kecil yang disebut polip adenoma [9], Pengobatan kanker dikenal dengan beberapa cara, salah satunya dengan kemoterapi, yaitu pengobatan dengan menggunakan obat-obatan yang dapat menghambat atau membunuh sel-sel kanker. Pengobatan dengan kemoterapi biasanya menggunakan satu atau lebih obat, sehingga dapat menimbulkan efek samping dan interaksi obat yang tidak diinginkan [10].

Dengan mengenal ketiga jenis kanker tersebut, penelitian ini bertujuan untuk menguji ketiga algoritma yaitu *Random Forest*, XGBoost, dan *Support Vector Machine* (SVM) pada sebuah Data Teks yang berisikan 2 kolom yaitu Target dan *Text*. Kolom *Text* berisikan Analisa gejala-gejala pasien yang mengalami ketiga jenis kanker tersebut. Hasil dari ketiga metode algoritma ini bisa menunjukan manakah metode yang memiliki kinerja terbaik dalam mengklasifikasi Data Teks ini.

2 TINJAUAN LITERATUR

Penelitian [11] mengenai perbandingan metode algoritma *Decision Tree* C4.5 dan Naïve Bayes untuk memprediksi Penyakit Tiroid dilakukan oleh beberapa peneliti Leli Safitri, Krista Cahayani Murtiwiyati, Siti Chodidjah, Deasy Indayanti yang di publikasikan November 2022. Penelitian ini berfokuskan pada perbandingan metode algoritma *Decision Tree* C4.5 dan Naïve Bayes untuk memprediksi penyakit tiroid. Melibatkan analisis dataset dari UCI *Machine Learning Repository*, hingga evaluasi menggunakan ten-folds cross-validation. Hasil analisis menunjukkan bahwa *Decision Tree* C4.5 memiliki tingkat akurasi yang lebih tinggi (97.12%) dibandingkan dengan Naïve Bayes (76.02%), didukung oleh nilai *Area Under Curve* (AUC) pada kurva *Receiver Operating Characteristic* (ROC) yang lebih tinggi.

Penelitian [12] yang dilakukan oleh Sri Widodo dari Fakultas Ilmu Kesehatan, Universitas Duta Bangsa Surakarta, fokus pada klasifikasi kanker paru dan arteri pada citra *Computed Tomography* (CT) menggunakan *Convolutional Neural Network* (CNN). Dalam upaya deteksi awal kanker paru, banyak penelitian telah dilakukan dengan sering dimulai dari preprosesing citra, segmentasi paru, dan langkah-langkah lainnya, yang memakan waktu cukup lama. Penelitian ini mengusulkan pendekatan baru dengan menggunakan CNN untuk mengklasifikasikan kanker paru dan arteri, dengan akurasi tertinggi mencapai 95%. Metode pengujian dilakukan dengan membandingkan hasil deteksi kanker menggunakan CNN dengan hasil deteksi oleh dokter, menggunakan matriks konfusi. Uji coba dilakukan dengan tiga dataset berbeda: citra *CT-Scan* asli, citra *CT-Scan* paru, dan citra yang sudah dilokalisasi.

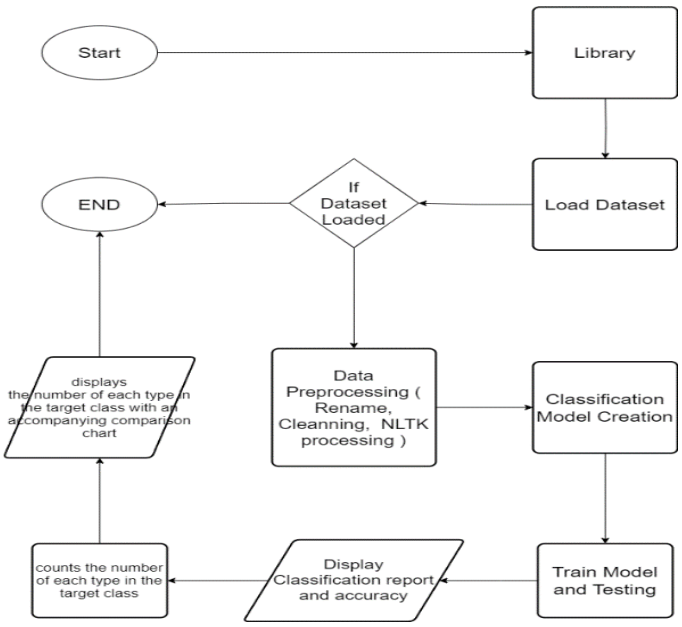
ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Penelitian [13] yang ditulis oleh Jaelani Muhammad Akbar, Muhammad Sabirin, Gibran Satya Nugraha, dan Noor Alamsyah. bertujuan untuk mengaplikasikan metode *Principal Component Analysis* (PCA) dan *K-Nearest Neighbors* (KNN) dalam klasifikasi data kanker paru-paru. Fokus penelitian ini adalah pada identifikasi awal penyakit pernapasan terkait paru-paru melalui analisis citra CT-Scan menggunakan pendekatan kecerdasan buatan. Proses penelitian ini melibatkan langkah-langkah seperti *pre-processing* citra, ekstraksi fitur menggunakan PCA, dan klasifikasi menggunakan metode KNN dengan variasi nilai K. Melalui penelitian ini, hasil akurasi tertinggi sebesar 98% diperoleh pada K = 9. Hasil ini mendukung kesimpulan bahwa metode PCA untuk ekstraksi fitur dan KNN untuk klasifikasi cocok digunakan dalam pengolahan dataset kanker paru-paru, memberikan kontribusi penting dalam diagnosis dini dan evaluasi nodul paru-paru.

