Prediksi Kekambuhan Kanker Tiroid Menggunakan Algoritma *Random Forest*

Egi Safitri[1], Dani Rofianto[2*], Sri Karnila[3], Nurjoko[4], Hendra Kurniawan[5], Yuni Arkhiansyah[6], Ruki Rizal[7]

[1,3,4,5,6,7]Program Studi Sains Data, Institut Informatika dan Bisnis Darmajaya, Bandar Lampung
[2]Program Studi Teknologi Rekayasa perangkat Lunak, Politeknik Negeri Lampung, Bandar Lampung
*danirofianto@polinela.ac.id

Diterima : 25 April 2025
Disetujui : 26 Mei 2025

***Abstract*—Kekambuhan kanker tiroid pasca terapi *Radioactive Iodine* (RAI) merupakan tantangan penting dalam penatalaksanaan jangka panjang pasien. Penelitian ini bertujuan membangun model prediktif untuk mengidentifikasi potensi kekambuhan dengan memanfaatkan data klinis dan patologis menggunakan algoritma *Random Forest. Dataset* terdiri atas 383 data pasien dengan 13 atribut, termasuk usia, jenis kelamin, staging kanker, jenis patologi, klasifikasi risiko, dan respons terhadap terapi. Proses pra-pemrosesan meliputi penyandian data kategorik, eksplorasi fitur, dan pembagian data latih dan uji secara stratifikasi. Hasil evaluasi menunjukkan performa tinggi dari model, dengan akurasi 96,5%, presisi 96,7%, recall 90,6%, dan AUC 99%. Analisis fitur menggunakan SHAP mengungkap bahwa Stage, Response, dan Risk merupakan faktor paling berkontribusi terhadap prediksi kekambuhan. Penelitian ini menunjukkan bahwa model *Random Forest* tidak hanya efektif dalam klasifikasi biner, tetapi juga dapat diinterpretasikan secara klinis untuk mendukung pengambilan keputusan medis yang lebih personal dan preventif..**

Keywords — kanker tiroid, kekambuhan, Random Forest, prediksi medis, fitur klinis dan patologis

I. PENDAHULUAN

Kanker tiroid merupakan salah satu bentuk keganasan endokrin yang paling umum dan menunjukkan tren peningkatan insiden secara global dalam beberapa dekade terakhir [1], [2], [3]. Meskipun memiliki tingkat kelangsungan hidup yang tinggi, kekambuhan tetap menjadi tantangan klinis yang serius karena dapat menyebabkan peningkatan morbiditas dan intervensi medis lanjutan. Deteksi dini terhadap risiko kekambuhan memiliki peranan penting dalam menentukan strategi terapi lanjutan dan pemantauan pasien secara lebih intensif.

Dalam beberapa tahun terakhir, pendekatan berbasis pembelajaran mesin mulai digunakan secara luas untuk meningkatkan akurasi diagnosis dan prediksi penyakit, termasuk kanker tiroid [4], [5], [6]. Berbagai variabel klinis dan patologis—seperti usia, jenis kelamin, stadium TNM (Tumor-Node-Metastasis), jenis patologi, serta respons terapi—dapat dimanfaatkan untuk membangun model prediksi yang andal.

Berbagai studi terdahulu juga menyoroti pentingnya integrasi antara data klinis dan patologis. Penelitian oleh Mishra et al. [7] menggunakan *Random Forest* untuk menganalisis gangguan tiroid dan menunjukkan kinerja yang sangat baik pada dataset diagnosis tiroid umum.

Chaganti et al. [8] menyatakan bahwa teknik seleksi fitur berbasis *Random Forest* menghasilkan akurasi tinggi dalam mendeteksi berbagai jenis penyakit tiroid, dan teknik ini juga dapat digunakan untuk memaksimalkan efisiensi prediksi kekambuhan.

