Paaiškinamasis dirbtinis intelektas: apžvalga

Greta Juškaitė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Duomenų mokslo ir skaitmeninių technologijų institutas,
Akademijos g. 4, LT-08412 Vilnius, Lietuva
gretaju1314@gmail.com

Santrauka. Šiame straipsnyje nagrinėjamas paaiškinamasis dirbtinis intelektas, aptariama jo samprata, pagrindiniai principai ir taikymo sritys. Taip pat išskiriamos bei aprašomos galimos paaiškinamojo dirbtinio intelekto metodų kategorijos, analizuojama jų sandara. Straipsnio pabaigoje apžvelgiami aktualūs iššūkiai ir problemos, su kuriomis šiandien susiduria paaiškinamojo dirbtinio intelekto tyrėjai ir praktikai.

Raktiniai žodžiai: Paaiškinamasis dirbtinis intelektas, XAI, Explainable AI, XAI samprata, Dirbtinis intelektas, XAI kategorijos.

1 Įvadas

Sparčiai vystantis dirbtinio intelekto (DI) technologijoms, DI tapo pritaikomas ne tik specializuotose srityse, tokiose, kaip medicina, inžinerija ar moksliniai tyrimai, bet ir kasdieniame gyvenime. Didelis technologijų prieinamumas bei spartus tobulėjimas leidžia plačiajai visuomenei pasiekti DI sprendimus naudojantis išmaniaisiais telefonais ar įvairiomis internetinėmis platformomis. Dėl to vis daugiau žmonių pradėjo naudotis DI siekdami supaprastinti kasdienes užduotis, automatizuoti pasikartojančius procesus ar tiesiog efektyviau planuoti savo laiką. Vis dėlto, nepaisant, kad DI gali būti puikus pagalbininkas svarbu suprasti, kad ne visada jo pateikiama informacija gali būti teisinga ar patikima. Nors DI sistemos geba apdoroti milžiniškus kiekius duomenų ir pateikti atsakymus per labai trumpą laiką, jų veikimas priklauso nuo turimų duomenų kokybės, atnaujinimo dažnumo bei pačių algoritmų tikslumo. Kartais DI gali pateikti netikslius faktus, pasenusią informaciją arba suformuluoti atsakymus, kurie skamba įtikinamai, tačiau neturi realaus pagrindo. Iš to kyla problema – kaip atskirti, kada DI pateikta informacija yra patikima, o kada ja reikėtų suabejoti? Kaip vartotojui įvertinti DI atsakymo pagrįstumą, jei pats sprendimo priėmimo procesas nėra tiesiogiai matomas?

DOI: https://doi.org/10.15388/LMITT.2026.12

Čia galima pasitelkti paaiškinamąjį DI (angl. *Explainable Artificial Intelligence*, XAI), kuris ir bando spręsti šią problemą – paaiškinti naudotojui, kodėl DI nusprendė pasiūlyti būtent tokį atsakymą, kokį gavo, arba paaiškinti, kaip modelis „galvoja“. Kitaip tariant paaiškinamasis DI yra DI metodų visuma, kuri gali paaiškinti savo loginį pagrindą žmogui, apibūdinti jų stipriąsias ir silpnąsias puses bei perteikti supratimą apie tai, kaip jie elgsis ateityje [1].

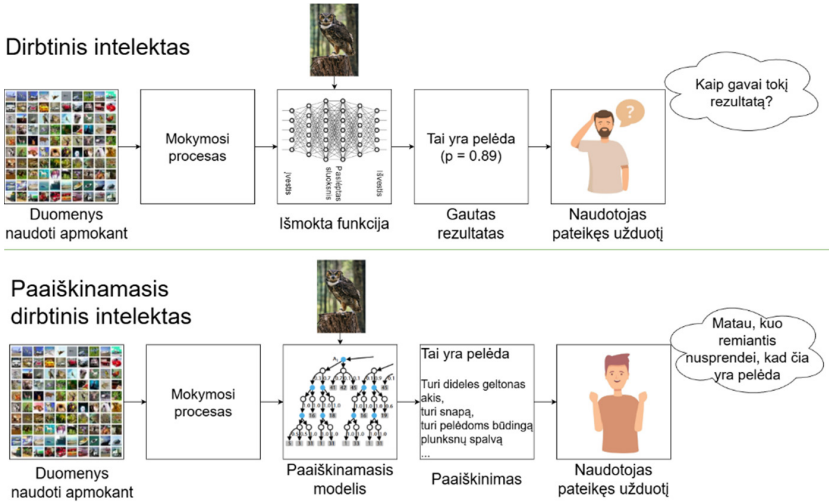
Tad šio straipsnio tikslas – išanalizuoti paaiškinamojo DI sampratą, pagrindinius principus, taikymo sritis bei išryškinti iššūkius, su kuriais susiduriama. Siekiant įgyvendinti numatytą tikslą, keliami šie uždaviniai:

