

Peran Explainable Machine Learning dalam Mendukung Pengambilan Keputusan yang Transparan dan Berkeadilan

Wiki Lofandri¹, Rizki Adam², Ifnalia Rahayu³

Afiliasi (Vokasi¹², Ilmu Kesejahteraan Keluarga³ - Universitas Negeri Padang ¹²³)

Alamat (Jl. Prof. Dr. Hamka, Air Tawar, Padang, Sumatera Barat, dengan kode pos 25132, Kota Padang, Indonesia)
wiloleaks@unp.ac.id¹, rioneeradam@gmail.com², ifnalia@unp.ac.id³

Received: 24 Februari 2026 | Revised: 27 Agustus 2026

Accepted: 28 Agustus 2026 | Published: 28 Agustus 2026

Abstrak

Pemanfaatan machine learning dalam mendukung pengambilan keputusan semakin meluas, namun kompleksitas sebagian model menyulitkan pengguna memahami alasan di balik prediksi sehingga menimbulkan tantangan transparansi, akuntabilitas, dan fairness, terutama ketika keluaran model memengaruhi individu atau kelompok. Penelitian ini menganalisis peran Explainable Machine Learning (XAI) dalam mendukung transparansi dan evaluasi fairness pada pengambilan keputusan berbasis machine learning. Penelitian menggunakan pendekatan kajian konseptual dan deskriptif-analitis melalui sintesis literatur mengenai interpretabilitas model, metode explainability, fairness, dan pengambilan keputusan yang berpusat pada manusia, tanpa pengujian empiris pada dataset tertentu. Hasil sintesis menunjukkan bahwa XAI dapat memperjelas hubungan antara fitur dan prediksi, mendukung pemeriksaan keputusan pada tingkat individual, serta menyediakan informasi untuk mengidentifikasi potensi bias. Namun, explainability tidak secara otomatis menjamin fairness; penilaian fairness tetap memerlukan metrik dan analisis antarkelompok yang sesuai konteks. Dengan demikian, XAI lebih tepat diposisikan sebagai salah satu komponen dalam kerangka pengambilan keputusan yang transparan, dapat diaudit, dan berorientasi pada manusia.

I. PENDAHULUAN

Pemanfaatan machine learning dalam mendukung pengambilan keputusan berbasis data terus berkembang seiring dengan meningkatnya ketersediaan data dan kemampuan komputasi. Pendekatan ini telah diterapkan pada berbagai bidang, seperti kesehatan, pendidikan, keuangan,

Kata kunci: Explainable Machine Learning; Fairness; Transparansi

Abstract

The use of machine learning to support decision-making is increasingly widespread, yet the complexity of some models makes it difficult for users to understand the reasons behind predictions, creating challenges of transparency, accountability, and fairness, particularly when model outputs affect individuals or groups. This study analyzes the role of Explainable Machine Learning (XAI) in supporting transparency and fairness assessment in machine-learning-based decision-making. It adopts a conceptual and descriptive-analytical approach through a synthesis of literature on model interpretability, explainability methods, fairness, and human-centered decision-making, without empirical testing on any specific dataset. The synthesis indicates that XAI can clarify relationships between features and predictions, support the examination of individual decisions, and provide information relevant to identifying potential bias. However, explainability does not automatically guarantee fairness; fairness still requires appropriate metrics and group-based analysis. Therefore, XAI is best positioned as one component of a transparent, auditable, and human-centered decision-making framework.

Keywords: Explainable Machine Learning; Fairness; Transparency

industri, serta kebijakan publik, untuk membantu proses analisis, prediksi, dan rekomendasi secara lebih sistematis. Dengan memanfaatkan pola yang dipelajari dari data historis, machine learning mampu memberikan dukungan keputusan yang lebih konsisten dibandingkan pendekatan konvensional [1], [2].

Namun demikian, peningkatan kinerja prediksi sering kali dicapai melalui penggunaan model dengan tingkat kompleksitas yang tinggi. Model-model tersebut cenderung bersifat black-box, yaitu proses pengambilan keputusannya sulit dipahami oleh pengguna. Kurangnya pemahaman terhadap cara kerja model dapat menimbulkan permasalahan transparansi dan akuntabilitas, terutama ketika hasil prediksi digunakan sebagai dasar keputusan yang berdampak langsung terhadap individu maupun kelompok masyarakat [3], [4].