Sementara itu, Alshayeji [9] dalam studinya tentang prediksi awal risiko penyakit tiroid menunjukkan bahwa metode ensembel seperti *Random Forest* dan *boosting* mampu

menghasilkan akurasi di atas 99% dalam mendeteksi gangguan tiroid. Namun, model-model tersebut cenderung berfokus pada klasifikasi awal fungsi tiroid (hipo/hipertiroid) dan belum secara spesifik diterapkan untuk mengantisipasi kekambuhan kanker tiroid yang memiliki kompleksitas klinis tersendiri.

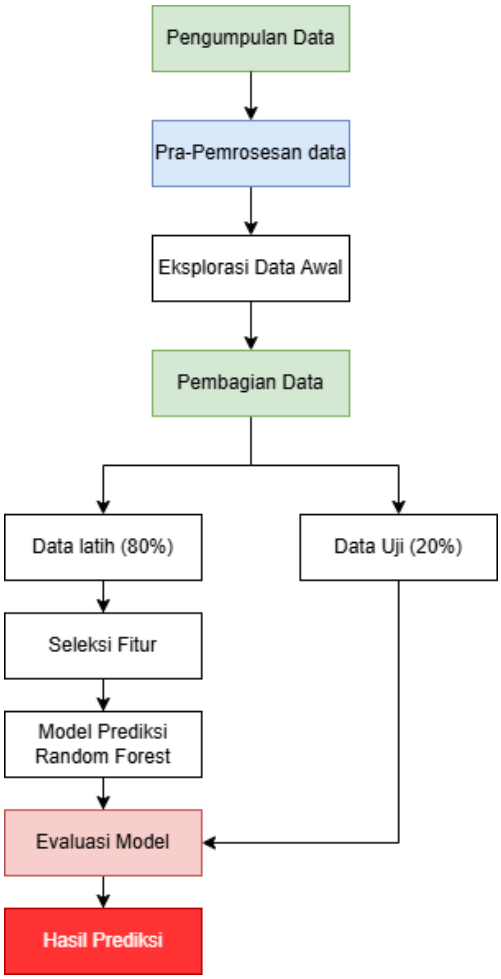
Dalam konteks kanker tiroid, kekambuhan dapat terjadi bahkan pada pasien dengan respons terapi yang awalnya baik [10]. Oleh karena itu, penelitian ini dilakukan untuk mengisi celah tersebut, dengan mengembangkan model prediksi kekambuhan kanker tiroid berbasis data klinis dan patologis yang kaya konteks lokal. Tidak seperti studi sebelumnya, penelitian ini menitikberatkan pada analisis pasca-terapi terhadap pasien kanker tiroid yang telah menjalani *Radioactive Iodine* (RAI), dengan memperhatikan fitur-fitur kunci seperti respons terapi dan klasifikasi risiko klinis yang berperan penting dalam kekambuhan. Dengan fokus ini, penelitian bertujuan memberikan kontribusi nyata dalam pengembangan sistem pendukung keputusan medis yang tidak hanya akurat, tetapi juga memiliki interpretabilitas melalui fitur penting hasil seleksi berbasis *Gini Index* dari algoritma *Random Forest*.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk membangun model prediksi kekambuhan kanker tiroid dengan memanfaatkan algoritma *Random Forest* berbasis fitur-fitur klinis dan patologis pasien. Melalui pendekatan ini, diharapkan penelitian dapat memberikan kontribusi pada pengembangan sistem pendukung keputusan medis yang lebih presisi dan berbasis data untuk penanganan kanker tiroid. Model yang akurat dan dapat dijelaskan ini akan membantu dokter dalam menetapkan strategi monitoring pascaoperasi, serta memberikan gambaran risiko kekambuhan secara individual kepada pasien.

II. DATA DAN MODE

A. *Desain Penelitian*

Deteksi kanker tiroid dimulai dengan pengumpulan data klinis dan patologis yang relevan. Data kemudian diproses melalui tahap pra-pemrosesan seperti pembersihan dan encoding. Setelah itu dilakukan eksplorasi data awal untuk memahami pola dasar. Dataset dibagi menjadi dua bagian, yaitu data latih (80%) dan data uji (20%). Model prediktif dibangun menggunakan algoritma Random Forest setelah melalui proses seleksi fitur. Evaluasi model dilakukan dengan metrik seperti akurasi, presisi, *recall*, dan *F1-score*. Gambar 1 menunjukkan desain penelitian yang lebih jelas.