Dalam [14] penelitian yang publikasikan dalam Jurnal Masyarakat Informatika (JMASIF) dan di tulis oleh Muhammad Sofi Yuniarto dan Eko Adi Sarwoko, penelitian ini memfokuskan pada implementasi metode *K-Nearest Neighbor* (KNN) untuk diagnosis kanker kolorektal menggunakan biomarker Micro-RNA. Kanker kolorektal, yang berasal dari jaringan usus besar, dapat dideteksi melalui metode skrining menggunakan micro-RNA, terutama miR-21, miR-31, miR-135b, miR-183, miR-222, miR-145, dan miR-195. Metode KNN digunakan untuk mengklasifikasikan data *micro-RNA*, dengan dataset berjumlah 600 data yang terbagi antara normal dan kanker kolorektal. Pembagian dataset menggunakan metode *K-Fold Cross Validation.* Hasil pengujian menunjukkan bahwa metode KNN memiliki performa terbaik pada K=3 dengan akurasi 94,17%, spesifisitas 94,43%, dan sensitivitas 94,41%. Penelitian ini menggambarkan bahwa perubahan nilai K pada KNN dapat mempengaruhi performa, dan melalui pengujian *K-Fold Cross Validation*, ditemukan bahwa K=3 memberikan model terbaik dengan pertimbangan nilai sensitivitas yang tinggi. Kurva ROC juga menunjukkan bahwa pada K=3, model memiliki performa terbaik pada fold 9 dengan akurasi 100%, spesifisitas 100%, dan sensitivitas 100%, sementara model dengan performa terendah terdapat pada *fold* 2.

3 METODE PENELITIAN

Dalam perencanaan melakukan klasifikasi Data Teks penderita kanker, ada beberapa hal yang harus dipersiapkan sebelum melakukan klasifikasi. Ada beberapa persiapan yang diperlukan dalam perencanaan klasifikasi ini Dataset, Pra Pemrosesan, metode evaluasi, Skema Eksperimen, serta penjelasan tentang algoritma yang akan digunakan.



Gambar 1. Skema Eksperimen dalam klasifikasi Data Teks Penderita Kanker

3.1 Data

Dataset yang digunakan untuk melakukan klasifikasi bersumber dari platform *Kaggle* . jumlah dari data ini adalah 7569 baris dengan 3 kolom nomor urut, Target, dan Teks. Target *Class* disini adalah pada kolom Target klasifikasi Dimana kolom Target ini memiliki tiga nilai, yaitu "*Colon_Cancer*", "*Lung_Cancer*",dan "*Thyroid_Cancer*". Pada kolom Teks, kolom ini berisikan Analisa gejala-gejala pasien yang mengalami ketiga jenis kanker tersebut.

