JURNAL SISTEM INFORMASI DAN TEKNOLOGI
(S I N T E K)

Situs Jurnal
https://sintek.stmikku.ac.id/index.php/home

MODEL *MACHINE LEARNING* UNTUK PREDIKSI RISIKO PENYAKIT LIVER DENGAN RANDOM FOREST TEROPTIMASI

Rizky Andrea Arifa[*1], Nana Suarna[2], Agus Bahtiar[3] , Nining Rahaningsih[4] , Willy Prihartono[5]

[12345]STMIK IKMI CIREBON
Jl. Perjuangan No.10B, Karyamulya, Kec. Kesambi, Kota Cirebon, Jawa Barat 45135

Email: [1]rizkyandreaarifa@gmail.com,[2]st_nana@yahoo.com, [3]agusbahtiar038@gmail.com, [4]niningr157@yahoo.co.id, [5]willyprihartono@gmail.com

ABSTRAK

Penyakit liver merupakan salah satu kondisi kronis dengan tingkat mortalitas tinggi, sehingga diperlukan pendekatan prediksi yang akurat untuk mendukung deteksi dini. Penelitian ini bertujuan mengembangkan model *machine learning* untuk memprediksi risiko penyakit liver menggunakan algoritma *Random Forest* yang dioptimalkan dengan *RandomizedSearchCV*. Dataset yang digunakan terdiri dari 1.700 entri yang mencakup variabel klinis dan gaya hidup, termasuk usia, jenis kelamin, BMI, konsumsi alkohol, kebiasaan merokok, riwayat genetik, aktivitas fisik, diabetes, hipertensi, serta hasil *Liver Function Test*. Proses penelitian meliputi *preprocessing*, normalisasi skala, pembagian data menggunakan *train-test split* 80:20, pembangunan model baseline, dan optimasi hiperparameter. Hasil eksperimen menunjukkan bahwa optimasi menghasilkan peningkatan performa model, dengan akurasi 0.91, peningkatan recall sebesar 3.20%, dan AUC-ROC mencapai 0.96. Analisis *feature importance* menunjukkan bahwa *LiverFunctionTest*, BMI, dan *AlcoholConsumption* merupakan fitur paling berpengaruh terhadap prediksi risiko penyakit liver. Dengan demikian, *Random Forest* teroptimasi terbukti efektif dalam menghasilkan model prediksi yang akurat dan dapat digunakan sebagai alat pendukung keputusan dalam deteksi dini penyakit liver.

Kata Kunci: *Machine Learning, Penyakit liver, Random Forest, RandomizedSearchCV, Klasifikasi.*

1. PENDAHULUAN

Perkembangan teknologi informasi dalam bidang kesehatan telah memberikan dampak signifikan terhadap peningkatan kualitas diagnosis, efisiensi pelayanan, serta kemampuan prediksi terhadap berbagai penyakit kronis. Salah satu penyakit yang menjadi perhatian global adalah penyakit liver, yang prevalensinya terus meningkat dan berdampak besar terhadap morbiditas maupun mortalitas [1]. Seiring dengan meningkatnya kasus penyakit liver, kebutuhan akan metode prediksi yang cepat, akurat, dan non-invasif semakin mendesak. Berbagai biomarker serum dan skor klinis telah dikembangkan untuk mendukung evaluasi kondisi hepatologis pasien, termasuk dalam mendeteksi steatosis, fibrosis, serta gangguan fungsi hati lainnya [2], [3], [4].

Di sisi lain, kemajuan teknologi kecerdasan buatan dan *machine learning* telah membuka peluang besar dalam pengolahan data medis. Berbagai penelitian menunjukkan bahwa model *machine learning* mampu mengidentifikasi pola kompleks dalam data klinis dan laboratorium sehingga dapat meningkatkan ketepatan prediksi dan mendukung pengambilan keputusan klinis [5], [6],

[14]. Model berbasis *machine learning* seperti Logistic Regression, SVM, K-Nearest Neighbor, dan Decision Tree telah banyak digunakan dalam prediksi penyakit liver. Namun, tantangan utama dalam penerapannya adalah ketergantungan model terhadap kualitas fitur, teknik preprocessing, serta pemilihan hyperparameter yang tepat [7], [8], [9].