Gambar 1. Desain Penelitian Prediksi Thyroid

B. *Dataset*

Dataset Penelitian ini menggunakan dataset yang berfokus pada kasus kekambuhan kanker tiroid setelah terapi *Radioactive Iodine* (RAI). Dataset terdiri dari 383 data pasien dengan 13 atribut utama yang merepresentasikan informasi klinis dan patologis penting. Atribut-atribut ini sangat berguna dalam mendukung proses prediksi kekambuhan, analisis faktor risiko, serta evaluasi efektivitas pengobatan.

Setiap baris dalam dataset merepresentasikan satu pasien, tanpa adanya nilai kosong atau hilang, sehingga proses pemrosesan data dapat dilakukan secara langsung tanpa imputasi. Atribut-atribut dalam dataset dapat dilihat pada Tabel 1.

Tabel 1. Fitur dan Deskripsi Data

Fitur	Deskripsi
Age	Usia pasien (dalam tahun)
Gender	Jenis kelamin pasien (Laki-laki atau Perempuan)
Hx Radiotherapy	Riwayat radioterapi sebelumnya (Ya atau Tidak)
Adenopathy	Keterlibatan kelenjar getah bening (Ya atau Tidak)
Pathology	Jenis kanker tiroid (misalnya, Micropapillary)
Focality	Fokalitas tumor (Uni-Focal atau Multi-Focal)
Risk	Klasifikasi risiko kanker (Rendah, Menengah, Tinggi)
T	Klasifikasi tumor primer (T1, T2, dst.)
N	Klasifikasi kelenjar getah bening (N0, N1, dst.)
M	Klasifikasi metastasis (M0, M1)
Stage	Stadium kanker (Stadium I, II, III, IV)
Response	Respons terhadap terapi (Excellent, Indeterminate, dsb.)
Recurred	Status kekambuhan kanker (Ya atau Tidak)

C. *Pra-pemrosesan Data*

Pada tahap pra-pemrosesan data dilakukan untuk memastikan bahwa data dalam kondisi yang siap digunakan dalam proses pelatihan model. Seluruh nilai pada dataset telah dinyatakan lengkap, tanpa adanya missing values. Beberapa langkah utama yang dilakukan meliputi:

1. Penyandian variabel kategorik yakni kolom-kolom kategorik seperti *Gender, Pathology*, *Risk*, *Response*, dan lainnya diubah menjadi format numerik menggunakan teknik *one-hot encoding* agar dapat diterima oleh model machine learning.
2. Normalisasi dan konsistensi format dilakukan untuk memastikan integritas data.
3. Dataset dibagi menjadi dua yakni: data latih (80%) untuk proses pelatihan model dan data uji (20%) untuk proses evaluasi. Pembagian dilakukan secara stratifikasi berdasarkan kolom target *Recurred* untuk menjaga proporsi kelas tetap seimbang.

D. *Algoritma*

Random Forest merupakan algoritma untuk memprediksi kekambuhan kanker tiroid. *Random Forest* adalah metode yang menggabungkan beberapa *decision tree*, di mana hasil akhir ditentukan melalui mekanisme *voting* [11], [12], [13]. Algoritma ini efektif dalam menangani data dengan fitur numerik maupun kategorik, serta tahan terhadap *overfitting* karena melibatkan pengacakan pada data dan fitur saat pelatihan. Selain itu, *Random Forest* juga memberikan informasi penting terkait kontribusi setiap fitur terhadap prediksi.