3.2 Pra pemrosesan

Dalam Pra-Pemrosesan disini, Dataset yang awalnya adalah 3 kolom, di ubah menjadi 2 kolom hal itu dikarenakan kolom nomor unit tidak memiliki fungsi dan tekertaitan pada kolom lain. Kemudian setelah di hapus dilanjutkan dengan *Pre-Processing* data. Dataset yang digunakan adalah sebuah Data Teks maka dari itu *Pre-Processing* dilakukan untuk membersihkan Teks seperti Menghapus karakter khusus dan mengonversi teks menjadi huruf kecil, Menghilangkan kata-kata *stop* (kata umum yang tidak informatif), dan Menghapus kata-kata pendek (kurang dari 3 karakter). Dengan melakukan langkah-langkah ini dapat membantu menyederhanakan teks, menghilangkan informasi yang tidak relevan, dan mempersiapkan data untuk analisis lebih lanjut atau pelatihan model klasifikasi. Kemudian untuk pembagian data latih dan uji, disini dataset dibagi menjadi 80% data latih dan 20% data uji. Pembagian ini bertujuan untuk menguji performa setiap metode dalam mengklasifikasikan akurasi setiap metode algoritma pada Dataset

3.3 Random Forest

Random Forest merupakan salah satu algoritma machine learning untuk klasifikasi data dalamjumlah yang besar [15]. Metode ini merupakan sebuah *ensemble* (kumpulan) metode pembelajaran menggunakan pohon keputusan sebagai *base classifier* yang dibangun dan dikombinasikan *decision tree* atau pohon pengambil keputusan adalah sebuah diagram alir yang berbentuk seperti pohon yang memiliki sebuah *root node* yang digunakan untuk mengumpulkan data [16] . *Random Forest* biasanya digunakan untuk analisa klasifikasi dan regresi [17] .

3.3.1 Rumus Matematika pada metode *Random Forest*

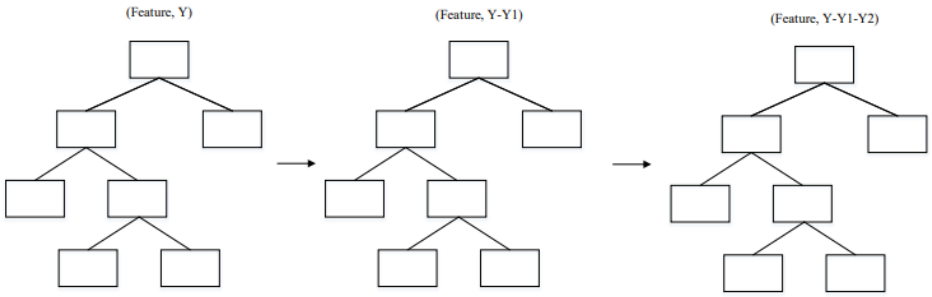
$$Gini = 1 - \sum_{i=1}^{c} (P_i)^2 \quad (1)$$

Saat melakukan *Random Forest* berdasarkan data klasifikasi, indeks Gini digunakan sebagai faktor penentu dalam membuat keputusan tentang percabangan node pada pohon keputusan. Indeks Gini memungkinkan perhitungan ketidakmurnian atau ketidakjelasan di setiap cabang node, dan ini membantu dalam pemilihan cabang yang memiliki kemungkinan lebih tinggi untuk menghasilkan prediksi yang akurat. Dalam rumus ini, "pi" mengacu pada frekuensi relatif kelas yang diamati dalam data, dan "c" adalah jumlah kelas yang ada dalam masalah klasifikasi.

3.4 XGBoost

Extreme Gradient Boosting (XGBoost) adalah metode *boosting* dengan cara menggabungkan kumpulan pohon keputusan yang akan digunakan untuk pembangunan pohon selanjutnya [18]. model tersebut merupakan algoritma ensemble tree yang terdiri dari beberapa pohon klasifikasi atau pohon regresi. Proses pembentukan pohon pertama diambil dari data latih (*feature,* Y) mendapatkan hasil estimasi pohon pertama (Y1). Selanjutnya pada pohon kedua dilakukan proses pembentukan pohon dari data latih (*feature,* |Y-Y1|), dimana nilai |Y-Y1| merupakan selisih dari label nyata dengan label prediksi tahap sebelumnya. Pohon ketiga melakukan proses pembentukan pohon dari data (*feature,* |Y-Y1-Y2|) dan memperoleh hasil estimasi Y3. Dari langkah tersebut, nilai *error* dapat direduksi dengan efektif. Berikut adalah stuktur dari XGBoost [19]

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)



Gambar 2. Struktur pohon keputusan pada XGBoost

3.5 *Support Vector Machine* (SVM)

