УДК 004.8:004.9 **DOI:** https://doi.org/10.30837/bi.2026.1(104).10

І. О. Лещинська
ХНУРЕ, м. Харків, Україна, iryna.leshchynska@nure.ua, ORCID iD: 0000-0002-8737-4595

УЗАГАЛЬНЕНА СИМВОЛЬНА МЕНТАЛЬНА МОДЕЛЬ РІШЕННЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ДЛЯ ЗОВНІШНІХ КОРИСТУВАЧІВ

Розглянуто проблему побудови ментальних моделей рішень інтелектуальних систем для зовнішніх користувачів, вирішення якої потребує побудови прозорих пояснень, адаптованих до предметної області та рівня підготовки користувача. Запропоновано узагальнену трирівневу символьну ментальну модель, що включає концептуальний рівень абстракцій рішень, рівень пояснень на основі каузальних правил і базу фактів, які об'єднані функцією вертикальної композиції, що забезпечує можливість керованої деталізації пояснень. Розроблено алгоритм автоматизованої побудови узагальненої моделі на основі доменної онтології, каузальних послідовностей міркування та даних користувачів, який забезпечує персоналізовані пояснення з урахуванням рівня підготовки цих користувачів. Модель забезпечує можливість формування пояснень для новачків, користувачів середнього рівня та експертів, що дає змогу узгоджувати глибину деталізації з когнітивними можливостями аудиторії та вимогами до верифікації рішень у високоризикових предметних областях, включаючи медичну діагностику, фінансові системи підтримки рішень та інші критичні застосування, де необхідні як інтерпретованість, так і відтворюваність ланцюжків міркувань.

УЗАГАЛЬНЕНА СИМВОЛЬНА МЕНТАЛЬНА МОДЕЛЬ, ПОЯСНЮВАНИЙ ШТУЧНИЙ ІНТЕЛЕКТ, ДОМЕННІ ОНТОЛОГІЇ, КАУЗАЛЬНІ ПРАВИЛА, БАЗА ФАКТІВ, ПЕРСОНАЛІЗОВАНІ ПОЯСНЕННЯ

I. O. Leshchynska. Generalised symbolic mental model of an intelligent system decision for external users. The paper addresses the problem of constructing mental models of intelligent system decisions for external users, which requires transparent explanations adapted to the application domain and the user's expertise level. A generalised three-layer symbolic mental model is proposed, comprising a conceptual layer of decision abstractions, an explanation layer based on causal rules, and a fact base, which are linked by a vertical composition function that enables controlled explanation granularity. An algorithm for automated construction of the generalised model is developed using a domain ontology, causal reasoning paths, and user data, providing personalised explanations tailored to users' expertise levels. The model supports explanation modes for novices, intermediate users, and experts, aligning the depth of detail with cognitive capabilities and verification requirements in high-risk domains where trust in intelligent systems is essential.

GENERALIZED SYMBOLIC MENTAL MODEL, EXPLAINABLE ARTIFICIAL INTELLIGENCE, DOMAIN ONTOLOGIES, CAUSAL RULES, FACT BASE, PERSONALIZED EXPLANATIONS

Вступ

Сучасні інтелектуальні системи використовують складні непрозорі моделі формування результатів, що потребують розробки підсистем пояснень для підвищення довіри зовнішніх користувачів до отриманих рішень. Побудова підсистем пояснень виконується в рамках наукового напрямку пояснювального штучного інтелекту. Розробка систем пояснювального штучного інтелекту потребує побудови формального опису сприйняття рішень інтелектуальних систем користувачами. Вирішення даної проблеми пов'язано із створенням символьних структур для побудови ментальних моделей рішень інтелектуальних систем. Такі структури мають забезпечувати адаптацію ментальної моделі до індивідуальних когнітивних особливостей зовнішніх користувачів-непрофесіоналів. Зовнішні користувачі – це особи, які не є розробниками системи та не володіють глибокими технічними знаннями про механізми роботи інтелектуальної системи, наприклад пацієнти, чиї дані використовуються в медичних діагностичних системах, клієнти банків у фінансових системах підтримки рішень, студенти, що використовують в освітні рекомендаційні системи.