E. *Evaluasi*

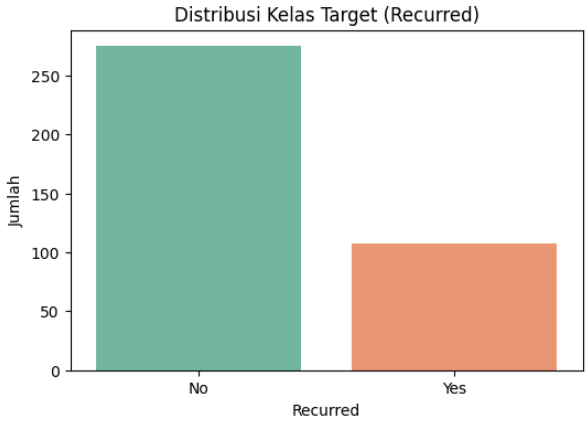
Evaluasi model dilakukan dengan mengukur sejauh mana model mampu mengklasifikasikan data secara benar. Metrik evaluasi yang digunakan dalam penelitian ini mencakup akurasi, presisi, recall, dan F1-score [14],[15].

III. HASIL DAN PEMBAHASAN

Pada tahap ini dibahas mengenai eksplorasi data awal, hasil analisis model dan interpretasi.

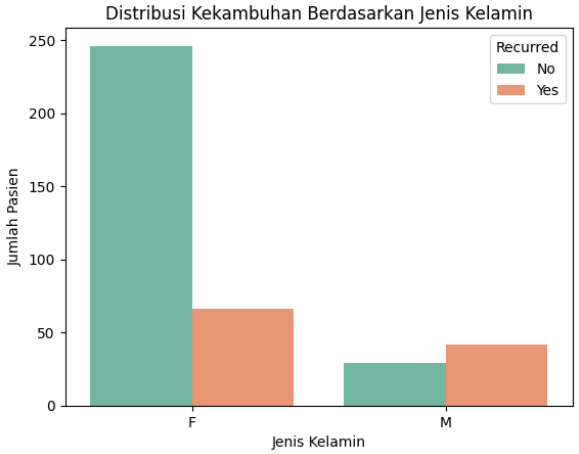
A. *Eksplorasi Data Awal*

Eksplorasi data awal dilakukan untuk memahami karakteristik umum dari dataset yang digunakan dalam penelitian ini. Dataset terdiri dari 383 data pasien kanker tiroid yang telah menjalani terapi *Radioactive Iodine* (RAI), dengan 13 atribut klinis dan patologis yang relevan terhadap risiko kekambuhan. Dari total 383 pasien, sebanyak 108 pasien (28,2%) tercatat mengalami kekambuhan kanker (label *Recurred = Yes*), sementara 275 pasien (71,8%) tidak mengalami kekambuhan. Distribusi yang tidak seimbang ini menjadi perhatian dalam proses pelatihan model, karena dapat mempengaruhi performa prediksi kelas minoritas. Distribusi target dapat dilihat pada Gambar 2.



Gambar 2. Distribusi Kelas Target

Sebagian besar pasien adalah perempuan, dengan persentase lebih dari 80%. Namun, jika dilihat dari proporsi kekambuhan, laki-laki cenderung memiliki risiko kekambuhan lebih tinggi dibandingkan perempuan, meskipun jumlahnya lebih sedikit secara keseluruhan. Distribusi kekambuhan berdasarkan jenis kelamin dapat dilihat pada Gambar 3.



Gambar 3. Distribusi Kekambuhan Berdasarkan Jenis Kelamin

B. Hasil Analisis Model

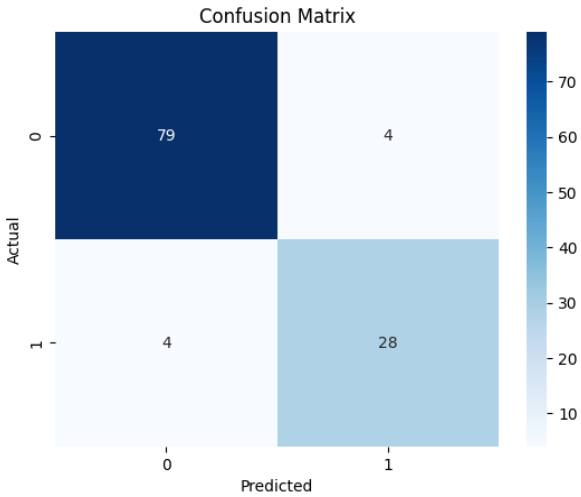
Setelah melalui proses pelatihan menggunakan algoritma Random Forest, model diuji menggunakan data uji yang telah dipisahkan sebelumnya (20% dari total data). Hasil evaluasi model terhadap prediksi kekambuhan kanker tiroid ditunjukkan melalui metrik pada Tabel 3.