Selain itu, keputusan yang dihasilkan model machine learning berpotensi mengandung bias yang berasal dari data maupun proses pembelajaran. Bias tersebut dapat memengaruhi keadilan (fairness) dalam pengambilan keputusan, khususnya pada sistem yang digunakan untuk mendukung keputusan penting. Tanpa mekanisme penjelasan yang memadai, bias dan ketidakseimbangan tersebut sulit diidentifikasi dan dievaluasi secara objektif [5]. Oleh karena itu, diperlukan pendekatan yang tidak hanya berfokus pada akurasi prediksi, tetapi juga pada kemampuan model untuk menjelaskan dasar pengambilan keputusannya.

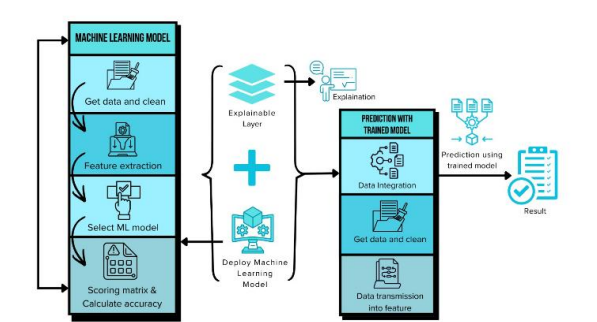
Explainable Machine Learning (XAI) merupakan pendekatan yang bertujuan meningkatkan interpretabilitas model machine learning dengan memberikan penjelasan mengenai kontribusi fitur terhadap hasil prediksi. Melalui pendekatan ini, pengguna dapat memperoleh pemahaman yang lebih baik mengenai faktor-faktor yang memengaruhi keputusan model. XAI menjadi penting dalam konteks sistem pendukung keputusan karena mampu meningkatkan kepercayaan pengguna serta mendukung proses evaluasi keputusan secara lebih transparan dan bertanggung jawab [6], [7].

Berbagai penelitian sebelumnya telah membahas penerapan machine learning maupun teknik explainability pada domain tertentu. Namun, kajian yang menempatkan Explainable Machine Learning sebagai pendekatan umum lintas bidang untuk mendukung pengambilan keputusan yang adil dan transparan masih perlu diperkuat. Pendekatan lintas domain

memungkinkan pemahaman yang lebih menyeluruh mengenai peran explainability sebagai fondasi keadilan dan transparansi dalam sistem pengambilan keputusan berbasis data.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan menganalisis bagaimana Explainable Machine Learning dapat mendukung transparansi dan evaluasi fairness dalam pengambilan keputusan berbasis machine learning. Secara khusus, kajian ini memetakan bentuk penjelasan global dan lokal, metode explainability yang relevan, keterkaitannya dengan evaluasi fairness, serta implikasinya terhadap pengawasan dan pengambilan keputusan oleh manusia. Kontribusi kajian terletak pada penyusunan kerangka konseptual yang menghubungkan explainability, fairness, transparansi, dan human oversight dalam satu alur analisis.

Kerangka konseptual yang menghubungkan data dan model machine learning dengan teknik explainability, evaluasi fairness, interpretasi manusia, dan keputusan ditunjukkan pada Gambar 1. Kerangka tersebut menempatkan XAI sebagai mekanisme penjelasan yang mendukung pemeriksaan keputusan, bukan sebagai jaminan bahwa suatu keputusan secara otomatis telah memenuhi prinsip fairness.



Gambar 1. Kerangka Konseptual Peran Explainable Machine Learning dalam Pengambilan Keputusan
Sumber: Olahan penulis (2026)

Posisi dan kontribusi kajian ini dibandingkan dengan penelitian terdahulu diringkaskan pada Tabel 1.

TABEL 1. POSISI DAN KONTRIBUSI KAJIAN

Aspek	Kajian Terdahulu	Kajian Ini
Fokus	Interpretabilitas, explainability, dan fairness umumnya dikaji secara terpisah [8], [9], [5]	Menghubungkan explainability, fairness, transparansi, dan human oversight dalam satu kerangka konseptual