SVM adalah model termasuk dalam *supervised learning* yang dapat menganalisis data dan mengenali pola untuk proses klasifikasi [20]. Prinsip dasar algoritma ini adalah Klasifikasi Linear, kemudian dikembangankan agar dapat berfungsi pada Klasifikasi Non-Linear. Prinsip dasar SVM adalah pengembangan yang mengklasifikasikan linear agar dapat diproses pada masalah non-linear. Prinsip dasar ini memakai metode *kernel trick* pada fitur berdimensi tinggi. Hasil akurasi data yang dihasilkan algoritma SVM ditentukan oleh parameter dan fungsi kernelyang digunakan [21].

$$K(x, y) = x, y \tag{1}$$

$$K(x, y) = (x, y + 1)p \tag{2}$$

$$K(x, y) = e^{-|x.y|2/2\sigma^2} \tag{3}$$

$$(x, y) = tanh(Kx. y - \delta \tag{4}$$

Rumus SVM linear ada pada (1) sedangkan rumus SVM nonlinear ada pada (2), (3), dan (4). Rumus polynomial ada pada (2), Rumus RBF ada pada (3), dan Rumus Sigmoid ada pada (4)

3.6 Metode Evaluasi

Metode evaluasi akan dilakukan dengan ketiga metode yang sudah ditentunkan, untuk pengujian metode *Random Forest* menggunakan estimator sebanyak 100 untuk melakukan pelatihan model, dan pada metode *Support Vector Machine* (SVM) menggunakan Rumus SVM Linear untuk melakukan pelatihan model. Setelah model ketiga metode tersebut sudah terbuat maka akan dilakukan report classification untuk melihat nilai akurasi yang di dapat dan melihat perbandingan jumlah *Thyroid Cancer*, *Lung Cancer* ,dan *Colon Cancer* secara keseluruhan.

3.7 Skema Eksperimen

Skema eksperimen dalam penelitian yang akan dilakukan Digambarkan pada gambar 1. Eksperimen akan dimulai dengan membuat Library disusul dengan memanggil dataset dalam format csv. Yang setelah dipanggil maka dilakukan Pra-pemrosesan dari mengubahan nama kolom data menjadi Target dan *Text* setelah pergantian nama kolom maka akan dilnjutkan dengan *Pre-Processing* data Dimana tujuan ini dilakukan untuk membersihkan Teks seperti Menghapus karakter khusus dan mengonversi teks menjadi huruf kecil, Menghilangkan kata-kata *stop* (kata umum yang tidak informatif), dan Menghapus kata-kata pendek (kurang dari 3 karakter).setelah melakukan Pra-pemrosesan maka dilanjutkan untuk melakukan proses klasifikasi dengan ketiga metode tersebut, hasil yang di dapat akan di evaluasi lebih lanjut pada Hasil dan pembahasan

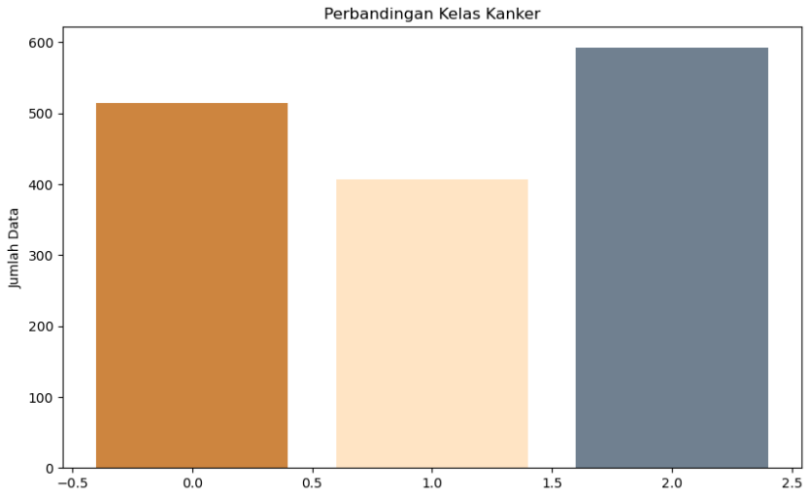
4 HASIL DAN PEMBAHASAN

Pada hasil klasifikasi dengan menggunakan ketiga metode ini, hasil akurasi tertinggi dimiliki oleh *Random Forest* dan XGBoost dengan akurasinya adalah 100% atau 0.9986789960369881% sedangkan pada *Support Vector Machine* (SVM) mendapatkan 92% atau 0.917437252311757%. Hasil ini menunjukan bahwa peforma dari *Random Forest* dan XGBoost dalam mengklasifikasi Data teks ini sangat bagus. Ketiga metode ini juga menghasilkan sebuah laporan klasifikasi yang dapat menjelaskan lebih detail tentang peforma model, berikut hasil laporan klasifikasi dari ketiga model.