Tabel 1. Hasil Evaluasi Model

Metrik Evaluasi	Nilai
Akurasi	96,5%
Presisi	96,7%
Recall	90,6%
F1-Score	93,5%
ROC AUC	99,0%

Berdasarkan Tabel 3, terlihat bahwa model memiliki performa yang sangat baik dalam mengklasifikasikan pasien yang mengalami kekambuhan maupun tidak. Nilai akurasi sebesar 96,5% menunjukkan tingkat ketepatan model yang tinggi dalam keseluruhan prediksi. Presisi yang tinggi menandakan bahwa sebagian besar pasien yang diprediksi mengalami kekambuhan benar-benar mengalami kekambuhan. Sementara itu, *recall* sebesar 90,6% menunjukkan bahwa model berhasil mendeteksi sebagian besar pasien yang benar-benar mengalami kekambuhan.

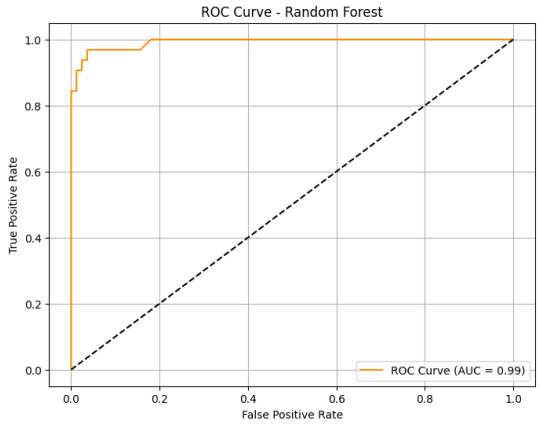
Visualisasi performa model ditampilkan pada Gambar 4 dan Gambar 5 berikut.



Gambar 4. Confusion Matrix Random Forest

Gambar 4 menunjukkan bahwa dari total data uji, sebanyak 79 pasien berhasil diklasifikasikan dengan benar sebagai tidak mengalami kekambuhan (*True Negative*), dan 28 pasien berhasil diklasifikasikan dengan benar sebagai mengalami kekambuhan (*True Positive*). Model hanya melakukan 4 kesalahan pada masing-masing kelas, yaitu 4 *False Positive* (pasien diprediksi kambuh padahal tidak) dan 4 *False Negative* (pasien diprediksi tidak kambuh padahal

kambuh). Hasil ini menunjukkan bahwa model memiliki tingkat kesalahan sangat rendah, serta mampu menyeimbangkan antara deteksi kasus positif dan negatif secara efektif.