Aspek	Kajian Terdahulu	Kajian Ini
Transparansi	Ditekankan sebagai syarat kepercayaan dan akuntabilitas model [3], [4], [7]	Dianalisis sebagai bagian dari proses pengawasan keputusan (human oversight)
Fairness	Difokuskan pada sumber bias dan metrik pengukuran keadilan [5], [10], [11]	Dihubungkan dengan informasi yang dihasilkan XAI untuk mendukung pemeriksaan potensi bias
Explainability	Dibedakan atas penjelasan global/lokal serta metode spesifik seperti SHAP, LIME, dan counterfactual [12], [13], [14]	Dipetakan berdasarkan fungsi dan penggunaannya dalam pengambilan keputusan
Cakupan	Sering berfokus pada satu metode atau satu domain tertentu [15], [16]	Kerangka konseptual yang dapat diterapkan lintas konteks

Sumber: Olahan penulis (2026)

II. TINJAUAN PUSTAKA

A. Machine Learning dalam Pengambilan Keputusan

Machine learning merupakan pendekatan komputasi yang memungkinkan sistem mempelajari pola dari data dan menghasilkan prediksi atau rekomendasi untuk mendukung pengambilan keputusan. Pendekatan ini banyak digunakan karena kemampuannya menangani data berskala besar dan kompleks serta membantu pengambil keputusan menganalisis data historis secara lebih sistematis [1], [2].

Namun, peningkatan performa prediksi sering kali dicapai melalui penggunaan model dengan tingkat kompleksitas yang tinggi. Model-model tersebut cenderung sulit dipahami oleh pengguna sehingga proses pengambilan keputusan menjadi kurang transparan. Beberapa studi menyoroti bahwa penggunaan model black-box dalam pengambilan keputusan dapat menimbulkan tantangan akuntabilitas dan kepercayaan pengguna terhadap sistem berbasis machine learning [7], [16].

B. Tantangan Transparansi dan Fairness dalam Machine Learning

Transparansi merupakan aspek penting dalam sistem pengambilan keputusan berbasis machine learning, terutama ketika sistem digunakan pada konteks yang berdampak langsung terhadap individu atau masyarakat. Kurangnya transparansi menyulitkan pengguna memahami dasar keputusan yang dihasilkan serta menghambat proses evaluasi dan pengawasan terhadap sistem [3], [7].

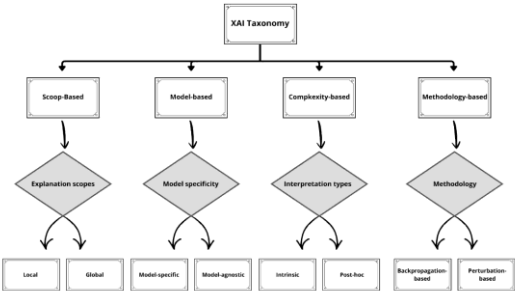
Selain transparansi, fairness atau keadilan menjadi isu krusial dalam pemanfaatan machine learning. Bias dalam model dapat muncul akibat ketidakseimbangan data, representasi fitur yang tidak memadai, maupun proses pembelajaran

model itu sendiri. Literatur menunjukkan bahwa bias algoritmik dapat berdampak signifikan terhadap kelompok tertentu jika tidak dievaluasi secara memadai [5], [10]. Oleh karena itu, upaya meningkatkan transparansi dan fairness perlu menjadi perhatian utama dalam pengembangan sistem pengambilan keputusan berbasis machine learning.

C. Explainable Machine Learning (XAI)

Explainable Machine Learning atau Explainable AI (XAI) merupakan pendekatan yang bertujuan meningkatkan interpretabilitas model machine learning dengan menyediakan penjelasan yang dapat dipahami oleh manusia. XAI memungkinkan pengguna mengetahui hubungan antara input dan output model, baik melalui penjelasan global yang menggambarkan perilaku umum model maupun penjelasan lokal yang menjelaskan keputusan pada satu instance tertentu [6].

Penelitian terbaru menunjukkan bahwa XAI tidak hanya berfungsi sebagai alat interpretasi teknis, tetapi juga sebagai komponen penting dalam pengembangan AI yang bertanggung jawab (responsible AI). Integrasi explainability dipandang mampu meningkatkan kepercayaan pengguna, transparansi sistem, serta mendukung evaluasi keputusan secara lebih akuntabel [17], [16].



Gambar 2. Taksonomi Pendekatan Explainable Machine Learning

Gambar 2 mengelompokkan pendekatan Explainable Machine Learning berdasarkan cakupan penjelasan, spesifisitas model, tingkat kompleksitas interpretasi, dan metodologi yang digunakan.