Традиційні неперсоналізовані підходи до побудови пояснень не враховують індивідуальні ментальні моделі користувачів та рівень їхньої експертизи, що утруднює розуміння пояснень зовнішніми користувачами, особливо коли ці користувачі мають різний фаховий бекграунд та когнітивні здібності. Також побудова ментальних моделей виконується із залученням експертів на основі трудомістких структурованих інтерв'ю. Наслідком використання трудомістких інтерв'ю при побудові опису сприйняття рішень системи людиною стає формування специфічних для предметної області ментальних моделей, які не можуть бути повторно використані в інших доменах через відсутність уніфікованої структури представлення знань. Побудова ментальних моделей експертами через структуровані інтерв'ю та картування концептів потребує значних витрат часу та характеризується низькою узгодженістю між різними експертами внаслідок суб'єктивної інтерпретації когнітивних структур користувачів.

Для побудови інтегрального опису ментальних моделей потрібно сформувати символьну архітектуру, яка явно виділяє концептуальний шар опису абстракцій рішень, пояснювальний шар каузальних правил виведення та базу знань доменних фактів, інтегруючи їх через онтологічний опис.

Внаслідок відсутності інтегрального опису ментальних моделей і відповідної відсутності персоналізованих пояснень знижується довіра до рішень інтелектуальних систем і можуть виникати обмеження при практичному застосуванні таких систем у високоризикових предметних областях.

Відповідно, проблема розробки узагальненого символьного опису ментальних моделей рішень інтелектуальної системи для персоналізації сприйняття пояснень є актуальною.

Серед існуючих підходів до побудови пояснень і до побудови ментальних моделей в рамках пояснень доцільно виділити три напрямки досліджень: пояснення на основі важливості ознак; нейро-символьний підхід; пояснення з використанням ментальних моделей користувачів інтелектуальних систем.

Перший напрямок охоплює методи пояснювального штучного інтелекту, які розраховують внесок окремих ознак у рішення моделі машинного навчання, яка використовується в інтелектуальній системі. В рамках даного напрямку в роботі [1] розроблено метод SHAP, що використовує значення Шеплі з кооперативної теорії ігор для обчислення справедливого розподілу внеску кожної ознаки в рішення моделі через принцип адитивності вкладів. В дослідженні [2] запропоновано метод LIME, який формує пояснення через локальні лінійні апроксимації в околиці конкретного рішення складних моделей, зокрема глибоких нейронних мереж. В [3] проведено порівняльний аналіз методів SHAP та LIME у контексті біомедичних застосувань та показано їхню залежність від типу моделі машинного навчання та характеристик даних, що обмежує узгодженість пояснень.

У монографії [4] систематизовано підходи до забезпечення інтерпретованості машинного навчання, включаючи методи для ранжування ознак за їх внеском в модель на всьому наборі даних та локальні методи для генерації контрфактичних сценаріїв, які відображають мінімальні зміни вхідних даних, необхідні для зміни рішення моделі. У огляді [5] класифіковано методи забезпечення глобальної та локальної пояснюваності. В [6] проведено аналіз обмежень методів забезпечення пояснюваності при клінічних застосуваннях. Дослідження показало суттєвий розрив між технічними метриками для оцінки пояснюваності в системах машинного навчання та реальним розумінням клініцистами механізмів роботи моделі.

У підсумку ці методи пояснювального штучного інтелекту орієнтовані на побудову неперсоналізованих пояснень без урахування індивідуальних ментальних моделей користувачів з різним рівнем експертизи.

Другий напрямок використовує онтології на нейросимволічному підході для представлення знань у предметних областях та їх інтеграції з методами машинного навчання. В [7] дослідили роль онтологій та показали, що онтології забезпечують структуровані фреймворки знань для забезпечення інтерпретованості процесів прийняття рішень в інтелектуальних системах. В дослідженні [8] проаналізовано множинні ролі онтологій у пояснювальному штучному інтелекті, в тому числі для формалізації доменних концептів, міркування для заповнення прогалин у знаннях з тим, щоб забезпечити розробку людиноцентричних пояснювальних систем.