1. Apibrėžti paaiškinamojo dirbtinio intelekto sąvoką ir pagrindinius principus.
2. Aptarti pagrindines paaiškinamojo DI taikymo sritis.
3. Išanalizuoti paaiškinamojo DI metodų kategorizavimą.
4. Išskirti pagrindinius iššūkius ir problemas, su kuriomis susiduria paaiškinamasis DI.

2 Paaiškinamojo dirbtinio intelekto samprata

Kaip jau buvo minėta, paaiškinamasis DI – tai DI metodai ir modeliai, kurie geba žmogui suprantamu būdu paaiškinti savo sprendimų ar pateiktų rezultatų loginį pagrindą [1]. Tai reiškia, kad vartotojas gali ne tik matyti galutinį rezultatą, bet ir suprasti, kodėl sistema pasirinko būtent tokį atsakymą. Toks skaidrumas leidžia lengviau įvertinti sprendimo patikimumą, pastebėti galimas klaidas ar šališkumą bei priimti labiau pagrįstus sprendimus. Paaiškinamojo DI koncepcija pateikiama 1 pav.

Kaip galime matyti 1 pav. yra vaizduojama paaiškinamojo DI sąvoka. Viršutinėje paveikslo dalyje pavaizduotas tradicinio DI veikimo principas – modelis apmokytas naudojant tam tikrą duomenų aibę. Gavęs ir išanalizavęs įvesties duomenis, pateikia galutinį rezultatą (pvz., nustato, kad nuotraukoje matomas paukštis yra pelėda), tačiau neatskleidžia, kokiais požymiais rėmėsi priimdamas šį sprendimą. Apatinėje dalyje parodyta, kaip paaiškinamasis DI praplečia šį procesą. Paaiškinamasis DI papildomai pateikia aiškius argumentus ir vizualines ar tekstines detales (pavyzdžiui, išryškina dideles akis, specifinį snapo tipą ar plunksnų raštą), kurios lėmė sprendimą, kad pavaizduotas paukštis yra pelėda. Taip paaiškinamasis DI padidina sprendimo skaidrumą, patikimumą ir leidžia vartotojui geriau suprasti modelio logiką.



1 pav. Paaiškinamojo DI koncepcija.

Paaiškinamojo DI algoritmai remiasi trimis pagrindiniais principais: paaiškinamumu (angl. *explainability*), interpretuojamumu (angl. *interpretability*) ir skaidrumu (angl. *transparency*). Modelis laikomas skaidriu, kai jis aiškiai parodo, kaip, remdamasis mokymo duomenimis, pasiekia tam tikrus rezultatus ir kaip, analizuodamas testavimo duomenis, sukuria atitinkamas žymas ar prognozes. Paaiškinamumas reiškia galimybę paaiškinti gautus rezultatus taip, kad juos suprastų kiti – ne tik sistemos kūrėjai, bet ir galutiniai vartotojai ar sprendimų priėmėjai [2]. O interpretuojamumas gali būti suprantamas, kaip algoritmo savybė ar bruožas, gebantis paaiškinti ar pateikti pakankamai išraiškingų duomenų žmogui suprantama kalba [13]. Tad jei algoritmas buvo parašytas laikantis šių principų, jį galima vadinti paaiškinamuoju. Tokiu atveju jis ne tik pateikia prognozę ar sprendimą, bet ir paaiškina, kokie veiksniai turėjo didžiausią įtaką rezultatui, leidžia įvertinti sprendimo pagrįstumą bei nustatyti galimus šališkumo ar klaidų šaltinius. Tokia savybė ypač svarbi realiose situacijose, kur sprendimai daro tiesioginę įtaką žmonėms ir aplinkai.