Tabel 1. Hasil Laporan Klasifikasi ketiga metode

	Random Forest			XGBoost			SVM		
	0	1	2	0	1	2	0	1	2
Precision	1.00	1.00	1.00	1.00	1.00	1.00	0.86	1.00	0.91
Recall	1.00	1.00	1.00	1.00	1.00	1.00	0.90	1.00	0.87
F1-Score	1.00	1.00	1.00	1.00	1.00	1.00	0.88	1.00	0.89
Accuracy	1.00			1.00			0.92		

Hasil laporan klasifikasi menunjukan *Precision, Recall*, dan *F1-Score* pada ketiga metode ini menunjukan hasil sangat baik. *Precision* merupakan sebuah ukuran yang mengukur tingkat proporsi jumlah dokumen yang dapat ditemukan kembali oleh sebuah proses pencarian dan dianggap relevan untuk kebutuhan pencarian informasi atau rasio jumlah dokumen relevan yang ditemukan dengan total jumlah dokumen yang ditemukan, dalam hasil laporan klasifikasi tersebut metode *Random Forest* dan XGBoost memiliki hasil *Precision* yang sama pada ketigas *Class* dan pada *Support Vector Machine* (SVM) memiliki hasil yang berbeda pada ketiga *Class*, *Class* 0 mendapatkan 0.86, *Class* 1 mendapatkan 1.00 dan *Class* 2 mendapatkan 0.91 . *Recall* adalah proporsi jumlah data yang dapat ditemukan Kembali oleh sebuah proses pencarian informasi [22], pada metode *Random Forest* dan XGBoost menghasilkan yang sama yaitu 1.00 sedangkan pada *Support Vector Machine* (SVM) *Class* 0 mendapatkan 0.90, *Class* 1 mendapatkan 1.00 dan *Class* 2 mendapatkan 0.87. *F1-score* adalah perbandingan presisi dan perolehan rata-rata tertimbang [23]. dalam hasil F1-Score pada metode *Random Forest* dan XGBoost tetap menghasilkan nilai yang sama sedangkan pada *Support Vector Machine* (SVM) memiliki hasil yang berbeda. Pada *Class* 0 mendapatkan 0.88, *Class* 1 mendapatkan 1.00 dan *Class* 2 mendapatkan 0.89. dalam klasifikasinya ketiga metode ini memiliki diagram batang untuk melihat nilai detail dari setiap *Class* yang menjadi target dalam klasifikasi ini.



Gambar 3. Perbandingan *Class* pada metode Random Forest

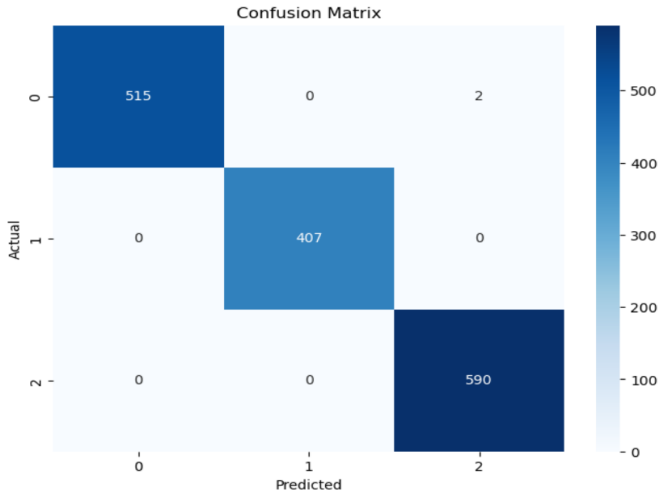
Dalam perbandingan setiap *Class* pada metode *Random Forest* diatas, menunjukan bahwa *Class* 2 memiliki nilai tertinggi dari *Class* 0 dan 1. Adapun detail nilai dari setiap *Class* berdasarkan perbandingan pada gambar [nomor gambar diatas].