D. Explainable Machine Learning dalam Sistem Pendukung Keputusan

Sistem pendukung keputusan berbasis machine learning dirancang untuk membantu pengambil keputusan dengan menyediakan prediksi atau rekomendasi berbasis data. Integrasi Explainable Machine Learning dalam sistem pendukung keputusan memungkinkan pengguna memahami alasan di balik rekomendasi yang diberikan sehingga keputusan akhir tidak sepenuhnya bergantung pada model [12].

Beberapa studi menunjukkan bahwa explainability berperan penting dalam meningkatkan kualitas interaksi antara manusia dan sistem pendukung keputusan. Penjelasan yang diberikan oleh XAI membantu pengguna menilai keandalan prediksi serta mengidentifikasi potensi bias atau ketidakwajaran dalam keputusan model [16], [17]. Dengan demikian, XAI mendukung pengambilan keputusan yang lebih transparan dan bertanggung jawab.

E. Explainability, Fairness, dan Keberuan Penelitian

Literatur menekankan bahwa explainability dan fairness merupakan dua aspek yang saling berkaitan dalam pengembangan sistem pengambilan keputusan berbasis AI. Explainability menyediakan mekanisme untuk memahami bagaimana keputusan dihasilkan, sementara fairness berfokus pada dampak keputusan tersebut terhadap berbagai kelompok. Meskipun explainability tidak secara otomatis menjamin fairness, keberadaan penjelasan yang jelas dapat membantu mengidentifikasi dan mengevaluasi potensi bias dalam model [10], [11].

Berdasarkan tinjauan pustaka tersebut, penelitian ini menempatkan Explainable Machine Learning sebagai pendekatan utama dalam mendukung pengambilan keputusan yang adil dan transparan secara lintas bidang. Fokus penelitian tidak hanya pada kinerja prediksi, tetapi juga pada interpretabilitas model sebagai dasar evaluasi transparansi dan keadilan dalam pengambilan keputusan berbasis data.

III. METODE PENELITIAN

A. Desain dan Pendekatan Penelitian

Penelitian ini menggunakan pendekatan kajian konseptual dan deskriptif-analitis. Pendekatan tersebut dipilih karena tujuan penelitian adalah mensintesis konsep dan temuan literatur mengenai Explainable Machine Learning, interpretabilitas, fairness, transparansi, dan pengambilan keputusan yang berpusat pada manusia, bukan melakukan pengujian empiris pada dataset tertentu. Analisis dilakukan dengan mengidentifikasi konsep utama, membandingkan fungsi penjelasan global dan lokal, memetakan metode explainability yang relevan, serta menghubungkannya dengan kebutuhan evaluasi fairness dan pengawasan keputusan.

Sebagai kajian konseptual, penelitian ini tidak menggunakan dataset tertentu, tidak menerapkan algoritma pembelajaran tertentu (misalnya random forest atau XGBoost), dan tidak menjalankan metode XAI tertentu (misalnya SHAP atau LIME) secara empiris. Dataset, algoritma, dan metode explainability tersebut dibahas pada tataran konseptual sebagai objek kajian literatur untuk memahami fungsi dan karakteristiknya, bukan sebagai instrumen yang diujikan pada data. Konsekuensinya, penelitian ini tidak menghasilkan nilai akurasi model, nilai metrik fairness, maupun waktu pengujian, dan setiap penyebutan metode bersifat deskriptif-konseptual.

B. Sumber dan Kriteria Literatur

Sumber yang dianalisis mencakup literatur fundamental dan penelitian yang membahas explainable AI, interpretabilitas model, fairness, bias algoritmik, serta human-centered AI. Literatur dipilih berdasarkan relevansi terhadap tema transparansi, fairness, dan pengambilan keputusan, serta keterwakilan konsep dan metode yang banyak dirujuk dalam bidang ini. Sintesis dilakukan secara tematik agar hubungan antarkonsep dapat dirumuskan menjadi kerangka analisis yang dapat digunakan sebagai dasar penelitian empiris pada tahap berikutnya.

C. Tahapan Analisis

Analisis dilakukan melalui lima tahapan utama, yaitu:

- a. identifikasi permasalahan transparansi dan fairness pada keputusan berbasis machine learning;
- b. pemetaan konsep interpretabilitas dan explainability;

- c. pemetaan metode penjelasan global dan lokal;
- d. analisis keterkaitan explainability dengan fairness dan potensi bias; serta
- e. sintesis kerangka pengambilan keputusan yang menggabungkan penjelasan model, evaluasi fairness, dan human oversight.