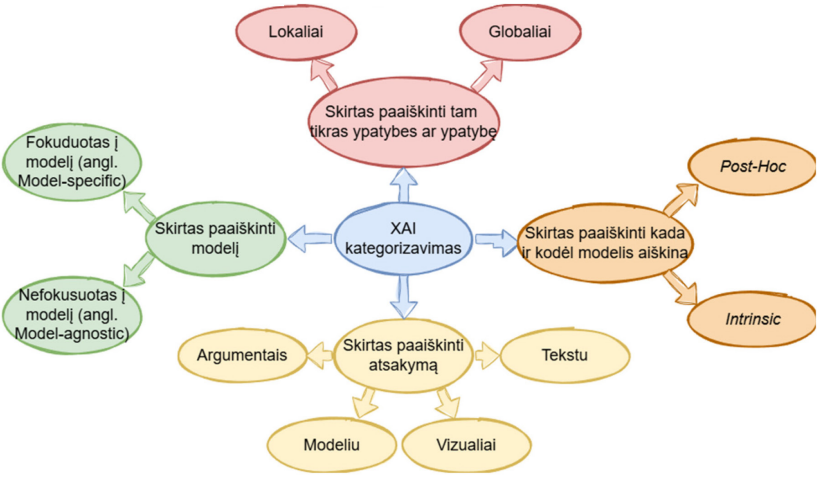
Paaiškinamasis DI gali būti taikomas įvairiose srityse. Pavyzdžiui, sveikatos priežiūroje – DI sistemose, analizuojančiose medicininius vaizdus ar prognozuojančiose ligų riziką. Paaiškinamasis DI gali paaiškinti, kokie klinikiniai rodikliai ar vaizdo sritys nulėmė sprendimą, taip padedant gydytojams priimti pagrįstus sprendimus [18]. Taip pat finansų sektoriuje paaiškinamasis

DI gali būti naudojamas kredito rizikos vertinimo, bankroto prognozavimo, sukčiavimo aptikimo modeliuose. Tai ypač aktualu bankuose, kur svarbu užtikrinti skaidrumą ir atitiktį reguliavimo reikalavimams [19]. Be finansų sektoriaus, būtų galima paminėti ir gamybos sektorių. Gamybos sektoriuje paaiškinamasis DI gali būti taikomas mašinos ir žmogaus bendradarbiavimo procesuose, siekiant užtikrinti sklandų, saugų ir efektyvų sprendimų priėmimą realiuoju laiku [5]. Saugumo industrijoje paaiškinamasis DI taip pat gali būti pritaikytas tokioms užduotims, kaip vizualinė žmogaus kūno analizė, transporto priemonės analizė ar elgesio analizė, užtikrinant didesnį sprendimų patikimumą ir skaidrumą [2].

Apibendrinant, paaiškinamasis DI įgauna vis didesnę reikšmę kaip esminis pažangių sprendimų priėmimo sistemų komponentas įvairiose taikymo srityse. Paaiškinamasis DI ne tik didina sprendimų skaidrumą ir patikimumą, bet ir užtikrina atskaitomybę.

3 Paaiškinamojo dirbtinio intelekto kategorijos

Remiantis straipsniais [3], [4] galima pastebėti, jog šiai dienai paaiškinamojo DI metodų sukurta yra ne vienas ir juos galima skirstyti į daugybę kategorijų. Vis dėlto, nors metodų įvairovė yra didelė ir jų klasifikavimo būdai gali skirtis priklausomai nuo pasirinktų kriterijų, bendrais bruožais juos būtų galima suskirstyti į kelias pagrindines grupes, kaip tai yra pateikiama 2 pav.



2 pav. Paaiškinamojo DI kategorizavimas.

Paaiškinamojo DI metodai 3 pav. yra suskirstyti į keturias kategorijas tokias, kaip: metodus, skirtus paaiškinti tam tikras ypatybes ar vieną iš ypatybių; skirtus paaiškinti kada ir kodėl modelis aiškina; skirtus suformatuoti atsakymus ir skirtus paaiškinti modelį. Taip pat šis kategorizavimas nėra baigtinis, vis dar yra diskutuojama, kaip turėtų atrodyti tikrasis kategorijų sąrašas ir įvairiuose šaltiniuose [3], [4], [5], [6], [19] šis sąrašas atrodo vis kitaip, tačiau šiame straipsnyje bus naudojamas būtent toks paaiškinamojo DI kategorizavimas.