Random Forest merupakan algoritma *ensemble* yang banyak digunakan dalam bidang medis karena stabilitas, ketahanan terhadap overfitting, serta kemampuannya mengolah data klinis yang bersifat heterogen [10], [11]. Selain menghasilkan performa yang baik, Random Forest juga menyediakan *feature importance* yang membantu dalam mengidentifikasi faktor-faktor klinis yang berpengaruh terhadap risiko penyakit liver. Namun, performa model ini sangat dipengaruhi oleh konfigurasi hyperparameter, sehingga diperlukan teknik optimasi yang tepat untuk meningkatkan akurasi prediksi [12], [13].

Metode optimasi hyperparameter seperti *RandomizedSearchCV* terbukti efektif untuk menemukan kombinasi parameter terbaik dengan waktu komputasi lebih efisien dibanding metode grid search [14]. Berbagai penelitian terkini juga melaporkan bahwa penerapan optimasi hyperparameter dapat meningkatkan akurasi model dalam memprediksi penyakit liver dan penyakit kronis lainnya [15], [16], [17].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengembangkan model prediksi risiko penyakit liver menggunakan algoritma Random Forest yang dioptimasi dengan RandomizedSearchCV, serta mengevaluasi performanya berdasarkan akurasi, precision, recall, dan F1-score. Selain itu, penelitian ini juga menganalisis *feature importance* untuk mengidentifikasi variabel-variabel klinis yang paling berpengaruh terhadap prediksi. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengembangan sistem prediksi medis yang akurat, interpretable, dan dapat mendukung deteksi dini penyakit liver

2. LANDASAN TEORI

2.1 Penyakit Liver

Penyakit liver mencakup berbagai kondisi yang memengaruhi fungsi hati, termasuk steatosis, fibrosis, sirosis, dan kerusakan hepatoseluler. Kondisi ini dapat dipicu oleh gaya hidup tidak sehat, gangguan metabolik, infeksi, atau faktor genetik. Biomarker non-invasif seperti ALT, AST, bilirubin, dan skor klinis lainnya berperan penting dalam deteksi dan pemantauan penyakit liver [2], [3], [4].

2.2 Machine Learning dalam Kesehatan

Machine learning merupakan pendekatan komputasi yang memungkinkan sistem mempelajari pola dari data dan menghasilkan prediksi secara otomatis. Dalam bidang kesehatan, machine learning digunakan untuk diagnosis, prediksi penyakit, analisis data klinis, serta pengembangan sistem pendukung keputusan medis [8], [6], [1]. Metode ini memberikan keunggulan karena mampu mengolah data kompleks secara cepat dan presisi.

2.3 Algoritma Random Forest

Random Forest adalah algoritma *ensemble learning* berbasis Decision Tree yang bekerja dengan menggabungkan banyak pohon untuk meningkatkan akurasi dan stabilitas prediksi. Model ini efektif dalam menangani data klinis karena mampu mengurangi overfitting dan memberikan analisis *feature importance* untuk melihat kontribusi setiap variabel [10].

2.4 Hyperparameter Optimization

Hyperparameter merupakan parameter yang mengatur proses pelatihan model, seperti jumlah pohon (*n_estimators*), kedalaman maksimum pohon (*max_depth*), dan jumlah fitur dalam pemisahan node. Pemilihan hyperparameter yang tepat dapat meningkatkan akurasi model secara signifikan. Berbagai studi melaporkan bahwa teknik optimasi hyperparameter berperan penting dalam meningkatkan performa model *machine learning* [12],

2.5 RandomizedSearchCV

RandomizedSearchCV adalah metode optimasi hyperparameter yang memilih kombinasi parameter secara acak dari ruang pencarian tertentu. Teknik ini lebih efisien dibandingkan GridSearchCV karena tidak menguji semua kemungkinan kombinasi, sehingga lebih cepat namun tetap menghasilkan konfigurasi optimal [24], [14].