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Tabel 2. Detail nilai perbandingan setiap *Class* pada metode *Random Forest*

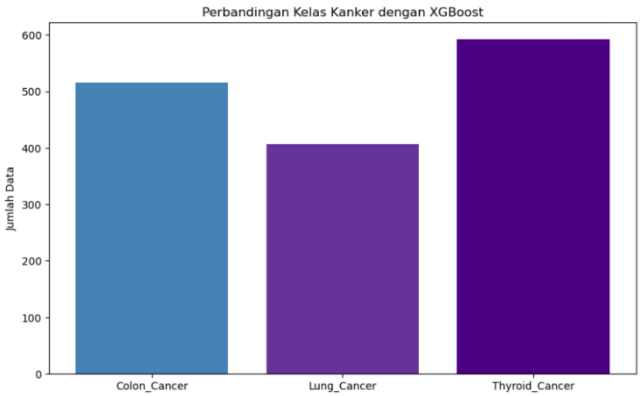
Random Forest		
0	*Colon Cancer*	515
1	*Lung Cancer*	407
2	*Thyroid Cancer*	592

Pada tabel [tabel *random forest*], menunjukan bahwa *Class* 2 memiliki hasil prediksi sebesar 592, pada *Class* 0 memiliki hasil prediksi sebesar 515 dan pada *Class* 1 memiliki hasil prediksi sebesar 407. Hasil prediksi ini juga disertakan *Confussion Matrix* dari metode *Random Forest*.



Gambar 4. *Confusion Matrix* dari metode *Random Forest*

Sama halnya pada metode *Random Forest*, XGBoost menghasilkan diagram dan *Confussion Matrix* yang sama dengan *Random Forest.*



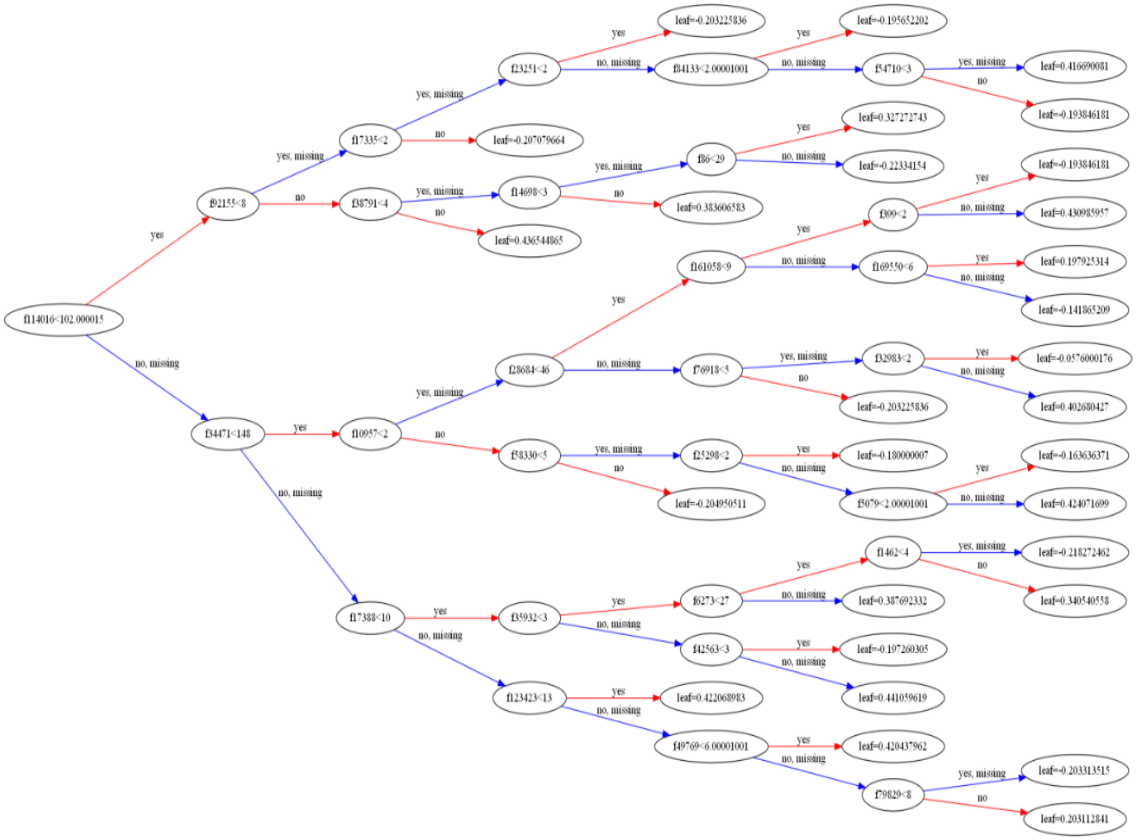
Gambar 5. Perbandingan setiap *Class* pada metode XGBoost

Detail nilai yang dihasilkan dari perbandingan setiap *Class* pada metode XGBoost juga sama dengan metode Random Forest.