Metodai, skirti paaiškinti tam tikras ypatybes ar vieną iš ypatybių, gali būti skirstomi į lokalų paaiškinamumą arba globalų. Lokalus paaiškinamumas apibūdina sistemos gebėjimą parodyti vartotojui, kodėl buvo priimtas konkretus sprendimas ar pasirinkimas. Populiarūs tokio tipo paaiškinimo metodai yra LIME [7], SHAP [8] bei kontrafaktinės analizės metodai [9]. Lokalaus paaiškinamumo metodai laikomi pirmuoju svarbiausiu modelio skaidrumo komponentu [10]. Tuo tarpu globalus paaiškinamumas reiškia viso mokymosi algoritmo veikimo paaiškinimą, atsižvelgiant į panaudotus mokymo duomenis, algoritmų tinkamą taikymą bei galimas klaidas ar netinkamas panaudojimo situacijas. Pavyzdžiui, globalus požymių atvaizdavimas (angl. *Global Attribution Mapping,* GAM) [11] yra globalus paaiškinamumo metodas, skirtas analizuoti ir paaiškinti neuroninių tinklų prognozių pasiskirstymą skirtingose populiacijos subgrupėse.

Modeliai, skirti paaiškinti kada ir kodėl modelis aiškina, gali būti išskirti į ribojantį DI modelio sudėtingumą (angl. *intrinsic*) arba analizuojantį modelio veikimą po apmokymo (angl. *post-hoc*) [10]. *Intrinsic* paaiškinamasis DI metodas pateikia žmogui suprantamus paaiškinimus kartu su originaliais rezultatais [12]. Kai kurie mašininio mokymosi (angl. *machine learning*, ML) modeliai, tokie kaip sprendimų medžiai (angl. *Decision Trees*) ir retų tiesinių struktūrų modeliai (angl*. Sparse Linear Models*), laikomi *Intrinsic* tipo paaiškinamaisiais DI metodais, nes jie yra savaime interpretuojami [10]. Tuo tarpu *post-hoc* paaiškinimai taikomi po to, kai modeliai jau buvo apmokyti ir sprendimai priimti. Populiarūs *post-hoc* metodai yra lokaliai interpretuojami nuo modelio nepriklausomi paaiškinimai (angl. *Local Interpretable Model-agnostic Explanations, LIME*) [7] ir permutacijos svarba (angl. *Permutation Importance*) [14].

Metodai, skirti paaiškinti modelį, galėtų būti subkategorizuoti į fokusuotus į modelį (angl. *Model-specific*) ir nefokusuotus į modelį (angl. *Model-agnostic*). Fokusuoti į modelį paaiškinimo įrankiai yra skirti konkrečiam mo-

deliui arba modelių grupei. Pavyzdžiui, GNN paaiškinimo įrankis (angl. *Graph Neural Network Explainer*) [15] yra metodas, leidžiantis pateikti suprantamus paaiškinimus bet kurio GNN pagrindu veikiančio modelio prognozėms bet kurioje grafais paremtoje mašininio mokymosi užduotyje. Priešingai, į modelį nefokusuoti paaiškinimo įrankiai teoriškai gali būti įgyvendinami su bet kokiu ML modeliu. Be to, nefokusuoti į modelį paaiškinimo metodai dažniausiai veikia analizuodami įvesties ir išvesties požymius ir pagal apibrėžimą neturi prieigos prie modelio vidinės informacijos, tokios kaip svoriai ar struktūrinė informacija [10].

Metodai, skirti paaiškinti atsakymą, taip pat yra labai svarbūs ir gali būti skirstomi į: skirtus paaiškinti tekstu; skirtus paaiškinti vizualiai; skirtus paaiškinti, kaip tam tikrą modelį ar skirtus paaiškinti pateikiant argumentus. Pavyzdžiui, teksto pagrindu veikiantys paaiškinimo metodai yra plačiai naudojami natūralios kalbos apdorojimo (angl. *Natural Language Processing*, NLP) srityje, siekiant gauti detalią informaciją ir sukurti žmogui suprantamus paaiškinimus [16]. Kita vertus, vizualizuoti paaiškinimo metodai taikomi platesnėse srityse, įskaitant NLP, neuroninius tinklus ir sveikatos priežiūrą. Argumentais pagrįsti paaiškinimai apima požymių pateikimą tokiu būdu, kokiu žmonės priima sprendimus, siekiant geriau suprasti požymio svarbą [17]. Modelio pagrindu grindžiami paaiškinimo metodai turi apibūdinti „juodosios dėžės“ modelio vidinę veikimo logiką. Tai dažnai pasiekiama „juodosios dėžės“ modelio elgesį apytikriai atkartojant kitu modeliu, kuris yra lengviau suprantamas ir skaidresnis [10].