2.6 Feature Importance

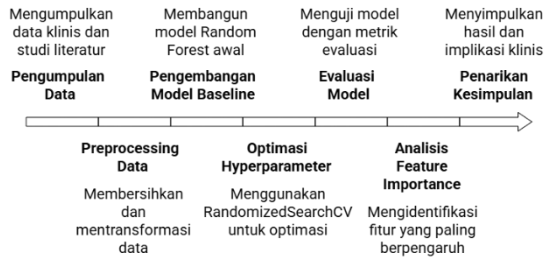
Feature importance digunakan untuk mengidentifikasi fitur atau variabel yang memberikan pengaruh terbesar pada hasil prediksi. Dalam konteks prediksi penyakit liver, analisis ini penting untuk mengetahui biomarker mana yang paling signifikan dalam menentukan risiko penyakit [11].

3. METODOLOGI PENELITIAN

Penelitian ini menggunakan pendekatan eksperimental (*experimental research*) untuk mengembangkan model prediksi penyakit liver berbasis algoritma Random Forest dengan optimasi hyperparameter menggunakan RandomizedSearchCV. Tahapan penelitian disusun secara sistematis mulai dari pengumpulan data hingga analisis hasil.

3.1 Tahapan Penelitian

Tahapan penelitian yang dilakukan dalam pengembangan model prediksi penyakit liver ditunjukkan pada Gambar 1 berikut.



Gambar 1 Tahapan Penelitian

3.1.1 Pengumpulan Data

Dataset penyakit liver diperoleh dari sumber data publik yang berisi variabel klinis seperti usia, jenis kelamin, enzim hati (ALT, AST), bilirubin, albumin, dan parameter laboratorium lainnya. Pengumpulan data dilakukan melalui:

1. Studi Literatur, untuk memahami teori terkait penyakit liver, machine learning, dan algoritma yang digunakan [2].
2. Dataset Sekunder, yaitu data siap pakai yang umum digunakan dalam penelitian kesehatan berbasis machine learning [1].

3.1.2 Preprocessing Data

Pada tahap ini dilakukan proses:

1. Pembersihan data (handling missing values dan outlier).
2. Transformasi data sesuai kebutuhan model.
3. Penanganan ketidakseimbangan kelas bila ditemukan (oversampling/undersampling).
4. Pembagian data menjadi data latih dan uji.

Preprocessing sangat penting karena kualitas data memengaruhi performa model secara signifikan.

3.1.3 Pengembangan Model Baseline

Model awal (baseline) dibangun menggunakan algoritma Random Forest dengan parameter default. Tahap ini bertujuan melihat performa awal model tanpa optimasi [10].

3.1.4 Optimasi RandomizedSearchCV

Optimasi dilakukan untuk meningkatkan performa model dengan menguji kombinasi hyperparameter secara acak, seperti:

1. n_estimators
2. max_depth
3. max_features
4. min_samples_split
5. min_samples_leaf

RandomizedSearchCV dipilih karena lebih efisien dalam waktu komputasi dibanding GridSearchCV namun tetap memberikan hasil yang optimal [13].

3.1.5 Evaluasi Model

Model hasil optimasi diuji menggunakan beberapa metrik evaluasi yang umum digunakan dalam model prediksi penyakit, yaitu:

1. Akurasi
2. Recall
3. AUC ROC
4. Confusion matrix

Metrik ini direkomendasikan dalam penelitian medis untuk menilai kualitas model prediksi [11].

3.1.6 Analisis Feature Importance

Tahap ini bertujuan mengetahui fitur mana yang memiliki kontribusi terbesar terhadap keputusan model. Analisis ini penting dalam konteks kesehatan untuk memahami variabel klinis yang paling memengaruhi risiko penyakit liver [4].