Kelima tahapan tersebut menghasilkan kerangka konseptual yang menjelaskan posisi XAI dalam proses pengambilan keputusan. Kerangka ini tidak dimaksudkan sebagai hasil pengujian empiris, melainkan sebagai sintesis konseptual yang dapat menjadi dasar perancangan eksperimen pada domain tertentu.

D. Kerangka Analisis Explainability dan Fairness

Analisis difokuskan pada dua dimensi utama. Pertama, explainability, yang mencakup penjelasan global untuk memahami perilaku model secara keseluruhan dan penjelasan lokal untuk memahami alasan prediksi pada instance tertentu. Metode yang relevan dalam literatur antara lain SHAP, LIME, dan pendekatan counterfactual. Kedua, fairness, yang menilai apakah keluaran model menunjukkan perbedaan yang tidak semestinya antarkelompok. Pada penelitian empiris lanjutan, metrik fairness perlu dipilih sesuai konteks, misalnya demographic parity, equal opportunity, atau equalized odds.

Aspek analisis dalam kajian ini meliputi:

- a. identifikasi bentuk penjelasan global dan lokal;
- b. pemetaan metode XAI dan fungsi penjelasannya;
- c. identifikasi pendekatan dan metrik evaluasi fairness; serta
- d. analisis hubungan antara penjelasan model, potensi bias, dan human oversight.

Dengan demikian, keluaran metodologis penelitian berupa sintesis konsep dan kerangka analisis, bukan model prediktif atau nilai evaluasi numerik. Kerangka tersebut dapat diterapkan pada penelitian lanjutan dengan memilih dataset, model, metode XAI, atribut sensitif, dan metrik fairness yang sesuai dengan domain.

E. Pemetaan Metode Explainable Machine Learning

Literatur menunjukkan bahwa metode explainability dapat dibedakan berdasarkan cakupan penjelasan dan ketergantungannya terhadap model. Penjelasan global digunakan untuk memahami pola umum model, sedangkan penjelasan lokal digunakan untuk menjelaskan keputusan pada instance tertentu. SHAP menyediakan atribusi fitur berbasis kontribusi terhadap prediksi, LIME memberikan pendekatan lokal melalui model yang lebih sederhana di sekitar instance, sedangkan counterfactual menjelaskan perubahan yang diperlukan agar keluaran model berubah. Pemetaan ini digunakan untuk memahami fungsi masing-masing metode, bukan untuk menetapkan satu metode sebagai pilihan terbaik secara umum [12], [13], [14].

Secara konseptual, evaluasi model dan evaluasi explainability perlu dipisahkan. Kinerja prediksi dapat dinilai menggunakan metrik seperti accuracy, precision, recall, atau F1-score sesuai tugas, sedangkan kualitas penjelasan perlu dianalisis dari aspek kesesuaian terhadap model, konsistensi, stabilitas, dan keterpahaman pengguna. Pemisahan tersebut penting agar interpretabilitas tidak disamakan dengan kinerja prediksi.

IV. HASIL DAN PEMBAHASAN

A. Hasil Sintesis Literatur

Hasil penelitian berupa sintesis konseptual dari literatur mengenai explainability, fairness, dan pengambilan keputusan berbasis machine learning. Sintesis menunjukkan bahwa transparansi dan fairness merupakan kebutuhan yang berbeda tetapi saling berkaitan. Explainability membantu menjelaskan bagaimana model menghasilkan prediksi, sedangkan fairness menilai bagaimana keluaran tersebut berdampak pada kelompok yang berbeda.

Sejalan dengan desain penelitian yang bersifat konseptual, kajian ini tidak menyajikan nilai numerik performa model karena memang tidak dilakukan eksperimen pada dataset tertentu. Kontribusi utama kajian ini adalah kerangka hubungan antara penjelasan model, evaluasi fairness, dan human oversight yang dapat menjadi dasar bagi pengujian empiris pada penelitian lanjutan. Ringkasan kerangka analisis disajikan pada Tabel 2.