Tabel 2. Detail nilai perbandingan setiap *Class* pada metode XGBoost

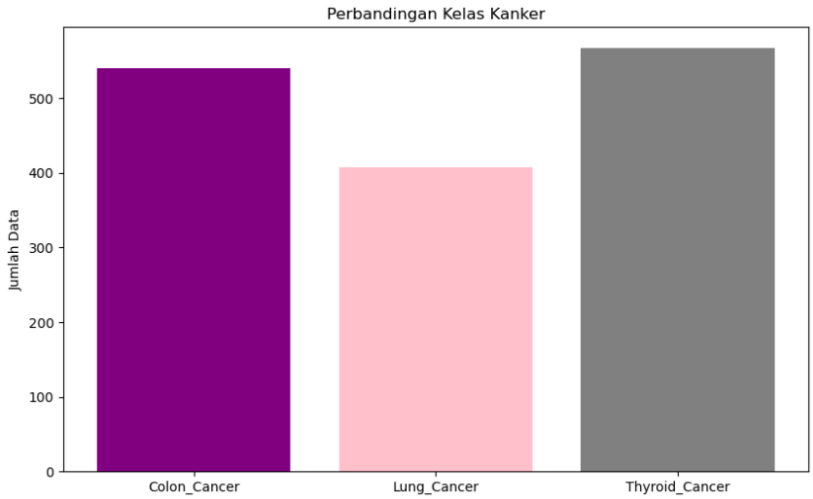
XGBoost		
0	*Colon Cancer*	515
1	*Lung Cancer*	407
2	*Thyroid Cancer*	592

Selain diagram batang, XGBoost juga memiliki pohon Keputusan. Dalam pohon Keputusan ini terdapat perumusan klasifikasi yang di lakukan pada metode XGBoost.



Gambar 6. Pohon Keputusan XGBoost

Pada perbandingan setiap *Class* dengan metode *Support Vector Machine* (SVM) hanya *Class* 1 yang memiliki kesamaan dengan kedua metode sebelumnya, hal ini dikarenakan ketiga metode ini memiliki nilai yang sama pada *Precision, Recall, F1-Score* yaitu 1.00.



Gambar 7. Perbandingan setiap *Class* pada metode *Support Vector Machine* (SVM)

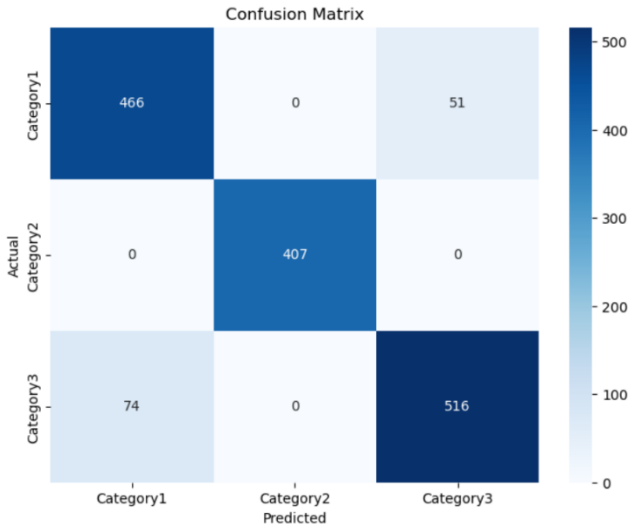
Dalam perbandingan setiap *Class* dengan metode SVM ini. Detail nilai yang di dapat memiliki perbedaan namun jika dilihat dari diagram batangnya, *Class* 0 dan *Class* 2 memiliki perbandingan hampir serupa.