В роботі [9] проведено систематичний огляд нейро-символьного штучного інтелекту в охороні здоров'я на основі аналізу біомедичних досліджень та показано потенціал інтеграції символьного міркування з глибоким навчанням для підвищення інтерпретовності рішень в клінічних застосуваннях, де зрозумілість міркування є критично важливою для прийняття рішення клініцистами. В [10] розроблено нейро-символьну систему, що об'єднує великі мовні моделі з експертною системою на основі правил для витягування структурованих даних із текстових радіологічних звітів, що забезпечує аудит ланцюжків виведення від кінцевих міток до початкових токенів моделі через отриману явну трасу міркування.

В дослідженні [11] розробили онтологію людських пояснень щодо поведінки робітників, що дає можливість роботам генерувати контекстно-релевантні пояснення своїх дій для людей – спостерігачів. В [12] запропонували онтологічний фреймворк, що інтегрує машинне навчання та імовірнісне планування в межах міркування на основі здорового глузду, де онтології використовуються для формалізації контексту активності користувача для підтримки рішень щодо допоміжних утручань.

Ключове обмеження даного підходу полягає у використанні специфічних онтологій для окремих галузей застосування без розробки узагальненої незалежної від предметної області схеми, що могла б бути повторно використана для нових предметних областей без повного переписування онтологічної структури.

Третій напрямок орієнтований на моделювання взаємодії людини з інтелектуальними системами з використанням ментальних моделей. Останні описують, як користувачі формують внутрішні представлення процесу роботи та рішень інтелектуальних систем на основі отриманих пояснень. В [13] дослідили ментальні моделі в рамках пояснювального штучного інтелекту, включивши призначення пояснення, аудиторію користувачів та мову представлення, з акцентом на зрозумілість пояснення. Тобто, дослідили ментальні моделі як основу людиноцентричного підходу до пояснень. В дослідженні [14] проведено огляд літератури з соціальних наук щодо процесу формування пояснень людьми в повсякденній комунікації та виявлено істотний розрив між методами

пояснювального штучного інтелекту, що фокусуються на повноті представленої інформації, та когнітивними теоріями людських пояснень, що підкреслюють селективність та контекстуальність представлення інформації.

В роботі [15] досліджували ефект прозорості штучного інтелекту на формування ментальних моделей користувачів та виявили, що відображення рівнів впевненості моделі в своїх передбаченнях зменшує надмірну довіру користувачів у медичних діагностичних системах. В роботі [16] провели дослідження щодо впливу інформації про відмови системи на рівень довгострокової довіри користувачів. Користувачі, що отримували чіткі пояснення причин відмов, демонстрували вищі показники довіри через кілька місяців порівняно з користувачами без досвіду відмов.

В [17] досліджено поведінкові метрики довіри в системах штучного інтелекту, виявивши ключові індикатори, зокрема частоту прийняття рекомендацій системи, час між отриманням рекомендації та дією користувача, час освоєння нового функціоналу та швидкість відновлення довіри після помилок системи. В [18] провели кількісне дослідження впливу інтерактивних та контекстуальних пояснень на довіру користувачів, ра продемонстрували, що залученість користувачів у процесу формування пояснень та розуміння контексту відіграють додаткову роль у формуванні довіри крім забезпечення прозорості механізмів роботи інтелектуальної системи.

Однак даний підхід має обмеження ,пов'язане із відсутністю формальних схем побудови символьних ментальних моделей із підтримкою механізмів деталізації від абстрактних концептів до конкретних фактів предметної області з використанням формалізованих правил логічного виведення.

Таким чином, існуючі підходи до побудови ментальних моделей рішень інтелектуальних систем для зовнішніх користувачів мають ряд обмежень щодо формування символьного представлення таких моделей, які перешкоджають побудові персоналізованих пояснень для користувачів-непрофесіоналів, оскільки останні потребують інтерпретованих пояснень рішень у зрозумілій формі без деталізації механізмів роботи алгоритмів машинного навчання. Зазначений недолік обґрунтовує необхідність розробки узагальненої символьної ментальної моделі рішення інтелектуальної системи для зовнішніх користувачів.

1. Постановка задачі

Метою дослідження є розробка узагальненої символьної ментальної моделі рішення інтелектуальної системи для зовнішніх користувачів з явним розділенням концептуального шару, пояснювального шару та фундаментальної бази знань для скорочення витрат часу на побудову ментальних моделей і можливості повторного використання для інших предметних областей за рахунок незалежної від домену схеми онтологічного представлення.