TABEL 2. RINGKASAN KERANGKA ANALISIS KONSEPTUAL

Tahap Analisis	Fokus Analisis	Hasil Sintesis
Identifikasi masalah	Transparansi, black-box, bias, dan fairness	Kebutuhan explainability dan evaluasi fairness
Pemetaan explainability	Global, lokal, SHAP, LIME, counterfactual	Fungsi dan karakteristik penjelasan
Pemetaan fairness	Perbandingan antarkelompok dan metrik fairness	Kerangka evaluasi potensi bias
Human oversight	Keterpahaman, pemeriksaan, dan penggunaan penjelasan	Pengambilan keputusan yang dapat diaudit
Sintesis	Integrasi XAI, fairness, transparansi, dan manusia	Kerangka konseptual pengambilan keputusan
Evaluasi transparansi	Analisis kualitas penjelasan	Dukungan keputusan yang transparan

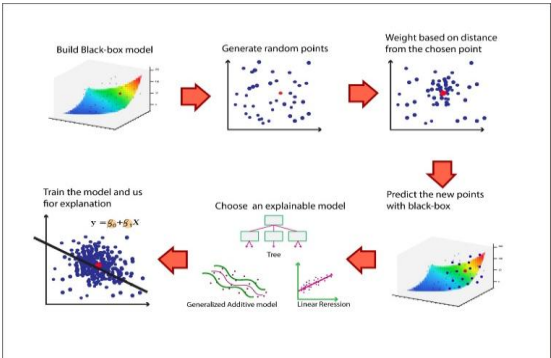
Sumber: Olahan penulis (2026)

Gambar 3. Ilustrasi Konseptual Penggunaan Explainable Machine Learning dalam Pengambilan Keputusan

B. Peran Explainable Machine Learning

Literatur menunjukkan dua bentuk penjelasan yang penting. Penjelasan global membantu memahami pola umum model dan fitur yang berkontribusi terhadap prediksi, sedangkan penjelasan lokal memberikan alasan yang lebih spesifik untuk satu instance. SHAP dan LIME merupakan contoh pendekatan yang dapat digunakan untuk menghasilkan atribusi atau penjelasan lokal, sedangkan pendekatan lain dapat digunakan untuk melihat perilaku model secara lebih luas. Counterfactual explanation memberikan perspektif berbeda dengan menunjukkan perubahan kondisi yang secara hipotetis dapat mengubah keluaran model.

Kombinasi penjelasan global dan lokal dapat memperkuat transparansi karena pengguna memperoleh informasi pada dua tingkat analisis. Namun, keterjelasan teknis belum tentu sama dengan keterpahaman pengguna. Oleh sebab itu, penjelasan perlu disesuaikan dengan kebutuhan dan konteks pengambilan keputusan. Ilustrasi konseptual penggunaan XAI dalam pengambilan keputusan ditunjukkan pada Gambar 3.



C. Keterkaitan Explainability dan Fairness

Sintesis literatur menunjukkan bahwa XAI dapat membantu mengidentifikasi fitur yang dominan dalam prediksi dan memberikan dasar untuk memeriksa kemungkinan bias. Informasi tersebut dapat menjadi titik awal audit terhadap model, terutama ketika fitur yang berkontribusi kuat berkaitan dengan atribut sensitif atau variabel yang menjadi proksi atribut sensitif.

Namun, explainability tidak secara otomatis membuktikan bahwa model telah memenuhi fairness. Sebuah model dapat memiliki penjelasan yang jelas tetapi tetap menghasilkan disparitas antarkelompok. Karena itu, pemeriksaan fairness perlu dilakukan secara terpisah menggunakan metrik dan prosedur yang sesuai dengan konteks. Hubungan yang lebih tepat adalah XAI membantu proses diagnosis dan audit, sedangkan fairness menyediakan kriteria untuk menilai dampak keputusan terhadap kelompok [11].

D. Implikasi terhadap Pengambilan Keputusan

Dalam pengambilan keputusan, hasil XAI dapat digunakan untuk membantu pengguna menilai apakah prediksi model memiliki alasan yang dapat dipahami dan relevan dengan konteks. Pada saat yang sama, hasil evaluasi fairness dapat menunjukkan apakah keluaran model memiliki perbedaan yang perlu diperiksa lebih lanjut. Kedua informasi tersebut dapat dipadukan dalam proses human oversight sebelum keputusan akhir ditetapkan.

Kerangka ini menempatkan manusia sebagai pihak yang menafsirkan informasi model, mempertimbangkan konteks, dan bertanggung jawab atas keputusan. Dengan demikian, XAI berfungsi sebagai mekanisme pendukung

transparansi dan audit, bukan sebagai pengganti penilaian manusia atau jaminan otomatis atas keadilan keputusan [18].