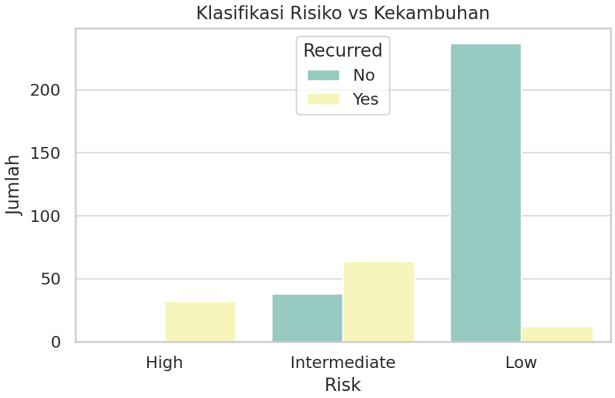


Gambar 5. ROC Curve Random Forest

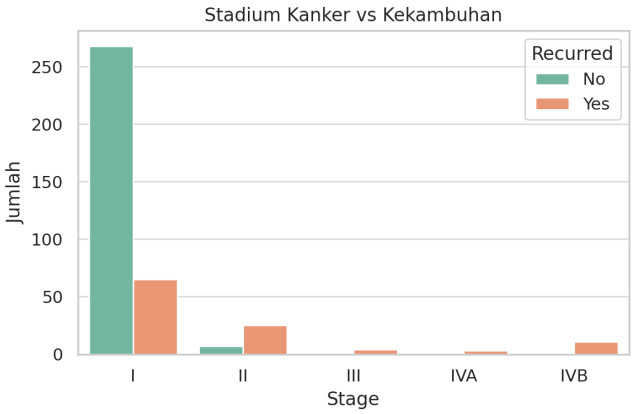
Gambar 5 menunjukkan kinerja model dalam membedakan antara pasien yang mengalami kekambuhan dan yang tidak. Kurva ROC mendekati sisi kiri atas grafik, yang merupakan indikasi kuat bahwa model memiliki tingkat sensitivitas dan spesifisitas yang tinggi. Nilai AUC sebesar 0,99 menandakan bahwa model *Random Forest* memiliki kemampuan diskriminatif yang hampir sempurna, sehingga sangat andal dalam memprediksi kekambuhan kanker tiroid secara akurat.

C. Diskusi

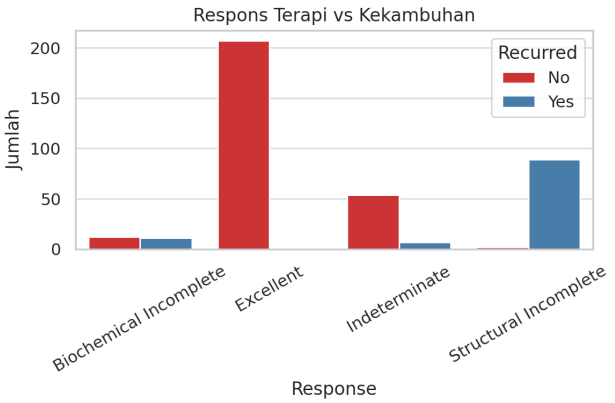
Hasil prediksi menggunakan model Random Forest menunjukkan performa yang sangat baik, dengan akurasi sebesar 96,5%, presisi 96,7%, dan recall 90,6%. Untuk memahami lebih lanjut kontribusi fitur terhadap prediksi kekambuhan kanker tiroid, dilakukan analisis visual terhadap tiga faktor utama: stadium kanker (*Stage*), respon terapi (*Response*), dan klasifikasi risiko kanker (*Risk*). Visualisasi kekambuhan dapat dilihat pada Gambar 6.



a). Klasifikasi Risiko vs Kekambuhan



b). Stadium Kanker vs Kekambuhan



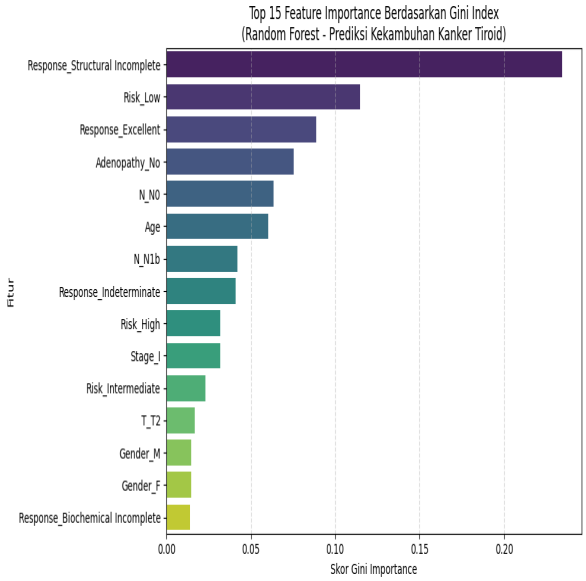
c). Respons Terapi vs Kekambuhan

Gambar 7. a), b), c) Kontribusi Fitur Terhadap Prediksi Kekambuhan Kanker Tiroid

Visualisasi a) menunjukkan pasien dengan klasifikasi risiko "*Intermediate*" dan "*High*" memiliki jumlah kekambuhan yang lebih tinggi dibandingkan pasien berisiko rendah. Ini konsisten dengan prediksi model, di mana fitur Risk juga memiliki kontribusi penting dalam menentukan kemungkinan kekambuhan.