Для досягнення мети вирішуються такі задачі:

– розробити трирівневу символьну ментальну модель рішення інтелектуальної системи для зовнішніх користувачів із концептуальним шаром для представлення абстракцій рішень через класи онтологій, пояснювальним шаром для представлення каузальних зв'язків між властивостями об'єктів онтологій, базою знань для представлення фактів предметної області;

– розробити алгоритм автоматизованої побудови трирівневої символьної ментальної моделі рішення інтелектуальної системи.

2. Трирівнева символьна ментальна модель рішення інтелектуальної системи для зовнішніх користувачів

Розроблена модель використовує чотири типи компонентів: класи онтологій, властивості об'єктів, властивості даних та індивіди онтологій.

Трирівнева архітектура моделі має вигляд:

$$M = (L_C, L_E, L_K, \Phi), \quad (1)$$

де L_C – концептуальний шар високорівневих абстракцій рішень, L_E – пояснювальний шар каузальних правил виведення, L_K – база знань фактів предметної області, а Φ – функція вертикальної композиції між рівнями для підтримки деталізації пояснень.

Концептуальний шар містить множину класів онтологій:

$$L_C = \{C_1, C_2, ..., C_n\}, \quad (2)$$

де кожен клас C представляє категорію рішень інтелектуальної системи з формальним визначенням необхідних та достатніх умов засобами дескриптивної логіки, що дає можливість відносити індивідів до класів на основі їхніх властивостей.

Пояснювальний шар містить множину правил виведення:

$$L_E = \{r_1, r_2, ..., r_m\}, \quad (3)$$

де кожне правило r_j представлено традиційною імплікацією:

$$r_j: \text{антецедент}_j \Rightarrow \text{консеквент}_j, \quad (4)$$

де антецедент містить кон'юнкцію атомів із порівняннями властивостей даних з пороговими значеннями з використанням булевих операторів, а консеквент містить або твердження про належність індивіда до класу концептуального шару або ж додаткові властивості об'єкта для формування каузального шляху міркування.

База знань містить множину екземплярів онтологій

$$L_K = \{i_1, i_2, ..., i_k\}, \quad (5)$$

де кожен екземпляр i представляє конкретного користувача системи з типізованими властивостями даних,

що відображають значення атрибутів користувача, які відповідають рішенню інтелектуальної системи.

Функція вертикальної композиції Ф відображає кожен концепт рішення з концептуального шару на множину пар із каузальних правил виведення пояснювального шару та фактів фундаментальної бази знань:

$$\Phi: L_C \to 2^{L_E \times L_K}. \quad (6)$$

Вирази (1) – (6) обґрунтовують належність індивіда до класу концепту, забезпечуючи механізм деталізації від абстрактних концептів через каузальні правила до конкретних значень атрибутів користувача, що забезпечує можливість адаптації з урахуванням рівня експертизи останнього.

У сукупності вертикальна композиція рівнів розробленої моделі забезпечує формування повних ланцюжків обґрунтування від високорівневих концептів рішень через каузальні правила до конкретних фактів про атрибути користувачів через функцію обґрунтування та механізм деталізації з адаптацією до рівня експертизи користувача.

Обчислення функції Ф базується на аналізі ланцюжків виведення для онтологій, отриманих під час віднесення індивідів до концептуальних класів через застосування правил виведення до властивостей індивідів. Такий підхід забезпечує прозорість процесу прийняття рішення інтелектуальною системою.

Механізм деталізації дозволяє переходити від концептуального шару через пояснювальний шар до бази знань відповідно до когнітивних потреб користувача та рівня його підготовки і практичних задач, які він вирішує.

На концептуальному рівні користувач може бачити лише високорівневий концепт рішення без деталей обґрунтування, що мінімізує когнітивне навантаження для користувачів-новачків.

На пояснювальному рівні користувач може деталізувати концепт для перегляду каузальних правил виведення, які обґрунтовують рішення, що орієнтовано на потреби користувачів середнього рівня експертизи.

На рівні бази знань розробленої моделі користувач може деталізувати правило для перегляду конкретних значень атрибутів свого профілю, які задовольняють умові правила. Така можливість забезпечує повну прозорість для експертів, які потребують верифікації всіх деталей міркування у процесі формування рішення в інтелектуальній системі.