ISSN 1410-7228 (Versi Cetak)
ISSN 2716-4233 (Versi Elektronik)

Tabel 3. Detail nilai setiap *Class* pada metode *Support Vector Machine* (SVM)

Support Vector Machine (SVM)		
0	*Colon Cancer*	540
1	*Lung Cancer*	407
2	*Thyroid Cancer*	567

Selain diagram batang, pada metode SVM ini juga memiliki *Confussion Matrix*\



Gambar 8. *Confussion Matrix* pada metode SVM

5 KESIMPULAN

Berdasarkan penelitian ini, penggunaan metode klasifikasi seperti *Random Forest*, XGBoost, dan *Support Vector Machine* (SVM) telah diimplementasikan pada dataset analisis gejala pasien kanker dengan hasil yang memuaskan. Setelah melewati tahap pra-pemrosesan data, pembentukan model, dan evaluasi, *Random Forest* dan XGBoost mencapai akurasi 100%, sementara SVM mencapai 92%. Hasil laporan klasifikasi menunjukkan bahwa semua metode memberikan Precision, *Recall*, dan *F1-Score* yang baik untuk setiap kelas kanker. Visualisasi seperti *Confusion Matrix* dan diagram batang memberikan gambaran yang jelas tentang performa metode *Random Forest* dan XGBoost. Keseluruhan, penelitian ini menyimpulkan bahwa *Random Forest* dan XGBoost efektif dalam mengklasifikasikan data teks terkait kanker, sedangkan SVM juga memberikan hasil yang baik.

DAFTAR PUSAKA

[1] R. Arania, R. Alfarisi, P. Rukmono and M. F. Mudtaghfirin, "Karakteristik Pasien Kanker Anak Berdasarkan Usia, Jenis Kelamin, Dan Jenis-Jenis

[2] Kanker Di Rsud Dr. H. Abdul Moeloek Tahun 2021," *Jurnal Medika Malahayati*, vol. Vol. 7 No. 2, p. 352, 2022.

[3] L. Rahayuwatri, I. A. Rizal, T. Pahria, M. Lukman and N. Juniarti, "Pendidikan Kesehatan tentang Pencegahan Penyakit Kanker dan Menjaga Kualitas Kesehatan," *MEDIA KARYA KESEHATAN*, vol. 3 No 1, p. 60, 2020

[4] N. Shafira and A. Wahyuni, "Manajemen Anestesi Pada Pasien Kanker Tiroid: *Sebuah Laporan Kasus," Laporan Kasus*, vol. Vol. 13 No. 1, p. 19, 2022.

[5] . M. P. Cardia, E. D. Martadiani and F. P. Sitanggang, "Karakteristik Ultrasonografi Pada Kecurigaan Klinis Kanker Tiroid Di Rsup Sanglah Denpasar Periode Januari 2015-Desember 2015," *Jurnal Medika Udayana*, vol. Vol. 9 No. 9, p. 76, 2020.

[6] F. I. Fathoni and A. S. Siwi, "Studi Kasus Asuhan Keperawatan pada Pasien Post Operasi Tiroidektomi atas Indikasi Kanker Tiroid," *Indogenius*, vol. Vol. 01 No. 02, p. 88, 2022.

[7] I. Buana and D. A. Harahap, "Asbestos, Radon Dan Polusi Udara Sebagai Faktor Resiko Kanker Paru Pada Perempuan Bukan Perokok," Averrous: *Jurnal Kedokteran dan Kesehatan Malikussaleh*, vol. Vol. 8 No. 1, p. 2, 2022.

[8] D. Septhya, K. Rahayu, S. Rabbani, V. Fitria, R. Y. Irawan and R. Hayami, "Implementasi Algoritma Decision Tree dan Support Vector Machine untuk Klasifikasi Penyakit Kanker Paru," MALCOM: Indonesian *Journal of Machine Learning and Computer Science*, vol. Vol. 3 No. 1, p. 15, 2023.

[9] . E. Putri and M. A. Rahman, "Sistem Pakar Mendiagnosa Stadium Dari Kanker Usus Besar Dengan Metode Certainty Factor," SYNTAX : *Journal of Software Engineering, Computer Science and Information Technology*, vol. Vol. 1 No. 2, pp. 63-64, 2020.

[10] R. D. Rambe, "Sistem Pakar Mendiagnosa Penyakit Kanker Usus Besar Pada Manusia Dengan Menerapkan Metode Hybrid Case Based," *Jurnal Riset Komputer (JURIKOM*), vol. vol. 6 No. 6, p. 606, 2019.

[11] N. K. Rusdi, E. N. Sari and N. Wulandari, "Ketepatan Obat, Dosis, dan Potensi Interaksi Obat pada Pasien Kanker Paru di Rumah Sakit X Jawa Barat Periode 2019-2021," *Jurnal Sains dan Kesehatan*, vol. Vol. 5 No. 3, p. 315, 2023.