Visualisasi b) menunjukkan bahwa kekambuhan lebih banyak terjadi pada pasien dengan stadium lanjut (Stage III dan IV). Jumlah kekambuhan pada stadium awal (Stage I–II) relatif kecil. Hal ini mendukung hasil prediksi model yang memberikan bobot penting pada variabel Stage.

Selanjutnya, visualisasi c) Pasien dengan respon terapi "*Indeterminate*" dan "*Biochemical Incomplete*" memiliki proporsi kekambuhan yang signifikan dibandingkan pasien dengan respons "*Excellent*". Model *Random Forest* berhasil mengenali pola ini dan menjadikannya sebagai salah satu indikator utama dalam klasifikasi. Hal ini tercermin dalam tingginya nilai *recall* model, yang menunjukkan kemampuan dalam mengidentifikasi pasien dengan risiko kekambuhan.



Gambar 8. Peringkat Pentingnya Fitur Berdasarkan Gini Index

Gambar 8 menunjukkan tingkat kepentingan setiap fitur dalam model Random Forest berdasarkan skor Gini Index. Fitur "*Risk*" (klasifikasi risiko klinis), "*Stage*" (stadium kanker), dan "*Therapy_Response*" (respons terhadap terapi) muncul sebagai tiga fitur paling signifikan dalam prediksi kekambuhan kanker tiroid. Hal ini sejalan dengan temuan klinis yang menyatakan bahwa pasien dengan risiko klinis menengah hingga tinggi, stadium lanjut, dan respons terapi yang tidak optimal memiliki kemungkinan kekambuhan yang lebih besar.

Model *Random Forest* menghitung pentingnya fitur dengan menjumlahkan kontribusi penurunan impuritas (dalam hal ini, *Gini impurity*) setiap kali fitur digunakan untuk membagi data di seluruh pohon dalam hutan. Oleh karena itu, fitur dengan skor tertinggi dianggap paling informatif dalam memisahkan kelas target (kambuh atau tidak kambuh).

Pengetahuan ini dapat dimanfaatkan oleh tenaga medis untuk memfokuskan perhatian pada fitur-fitur yang secara signifikan meningkatkan risiko kekambuhan. Dengan demikian, model tidak hanya berfungsi sebagai alat prediksi, tetapi juga mendukung interpretasi klinis yang lebih dalam.

IV. SIMPULAN

Penelitian ini menunjukkan bahwa model prediksi berbasis algoritma Random Forest mampu mengidentifikasi potensi kekambuhan kanker tiroid dengan akurasi yang sangat tinggi dan interpretabilitas yang kuat. Temuan ini menegaskan bahwa fitur-fitur klinis seperti stadium kanker, klasifikasi risiko, dan respons terhadap terapi memiliki peran penting dalam menentukan hasil jangka panjang pasien. Dengan demikian, pemanfaatan pembelajaran mesin tidak hanya berpotensi meningkatkan efisiensi prediksi klinis, tetapi juga mendukung pendekatan pengobatan yang lebih personal dan proaktif.

Untuk pengembangan ke depan, disarankan penelitian dilakukan dengan dataset yang lebih besar dan beragam secara geografis, serta mempertimbangkan integrasi data biomolekuler atau radiologis guna meningkatkan kedalaman prediksi. Selain itu, pengembangan sistem berbasis web atau aplikasi klinis berbasis model ini juga menjadi peluang besar dalam mendukung pengambilan keputusan dokter secara *real-time* di fasilitas layanan kesehatan.