3. Алгоритм побудови узагальненої символьної ментальної моделі рішення інтелектуальної системи

Алгоритм автоматизованої побудови узагальненої символьної ментальної моделі рішення інтелектуальної системи виконується для кожної нової предметної області застосування моделі через формування доменних класів без зміни архітектури.

Вхідними даними алгоритму є:

– доменна онтологія із ієрархією класів та відношеннями для відповідної предметної області;

– набір рішень інтелектуальної системи з описами категорій рішень та критеріїв класифікації;

набір даних користувачів із значеннями атрибутів для заповнення бази знань;

– каузальні послідовності міркування від атрибутів до концептів рішень, отримані на основі експертних знань щодо предметної області.

На першому етапі алгоритму виконується витягування релевантних класів концептів рішень з доменної онтології. Використовується аналіз семантичної подібності між описами категорій рішень системи та класами доменної онтології. Для оцінки семантичної подібності використовують векторні представлення цих описів.

На другому етапі виконується формалізація каузальних шляхів шляхом їх опису з використанням правил виведення. Для кожного каузального шляху від атрибутів до концепту рішення формується правило виведення з антецедентом, що містить булеві умови на значення властивостей даних індивіда, та консеквентом, що містить твердження про належність індивіда до класу концептуального шару.

На третьому етапі виконується заповнення фундаментальної бази знань даними користувачів із типізованими властивостями даних, які відображають конкретні значення атрибутів, релевантних для рішення інтелектуальної системи.

На четвертому етапі виконується автоматизоване виведення через механізм логічного виведення онтологій для віднесення індивідів до класів концептів рішень на основі правил виведення та означень класів, що дає можливість сформувати повну ментальну модель із ланцюжками виведення від концептів рішень до конкретних значень атрибутів користувачів.

Результатом виконання алгоритму є повна узагальнена символьна ментальна модель рішення інтелектуальної системи для зовнішніх користувачів із класифікованими індивідами, ланцюжками обґрунтування та обчисленими каузальними шляхами міркування.

Дана модель може бути використана для генерації персоналізованих пояснень відповідно до рівня експертизи кожного конкретного користувача.

При імплементації моделі адаптація до рівня експертизи може бути реалізована шляхом конфігурації інтерфейсу користувача відповідно до метаданих із профілю користувача щодо рівня його експертності та практичних задач, що він вирішує.

Новачкам може бути автоматично встановлено спрощений режим із відображенням лише концептуального шару без можливості деталізації для зменшення когнітивного навантаження та запобігання інформаційному перевантаженню.

Користувачам середнього рівня має бути встановлено стандартний режим із можливістю деталізації до пояснювального шару для вивчення каузальних правил обґрунтування рішення та відповідних ланцюжків міркувань.

Експертам встановлюється експертний режим з повною можливістю деталізації від концептуального шару через правила й до бази знань для верифікації конкретних значень даних та повного аудиту ланцюжків міркування системи.

4. Експериментальна перевірка розробленої моделі

Експериментальна перевірка узагальненої символьної ментальної моделі здійснюється на наборі даних медичної діагностики діабету типу 2 шляхом порівняння зрозумілості пояснень із двома базовими підходами SHAP та LIME за метриками частотності термінів у веб-корпусі, семантичної узгодженості з медичною літературою та читабельності текстів пояснень, що забезпечує об'єктивну комплексну оцінку на відміну від суб'єктивних даних за результатами опитування користувачів. Набір даних із репозиторію UCI Machine Learning Repository, що містить записи пацієнток з атрибутами – рівень глюкози, індекс маси тіла, вік, артеріальний тиск , рівень інсуліну, спадкова функція діабету тощо. Для експерименту випадковим чином відібрано підмножину записів для того, щоб забезпечити збалансований розподіл класів у вхідному наборі даних. Для запропонованої моделі пояснення генеруються на основі ланцюжків виведення у середньому шарі. Такий підхід забезпечує побудову структурованих пояснень з явними каузальними зв'язками між атрибутами та рішеннями.

Для методів SHAP та LIME пояснення генеруються через стандартні бібліотеки shap та lime з параметрами за замовчуванням. Ці бібліотеки повертають числові коефіцієнти важливості кожної із ознак або оцінки її впливу на передбачення без явної каузальної інтерпретації. Зрозумілість пояснень оцінюється через три метрики, які використовують веб-пошук та лінгвістичний аналіз, що забезпечує об'єктивність перевірки і не потребує залучення людей - користувачів до опитувань.