3.1.7 Penarikan Kesimpulan

Tahap akhir berupa penyimpulan hasil penelitian, performa model, dan implikasi klinis berdasarkan analisis model dan data.

4. HASIL DAN PEMBAHASAN

4.1 Hasil

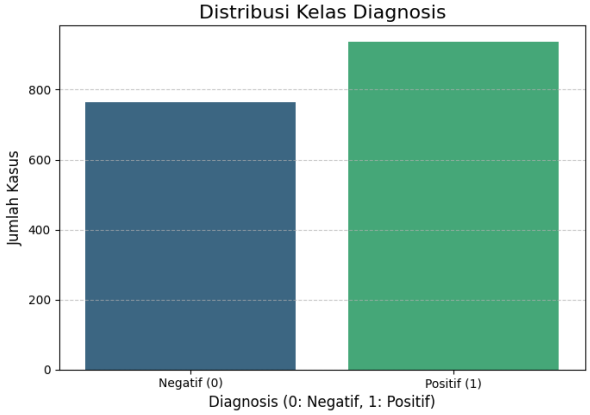
4.1.1 Pengumpulan Data

Dataset penyakit liver diperoleh dari sumber publik dengan total 1.700 entri yang berisi variabel klinis dan perilaku. Pada tahap ini dilakukan eksplorasi untuk memahami karakteristik data sebelum dilakukan preprocessing.

4.1.2 Preprocessing Data

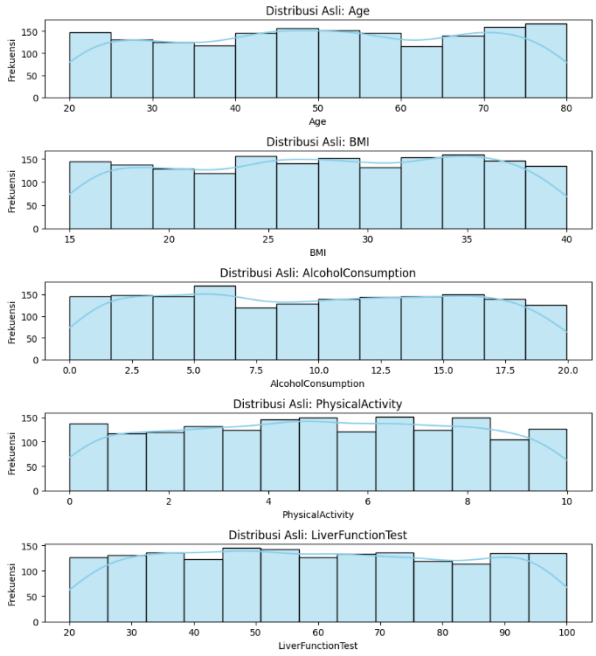
1. Analisis Distribusi Kelas Target: Gambar 2 menunjukkan visualisasi dari variabel target, Diagnosis. Keseimbangan atau ketidakseimbangan kelas ini dipertimbangkan

selama pembagian data menggunakan skema stratifikasi.



Gambar 2. Distribusi Kelas Target (Diagnosis)

2. Justifikasi Normalisasi Skala: Fitur-fitur numerik seperti Age, BMI, AlcoholConsumption, dan LiverFunctionTest memiliki rentang dan distribusi yang sangat bervariasi. Gambar 2 menunjukkan visualisasi gabungan distribusi fitur numerik awal, yang memperjelas adanya perbedaan skala yang ekstrem dan kemiringan data (*skewness*) signifikan. Hal ini membenarkan penggunaan StandardScaler untuk menstandarisasi data, mencegah fitur berskala besar mendominasi model.