4 Iššūkiai, su kuriais susiduria paaiškinamasis dirbtinis intelektas

Nors paaiškinamasis DI gali būti pritaikomas įvairiose srityse ir jo reikšmė nuolat auga, praktikoje vis dar išlieka reikšmingų metodologinių, vertinimo ir taikymo spragų ypač tada, kai DI sistemos tampa vis didesnės, multimodalinės ir autonomiškesnės.

Pagrindiniai iššūkiai, su kuriais susiduriama naudojant paaiškinamąjį DI:

1. Sudėtingas paaiškinamųjų DI metodų įtraukimas į esamas DI sistemas, nes daugelis metodų veikia kaip išoriniai priedai, o ne integruota modelio dalis. Toks požiūris gali sukelti efektyvumo nuostolių, papildomų klaidų ar šališkumo, o taip pat reikalauja daug techninių išteklių, ypač sistemoms, kurios iš pradžių nebuvo kuriamos interpretuojamumui užtikrinti [23].

2. Multimodaliniai modeliai integruoja skirtingų tipų duomenis – tekstą, vaizdą, garsą ar struktūrizuotą informaciją, o jų sprendimai dažnai kyla iš kryžminės tarpmodalinės sąveikos. Dabartiniai paaiškinamieji DI metodai dažniausiai yra sukurti vienam duomenų tipui ir sunkiai perkeliamai apibendrinami į multimodalinį kontekstą. Be to, trūksta efektyvių paaiškinimo metodų, kurie galėtų vienu metu veikti su multimodaliniais ir daugiakalbiais duomenimis. Todėl išlieka esminis klausimas: kaip formaliai modeliuoti ir paaiškinti skirtingų modalumų bei kalbų indėlį į galutinę išvadą, išvengiant perteklinio supaprastinimo ir užtikrinant paaiškinimų tikslumą bei patikimumą [10], [21].
3. Dabartiniai paaiškinamieji DI metodai daugiausia remiasi koreliacinėmis struktūromis ir neatskleidžia tikrų priežastinių mechanizmų, o tai ypač aktualu aukštos rizikos srityse, tokiose kaip medicina ar finansai, kur sprendimų paaiškinimai turi turėti priežastinį pagrindimą. Pereiti nuo statistinės asociacijos prie priežastinio aiškinimo reikalauja integruoti struktūrinius priežastinius modelius, intervencinius testavimus ir priešingus paaiškinimus (angl. *Counterfactual Explanations*), tačiau tai išlieka metodologiškai sudėtinga, ypač didelio masto neuroninių tinklų kontekste [20]. Be to, trūksta standartizuotų paaiškinamųjų DI vertinimo metrikų, todėl sunku objektyviai palyginti metodų efektyvumą tarp skirtingų tyrimų ir taikymo sričių, kas dar labiau apsunkina patikimų ir saugių paaiškinimų diegimą praktikoje [4], [21], [22].
4. Teisiniai dokumentai, tokie kaip Europos Sąjungos dirbtinio intelekto aktas, įtvirtina skaidrumo ir atskaitomybės reikalavimus aukštos rizikos DI sistemoms. Tačiau „paaiškinamumo“ sąvoka teisiniame diskurse nebūtinai sutampa su techniniu jos apibrėžimu. Tai apsunkina tarpdisciplininį bendradarbiavimą: techniniai sprendimai turi būti suderinti su normatyviniais lūkesčiais, kartu išlaikant komercinį gyvybingumą ir intelektinės nuosavybės apsaugą [21].

Ir tai yra tik kelios esminės problemos, kurias būtina spręsti siekiant tobulinti paaiškinamąjį DI.

5 Išvados

Apžvelgus paaiškinamojo DI sampratą, aišku, kad paaiškinamasis DI yra metodų visuma, kuri paaiškina kaip ir kodėl DI priima vienokį ar kitokį sprendimą. Paaiškinamasis DI remiasi trimis pagrindiniais principais – paaiškina-