Метрика TFWC визначає частотність термінів у веб-корпусі (Term Frequency in Web Corpus). Для розрахунку метрики для кожного текстового пояснення виявляються ключові терміни з використанням алгоритму RAKE (Rapid Automatic Keyword Extraction), що реалізований у бібліотеці rake-nltk. Даний алгоритм ідентифікує 3–5 найважливіших термінів у складі пояснення, наприклад, «high BMI», «diabetes risk».

В подальшому кожен ключовий термін подається у складі пошукового запиту до англомовних медичних сайтів. В запиті використовується параметр site:*.edu або site:diabetes.org або аналогічні. Для кожного терміну у запиті фіксується кількість результатів пошуку, яка відображає його поширеність в мережі інтернет.

Метрика TFWC обчислюється як середнє логарифмічне значення кількості результатів за всіма ключовими термінами пояснення. Високі значення TFWC свідчать про поширеність термінів пояснення в медичній веб-літературі, що свідчить про зрозумілість вказаних слів для зовнішніх користувачів внаслідок знайомства з термінологією.

Метрика SCMC визначає семантичну узгодженість із медичним корпусом (Semantic Coherence with Medical Corpus). Для кожного текстового пояснення обчислюється векторне представлення через модель sentence-transformers all-MiniLM-L6-v2, що відображає текст у 384-вимірний вектор. Через веб-пошук знаходяться топ-10 результатів пошуку для ключових термінів пояснення з медичних сайтів. Ці результати також відображаються у векторі. Метрика SCMC обчислюється як середня косинусна подібність. Високі значення SCMC свідчать про семантичну узгодженість пояснення з реальною медичною літературою, відображають використання стандартної медичної термінології.

Метрика RS характеризує читабельність текстів пояснень (Readability Score). Для кожного текстового пояснення обчислюється індекс читабельності Flesch Reading Ease з використанням програмної бібліотеки textstat. Індекс читабельності оцінює складність тексту на основі довжини речень та кількості складів у словах. Значення FRE інтерпретуються за шкалою: 90–100 (дуже легко), 80–90 (легко), 70–80 (помірно легко), 60–70 (стандартно), 50–60 (помірно важко), 30–50 (важко), 0–30 (дуже важко). Вищі значення відповідають кращій читабельності для зовнішніх користувачів, що не мають спеціалізованої підготовки у визначеній предметній області.

Результати експериментальної перевірки наведено у таблиці 1.

Таблиця 1

Результати експериментальної перевірки зрозумілості узагальненої символьної ментальної моделі рішення інтелектуальної системи

Метрика	Запропонована модель	SHAP (XGBoost)	LIME
TFWC (логарифм кількості результатів)	6,42	4,87	5,13
SCMC (косинусна подібність)	0,784	0,521	0,598
FRE (індекс Flesch)	68,3	42,1	51,6
Час побудови моделі (сек)	5,3	142,7	38,4

Запропонована узагальнена модель має найвищі значення TFWC – 6,42. Такі значення відповідають середній кількості результатів пошуку більше одного мільйона ключових термінів пояснень, наприклад, «high BMI diabetes risk», що свідчить про високу поширеність цих термінів у медичній веб-літературі та їхнє використання великою кількістю користувачів.

Базові підходи демонструють значно нижчі значення TFWC, зокрема, для SHAP TFWC становить 4,87, що відповідає десяткам тисяч результатів для термінів типу «SHAP value diabetes prediction».

Для LIME TFWC становить 5,13, що відповідає більше ніж 100 тисяч результатів для термінів типу «BMI contribution diabetes model». Такі результати пояснюються використанням технічної термінології машинного навчання замість стандартної медичної термінології. Остання менш поширена на медичних сайтах.

Запропонована модель має найвищу семантичну узгодженість між векторними представленнями пояснень та топ-10 результатів пошуку з медичних сайтів, що свідчить про використання термінології, яка використовується в популярній медичній літературі.