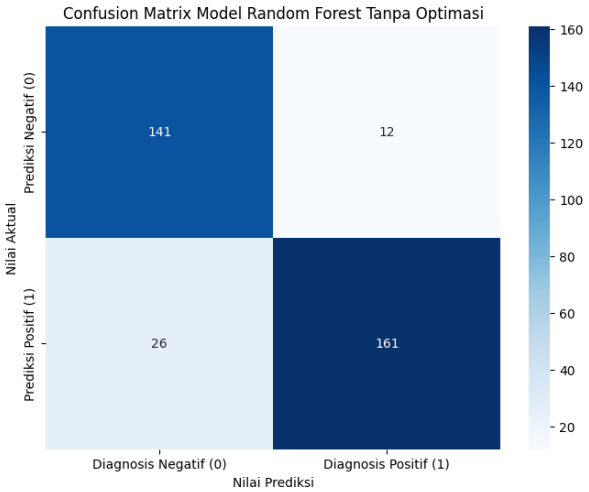


Gambar 3. Distribusi Gabungan Fitur Numerik Sebelum Normalisasi.

Fitur numerik seperti Age, BMI, AlcoholConsumption, dan LiverFunctionTest menunjukkan skala yang berbeda jauh dan distribusi yang tidak normal (*skewed*). Situasi ini sejalan dengan temuan Erol & Uzba tahun 2022 bahwa normalisasi skala penting untuk mencegah dominasi fitur berskala besar dalam model prediksi medis [16].

4.1.3 Hasil Model Baseline

Model baseline diuji menggunakan parameter default. Hasil confusion matrix ditampilkan pada Gambar 4.

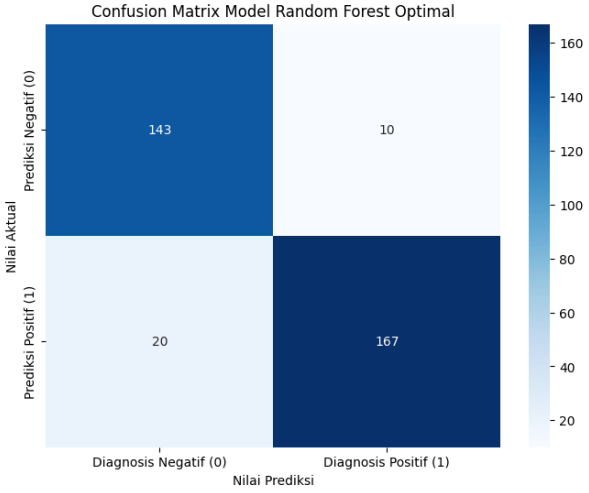


Gambar 4. Confusion Matrix Model Baseline

Model baseline (Random Forest default) menghasilkan *confusion matrix* seperti pada Gambar 4. Performa awal ini sejalan dengan laporan penelitian Dritsas & Trigka (2023), yang menyatakan bahwa Random Forest memberikan baseline akurasi yang baik pada data penyakit liver tanpa optimasi awal [18].

4.1.4 Hasil Model Setelah Optimasi

Setelah optimasi dengan RandomizedSearchCV, performa model meningkat. Confusion matrix model teroptimasi ditampilkan pada Gambar 5.



Gambar 5. Confusion Matrix Model Teroptimasi

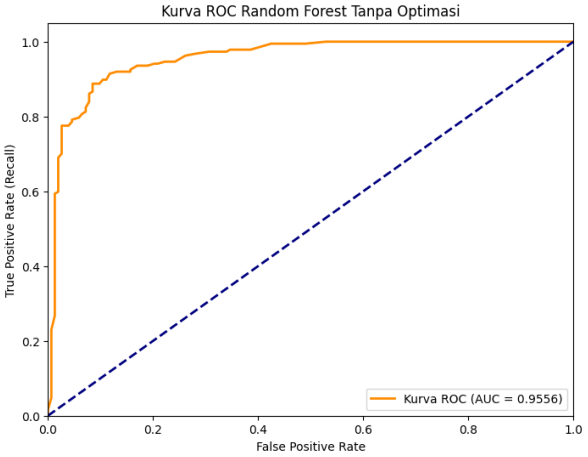
4.1.5. Evaluasi Model

Tabel 1 menyajikan perbandingan kinerja metrik utama, menunjukkan peningkatan yang konsisten setelah optimasi