mumu, interpretuojamumu ir skaidrumu, kurie leidžia vartotojams vertinti modelių logiką ir patikimumą. Paaiškinamasis DI gali būti skirstomas į keturias kategorijas – metodus skirtus paaiškinti tam tikras ypatybes ar vieną iš ypatybių, modelius skirtus paaiškinti kada ir kodėl modelis aiškina, metodus skirtus paaiškinti modelį bei metodus skirtus suformatuoti atsakymą. O šios kategorijos gali būti suskirstytos į dar siauresnes grupes. Taip pat buvo apžvelgti keli, iš daugumos, paaiškinamojo DI iššūkiai, kurie leido suprasti, kad vis dar išlieka metodologinės, vertinimo ir taikymo spragos, ypač didelėse DI sistemose, kad vis dar trūksta standartizuotų vertinimo metrikų bei patikimų metodų, leidžiančių objektyviai vertinti paaiškinimų kokybę ir jų praktinę vertę. Didelio masto kalbinėse ir multimodalinėse sistemose tradiciniai *post-hoc* metodai dažnai nebeatspindi realių vidinių sprendimų priėmimo mechanizmų, o paaiškinamojo DI sprendimai dažniausiai grindžiami koreliacijomis, o ne priežastiniais ryšiais. Papildomų iššūkių kelia skirtingų modalumų indėlio paaiškinimas bei teisiniai reikalavimai, tokie kaip Europos Sąjungos dirbtinio intelekto aktas, atskleidžiantys atotrūkį tarp techninės ir teisinės paaiškinamumo sampratos. Todėl ateityje būtina plėtoti standartizuotas vertinimo metrikas, stiprinti priežastinį paaiškinamumą ir geriau derinti techninius paaiškinamuosius DI sprendimus su praktiniais ir normatyviniais reikalavimais.

Siekiant tobulinti atliekamą tyrimą, būtų galima išplėsti nagrinėjamų paaiškinamojo DI metodų spektrą, įtraukiant naujausius taikomus sprendimus bei atlikti eksperimentinį palyginimą skirtinguose taikymo scenarijuose. Taip pat būtų naudinga giliau analizuoti paaiškinimų kokybės vertinimo metodikas, siekiant nustatyti objektyvius ir standartizuotus kriterijus, leidžiančius įvertinti paaiškinimų naudingumą galutiniams vartotojams. Be to, tyrimą būtų galima papildyti praktiniu taikymu, įgyvendinant pasirinktus paaiškinamojo DI metodus realioje sistemoje ir įvertinant jų poveikį sprendimų priėmimo procesui. Ateities darbuose taip pat tikslinga daugiau dėmesio skirti priežastinio paaiškinamumo integravimui bei multimodalinių modelių analizės metodų kūrimui, siekiant geriau suprasti sudėtingų DI sistemų veikimą.

Literatūra

[1] Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI magazine*, *40*(2), 44-58.

[2] Abdullah Al Nasim, M. D., Anas Ferdous, A. S., Rashid, A., Tuj Johura Soshi, F., Biswas, P., Biswas, A., & Datta Gupta, K. (2024). Trustworthy XAI and Application. *arXiv e-prints*, arXiv-2410.

[3] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82-115.
[4] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.
[5] Abhilash, P. M., Luo, X., Liu, Q., Madarkar, R., & Walker, C. (2024). Towards next-gen smart manufacturing systems: the explainability revolution. *Npj Advanced Manufacturing*, *1*(1), 8.
[6] Nassar, M., Salah, K., Ur Rehman, M. H., & Svetinovic, D. (2020). Blockchain for explainable and trustworthy artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(1), e1340.
[7] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). „ Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
[8] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.
[9] Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, *31*, 841.
[10] Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. *IEEe Access*, *10*, 93104-93139.
[11] Ibrahim, M., Louie, M., Modarres, C., & Paisley, J. (2019, January). Global explanations of neural networks: Mapping the landscape of predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 279-287).
[12] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.
[13] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18.
[14] Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340-1347.
[15] Ying, Z., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, *32*.
[16] Liu, H., Yin, Q., & Wang, W. Y. (2019, July). Towards explainable NLP: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5570-5581).
[17] Amgoud, L., & Prade, H. (2009). Using arguments for making and explaining decisions. *Artificial Intelligence*, *173*(3-4), 413-436.
[18] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721-1730).
[19] Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. *Artificial Intelligence Review*, *57*(8), 216.
[20] Chinnaraju, A. (2025). Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability. *World Journal of Advanced Engineering Technology and Sciences*, *14*(3), 170-207.

[21] Abbas, Q., Jeong, W., & Lee, S. W. (2025, August). Explainable AI in clinical decision support systems: a meta-analysis of methods, applications, and usability challenges. In *Healthcare* (Vol. 13, No. 17, p. 2154). MDPI.
[22] Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., ... & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, *15*(2), 1-38.
[23] Butt, T., & Iqbal, M. (2024, December). Explainable ai: Applications, challenges, current solutions and future research directions. In *Proceedings of the 2024 the 12th International Conference on Information Technology (ICIT)* (pp. 108-113).