E. Diskusi Implikasi Lintas Bidang

Kerangka konseptual ini dapat diterapkan pada berbagai domain seperti kesehatan, pendidikan, keuangan, dan layanan publik, tetapi implementasinya tetap harus disesuaikan dengan karakteristik risiko dan tujuan keputusan. Pada setiap domain, pemilihan model, atribut sensitif, metode XAI, metrik fairness, dan bentuk penjelasan kepada pengguna perlu ditentukan secara kontekstual [15].

Implikasi praktisnya, pengembangan sistem pendukung keputusan sebaiknya tidak berhenti pada peningkatan kinerja prediksi. Sistem juga perlu menyediakan mekanisme untuk menjelaskan keputusan, mengevaluasi potensi ketimpangan antarkelompok, dan memberikan ruang bagi pengguna untuk melakukan pemeriksaan serta intervensi. Pendekatan tersebut lebih sejalan dengan prinsip AI yang transparan, dapat dipertanggungjawabkan, dan berpusat pada manusia.

V. KESIMPULAN

Penelitian ini menganalisis peran Explainable Machine Learning dalam mendukung pengambilan keputusan yang transparan dan berkeadilan melalui kajian konseptual dan sintesis literatur. Hasil sintesis menunjukkan bahwa penjelasan global dan lokal dapat membantu pengguna memahami perilaku model serta alasan prediksi pada tingkat individual. Informasi tersebut dapat digunakan untuk mendukung transparansi, audit, dan pemeriksaan terhadap keputusan berbasis machine learning.

Kajian juga menunjukkan bahwa explainability dan fairness merupakan dua dimensi yang saling melengkapi tetapi tidak identik. XAI dapat membantu menelusuri faktor yang berkontribusi terhadap prediksi dan mendukung identifikasi potensi sumber bias, sedangkan fairness tetap memerlukan evaluasi antarkelompok dengan metrik yang sesuai. Oleh karena itu, kerangka pengambilan keputusan yang transparan perlu mengintegrasikan XAI, evaluasi fairness, dan human oversight.

Secara keseluruhan, kontribusi penelitian ini adalah kerangka konseptual yang menempatkan XAI sebagai salah satu komponen dalam proses pengambilan keputusan yang dapat dijelaskan,

diaudit, dan ditinjau oleh manusia. Penelitian ini memiliki keterbatasan karena bersifat konseptual dan tidak melakukan eksperimen pada dataset tertentu sehingga belum menghasilkan nilai kinerja prediksi maupun nilai fairness. Penelitian selanjutnya perlu melakukan studi empiris pada domain tertentu dengan dataset nyata, membandingkan beberapa model dan metode XAI, mengukur fairness menggunakan metrik yang sesuai, serta melibatkan pengguna untuk mengevaluasi keterpahaman dan kegunaan penjelasan dalam proses pengambilan keputusan.

REFERENSI

- [1] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.
- [2] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
- [3] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 36–43, 2018.
- [4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, pp. 206–215, 2019.
- [5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, 2021.
- [6] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.
- [7] A. Rai, "Explainable AI: From black box to glass box," *J. Acad. Mark. Sci.*, vol. 48, pp. 137–141, 2020.
- [8] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv:1702.08608*, 2017.
- [9] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.
- [10] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Comput. Surv.*, vol. 55, no. 3, pp. 1–44, 2022.
- [11] S. Caton and C. Haas, "Fairness in machine learning: A survey," *AI Ethics*, vol. 4, pp. 1–31, 2024.
- [12] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

- [13] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 4765–4774.
- [14] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard J. Law Technol.*, vol. 31, no. 2, pp. 841–887, 2018.
- [15] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," in *Proc. Mach. Learn. Res.*, vol. 106, 2019, pp. 359–380.
- [16] T. Ahmad, J. Zhang, and X. Li, "Explainable artificial intelligence for decision support systems: A comprehensive survey," *Expert Syst. Appl.*, vol. 237, p. 121540, 2024.
- [17] P. Angelov and E. Soares, "Towards explainable deep neural networks (xDNN)," *Neural Netw.*, vol. 158, pp. 437–455, 2023.
- [18] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *Int. J. Human-Comput. Interact.*, vol. 36, no. 6, pp. 495–504, 2020.