Metrik	Model Baseline	Model teroptimasi	Peningkatan
Akurasi	0.88	0.91	+2.36%
Recall	0.86	0.89	+3.20%
AUC-ROC Score	0.95	0.96	+0.45%

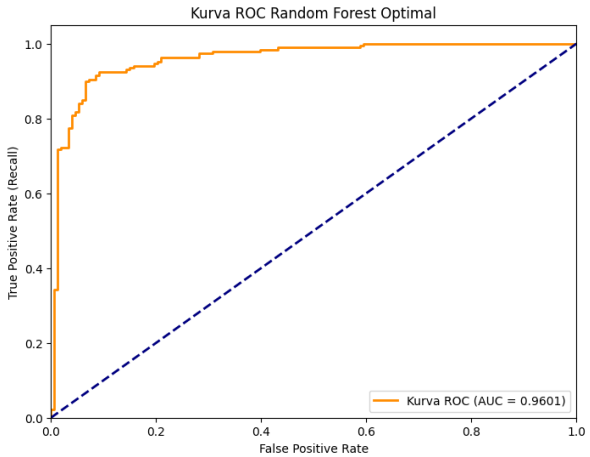
Peningkatan kapabilitas diskriminasi (AUC) divisualisasikan pada Kurva ROC Gambar 6 dan Gambar 7 :

Kurva ROC model baseline dapat dilihat pada Gambar 6.



Gambar 6. Kurva ROC Model Baseline

Kurva ROC model teroptimasi dapat dilihat pada Gambar 7.

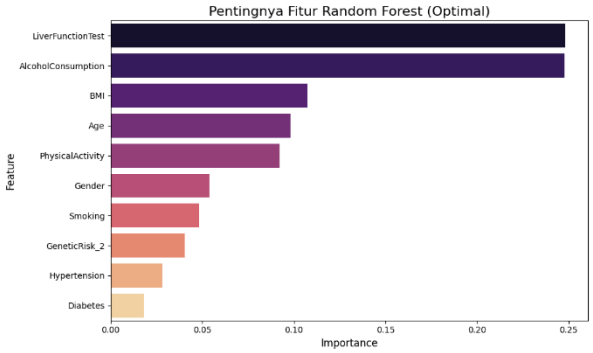


Gambar 7. Kurva ROC Model Teroptimasi

Peningkatan AUC-ROC mengindikasikan meningkatnya kemampuan diskriminatif model. Hal ini konsisten dengan temuan Hong tahun 2022 , yang menunjukkan bahwa optimasi parameter pada Random Forest meningkatkan AUC pada prediksi penyakit akut [19].

4.1.6 Analisis Feature Importance

Random Forest dikenal memiliki keunggulan dalam mengidentifikasi fitur klinis yang paling relevan untuk prediksi penyakit kronis Tahap ini bertujuan mengetahui fitur mana yang memiliki kontribusi terbesar terhadap keputusan model. *Feature importance* model teroptimasi ditampilkan pada Gambar 8 berikut.



Gambar 8. Feature Importance Algoritma Random Forest Teroptimasi.

4.2 Pembahasan

Hasil penelitian menunjukkan bahwa optimasi *hyperparameter* menggunakan RandomizedSearchCV meningkatkan kinerja model. Peningkatan nilai AUC-ROC menjadi 0,9601 (Gambar 6) membuktikan bahwa model teroptimasi memiliki kemampuan diskriminasi yang sangat tinggi dan stabil.

Secara klinis, peningkatan Recall (+3.20%) adalah hasil yang paling kritis. Peningkatan ini didorong oleh pengurangan False Negatives (dari 26 menjadi 20, lihat Gambar 4 dan 5), memastikan bahwa pasien yang positif risiko penyakit liver lebih kecil kemungkinannya untuk terlewatkan, mendukung intervensi kesehatan secara lebih cepat.