DAFTAR PUSTAKA

[1] J. Ramírez-Moya *et al.*, "Identification of an interactome network between lncRNAs and miRNAs in thyroid cancer reveals SPTY2D1-AS1 as a new tumor suppressor," *Sci Rep*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-11725-4.

[2] J. Sharifi-Rad *et al.*, "Plant natural products with anti-thyroid cancer activity," 2020. doi: 10.1016/j.fitote.2020.104640.

[3] A. Popławska-Kita *et al.*, "Thyroid carcinoma with atypical metastasis to the pituitary gland and unexpected postmortal diagnosis," *Endocrinol Diabetes Metab Case Rep*, vol. 2020, no. 1, 2020, doi: 10.1530/EDM-19-0148.

[4] E. Safitri, R. R. Fikri, and R. Nurlistiani, "Application of Ensemble Machine Learning for Infectious Diseases with Vaccine Intervention: A Global COVID-19 Case Study," *JURNAL INFOTEL*, vol. 16, no. 4, Dec. 2024, doi: 10.20895/infotel.v16i4.1263.

[5] E. Safitri, R. Nurlistiani, and R. R. Fikri, "COVID-19 Case Prediction in Indonesia Using Ensemble Machine Learning Techniques with Vaccinations," in *2024 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, IEEE, Nov. 2024, pp. 1–7. doi: 10.1109/CENIM64038.2024.10882687.

[6] T. Suresh, T. A. Assegie, S. Rajkumar, and N. K. Kumar, "A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model," *International Journal of Electrical and Computer Engineering*, vol. 12, no. 2, 2022, doi: 10.11591/ijece.v12i2.pp1831-1838.

[7] S. Mishra, Y. Tadesse, A. Dash, L. Jena, and P. Ranjan, "Thyroid disorder analysis using random forest classifier," in *Smart Innovation, Systems and Technologies*, 2021. doi: 10.1007/978-981-15-6202-0_39.

[8] R. Chaganti, F. Rustam, I. De La Torre Díez, J. L. V. Mazón, C. L. Rodríguez, and I. Ashraf, "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques," *Cancers (Basel)*, vol. 14, no. 16, 2022, doi: 10.3390/cancers14163914.

[9] M. H. Alshayeji, "Early Thyroid Risk Prediction by Data Mining and Ensemble Classifiers," *Mach Learn Knowl Extr*, vol. 5, no. 3, 2023, doi: 10.3390/make5030061.

[10] C. S. Grant, "Recurrence of papillary thyroid cancer after optimized surgery.," *Gland Surg*, vol. 4, no. 1, 2015, doi: 10.3978/j.issn.2227-684X.2014.12.06.

[11] A. Sekulić, M. Kilibarda, G. B. M. Heuvelink, M. Nikolić, and B. Bajat, "Random forest spatial interpolation," *Remote Sens (Basel)*, vol. 12, no. 10, 2020, doi: 10.3390/rs12101687.

[12] W. A. W. A. Bakar, N. L. N. B. Josdi, M. B. Man, and Y. S. Triana, "An Evaluation of Artificial Neural Networks and Random Forests for Heart Disease Prediction," *Journal of Hunan University Natural Sciences*, vol. 49, no. 2, 2022, doi: 10.55463/issn.1674-2974.49.2.4.

[13] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, no. 1, 2020, doi: 10.1177/1536867X20909688.

[14] D. Rofianto, E. Safitri, K. Amaliah, J. Fitra, and A. Hijriani, "Cyber Threat Detection Using an Ensemble Model Approach for Phishing Website Identification ARTICLE INFORMATION ABSTRACT," 2024. [Online]. Available: http://innovatics.unsil.ac.id

[15] E. Safitri, D. Rofianto, N. Purwati, H. Kurniawan, and S. Karnila, "Prediksi Penyakit Diabetes Melitus Menggunakan Algoritma Machine Learning," vol. 12, no. 4, 2024, doi: 10.26418/justin.v12i4.84620.