Implikasi Klinis dari Feature Importance

Analisis *feature importance* (Gambar 8) selaras dengan pengetahuan klinis, mengidentifikasi faktor-faktor risiko utama:

1. Prioritas Indikator Biokimia: LiverFunctionTest menduduki peringkat teratas, menegaskan peran utamanya sebagai *biomarker* deteksi kerusakan hepatik.
2. Faktor Risiko Gaya Hidup: Kontribusi signifikan dari AlcoholConsumption dan BMI menyoroti bahwa model secara efektif menggabungkan

risiko yang berasal dari gaya hidup (ALD dan NAFLD) dalam proses prediksinya.

Penelitian ini menunjukkan bahwa Random Forest teroptimasi mampu memberikan performa prediksi yang lebih baik dan dapat dijadikan model pendukung keputusan dalam identifikasi risiko penyakit liver.

5. PENUTUP

Penelitian ini dilakukan untuk mengembangkan model prediksi risiko penyakit liver berbasis *machine learning* dengan memanfaatkan algoritma *Random Forest* yang dioptimalkan melalui *RandomizedSearchCV*. Hasil yang diperoleh menunjukkan bahwa pendekatan optimasi hiperparameter mampu meningkatkan kemampuan model dalam mengolah data klinis yang kompleks, sekaligus memperkuat pemahaman mengenai variabel-variabel kesehatan yang paling relevan terhadap risiko penyakit liver. Kontribusi utama penelitian ini terletak pada integrasi metode optimasi model dengan analisis *feature importance*, sehingga tidak hanya menghasilkan model prediktif yang kuat, tetapi juga menyediakan wawasan interpretatif yang penting bagi praktisi kesehatan dalam memahami faktor risiko yang dominan. Pendekatan ini memperkaya literatur mengenai penerapan *machine learning* untuk deteksi dini penyakit kronis, khususnya dalam konteks penyakit liver yang membutuhkan diagnosis cepat dan penilaian risiko yang akurat.

Secara praktis, model yang dikembangkan berpotensi diimplementasikan dalam sistem pendukung keputusan klinis untuk membantu tenaga kesehatan melakukan skrining awal terhadap pasien berisiko tinggi. Integrasi lebih lanjut ke dalam platform kesehatan digital atau aplikasi pemantauan kesehatan juga memungkinkan peningkatan aksesibilitas bagi masyarakat. Untuk penelitian selanjutnya, perlu dilakukan pengujian menggunakan dataset yang lebih beragam dan representatif secara demografis agar model memiliki generalisasi yang lebih kuat. Selain itu, pengembangan analisis interpretasi yang lebih mendalam seperti SHAP atau LIME dapat memberikan pemahaman yang lebih kaya terkait mekanisme keputusan model. Ekstensi lain yang layak dieksplorasi meliputi penerapan model ke data longitudinal, integrasi biomarker tambahan, serta perbandingan dengan algoritma *deep learning* untuk mengevaluasi potensi peningkatan performa pada skala data yang lebih besar.

DAFTAR PUSTAKA

[1] H. Devarbhavi, S. K. Asrani, J. P. Arab, Y. A. Nartey, E. Pose, and P. S. Kamath, "Global burden of liver disease: 2023 update," *J. Hepatol.*, vol. 79, no. 2, pp. 516–537, 2023, doi: 10.1016/j.jhep.2023.03.017.

[2] F. Abdelhameed *et al.*, "Non-invasive scores and serum biomarkers for fatty liver disease," *Curr. Obes. Rep.*, vol. 13, no. 1, pp. 17–29, 2024, doi: 10.1007/s13679-024-00574-z.

[3] Q. M. Anstee, L. Castera, and R. Loomba, "Review Impact of non-invasive biomarkers on hepatology practice : Past , present and future," *J. Hepatol.*, vol. 76, no. 6, pp. 1362–1378, 2022, doi: 10.1016/j.jhep.2022.03.026.

[4] M. Iwasa, A. Hiraoka, and M. Kumagai, "Update on blood-based biomarkers for chronic liver diseases," *Gut Liver*, vol. 15, no. 3, pp. 318–329, 2021, doi: 10.5009/gnl21058.

[5] A. Al-Nafjan, A. Aljuhani, A. Alshebel, A. Alharbi, and A. Alshehri, "Artificial intelligence in predictive healthcare: A systematic review," *J. Clin. Med.*, vol. 14, no. 19, p. 6752, 2025, doi: 10.3390/jcm14196752.

[6] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges," *Sensors*, vol. 23, no. 9, p. 4178, 2023, doi: 10.3390/s23094178.

[7] Y. A. Ali, E. M. Awwad, and M. Al-razgan, "Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity," *Process. 2023*, vol. 11, no. 2, p. 349, 2023, doi: https://doi.org/10.3390/pr11020349.

[8] L. Alzubaidi, J. Zhang, A. J. Humaidi, and Y. Duan, "Review of machine learning models in medical diagnosis and their data processing requirements," *Front. Artif. Intell.*, vol. 4, p. 759972, 2021, doi: 10.3389/frai.2021.759972.

[9] D. Fernandes Prabhu, V. Gurupur, A. Stone, and E. Trader, "Integrating artificial intelligence, electronic health records, and wearables for predictive, patient-centered decision support in healthcare," *Healthcare*, vol. 13, no. 21, p. 2753, 2025, doi: 10.3390/healthcare13212753.

[10] F. Cappelli, G. Castronuovo, S. Grimaldi, and V. Telesca, "Random Forest and Feature Importance Measures for Discriminating the Most Influential Environmental Factors in

Predicting Cardiovascular and Respiratory Diseases," *Int. J. Environ. Res. Public Health*, vol. 21, no. 7, 2024, doi: 10.3390/ijerph21070867.

[11] S. M. Ganie, S. Ahmed, and N. Khan, "Improved liver disease prediction from clinical data through ensemble learning methods," *BMC Med. Inform. Decis. Mak.*, vol. 24, p. 160, 2024, doi: 10.1186/s12911-024-02550-y.

[12] W. El Atifi, O. El Rhazouani, F. M. Khan, and H. Sekkat, "Optimizing ensemble machine learning models for accurate liver disease prediction in healthcare," *PLoS One*, vol. 20, no. 8, p. e0330899, 2025, doi: 10.1371/journal.pone.0330899.

[13] N. M. F. S. Raiaan, Mohaimenul Azam Khan, Sadma, "A systematic review of hyperparameter optimization techniques in convolutional neural networks," *Decis. Anal. J.*, vol. 11, p. 100470, 2024, doi: 10.1016/j.dajour.2024.100470.

[14] O. Rainio, "Evaluation metrics and statistical tests for machine learning," *Sci. Rep.*, vol. 14, p. 56706, 2024, doi: 10.1038/s41598-024-56706-x.

[15] M. C. Bragazzi, R. Venere, G. Andriollo, and L. Ridola, "Diagnostic Applications of Artificial Intelligence in Liver Diseases," *J. Clin. Med.*, pp. 1–12, 2025, doi: https://doi.org/ 10.3390/jcm14176231.

[16] G. Erol, B. Uzbaş, and Ş. Yücelbaş, "Analysing the effect of data preprocessing techniques using machine learning algorithms on COVID-19 diagnosis," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 28, p. e7393, 2022, doi: 10.1002/cpe.7393.

[17] L. J. Kanbar, B. Wissel, Y. Ni, N. Pajor, T. Glauser, and J. W. Dexheimer, "Implementation of Machine Learning Pipelines for Clinical Practice : Development and Validation Study Corresponding Author :," vol. 10, pp. 1–9, 2022, doi: 10.2196/37833.

[18] E. Dritsas and M. Trigka, "Supervised Machine Learning Models for Liver Disease Risk Prediction," *Comput. Artic.*, 2023.

[19] W. Hong *et al.*, "Usefulness of Random Forest Algorithm in Predicting Severe Acute Pancreatitis," *Front. Cell. Infect. Microbiol.*, vol. 12, no. June, pp. 1–14, 2022, doi: 10.3389/fcimb.2022.893294.