Запропонована трирівнева символьна ментальна модель характеризується незначними обчислювальними витратами. Побудова моделі займає 5 секунд для експериментальної вибірки, що значно менше порівняно з методом SHAP, який потребує майже 143 секунди для перенавчання та обчислення значень Шеплі для всіх екземплярів. І також менше порівняно з методом LIME, який потребує більше 38 секунд для генерації локальних лінійних апроксимацій. Висока швидкість автоматизованої побудови моделі пояснюється ефективністю алгоритмів класифікації на основі логік для онтологій малого і середнього розміру. Також модель не потребує перенавчання для кожного нового набору даних користувачів.

Результати експериментальної перевірки свідчать про ряд переваг розробленої узагальненої символьної ментальної моделі порівняно з традиційними підходами до побудови пояснень. По-перше, модель дає можливість сформувати зрозумілі каузальні шляхи міркування через символьне представлення знань засобами онтологій та правил виведення, що підтверджується високими значеннями метрик TFWC та SCMC. Використовується термінологія предметної області доменної термінології замість технічної термінології машинного навчання. По-друге, модель забезпечує можливість автоматизації процесу побудови ментальних моделей через механізми логічного виведення онтологій без залучення експертів для структурованих інтерв'ю, що усуває потребу в трудомістких якісних методах побудови ментальних моделей на основі когнітивного картування. В третіх, компоненти моделі можуть бути використані повторно для інших предметних областей внаслідок застосування доменно-незалежної схеми онтологічного представлення, що скорочує витрати на адаптацію моделі до нових варіантів застосування інтелектуальної системи. В четвертих, механізм композиції рівнів для підтримки деталізації пояснень відповідно до рівня експертизи користувача дозволяє адаптувати складність представлення пояснень до індивідуальних особливостей та потреб користувачів. Новачки отримують спрощене представлення на концептуальному рівні без перевантаження деталями. Користувачі середнього рівня можуть деталізувати концепти для вивчення каузальних правил виведення. Експерти отримують повний доступ до конкретних значень атрибутів для верифікації всіх компонентів міркування. На відміну від розробленої моделі, традиційні методи побудови пояснень формують однаковий рівень складності пояснень для всіх користувачів незалежно від їхніх потреб у деталізації.

Проте дана модель має ряд обмежень, що визначають напрямки майбутніх досліджень для підвищення ефективності та розширення можливостей застосування представленого підходу.

Якість моделі залежить від повноти та коректності експертних знань про каузальні шляхи міркування від атрибутів користувачів до концептів рішень. Розроблена модель має суттєві затримки при виконанні механізмів логічного виведення онтологій, що обмежує можливості застосування підходу для систем із суттєвими вимогами до затримок при формуванні пояснень. Також дана модель не призначена для пояснення роботи глибоких нейронних мереж, призначених для обробки візуальної інформації, представленої векторами пікселів. Запропонована модель орієнтована на системи підтримки рішень із табличними даними користувачів, де атрибути мають явне семантичне значення для формування зрозумілих правил виведення.

Висновки

Розроблена узагальнена символьна ментальна модель рішення інтелектуальної системи для зовнішніх користувачів, що задається на основі трирівневої архітектури з явним розділенням концептуального шару високорівневих абстракцій рішень, пояснювального шару каузальних правил виведення та бази знань доменних фактів. Модель передбачає використання онтологічної формалізації та правил виведення ланцюжка міркувань, а також механізму вертикальної композиції рівнів для деталізації інформації відповідно до індивідуального рівня експертизи користувача, що забезпечує формування зрозумілих каузальних шляхів міркування від високорівневих концептів рішень через символьні правила логічного виведення з використанням доменної термінології й до конкретних значень атрибутів користувачів.

Модель забезпечує зниження трудомісткості формування пояснень за рахунок виключення трудомісткого процесу побудови ментальних моделей експертами з використанням структурованих інтерв'ю.

Експериментальна перевірка узагальненої символьної ментальної моделі на наборі даних з медичної діагностики діабету типу 2 з використанням метрик на основі веб-пошуку та лінгвістичного аналізу підтвердила покращення зрозумілості пояснень порівняно з базовими підходами до побудови пояснень – методами SHAP та LIME. Практичне значення отриманих результатів полягає у створенні умов для побудови інструментальних засобів розробки пояснювальних інтелектуальних систем у ризикових предметних областях, таких як медичні діагностичні системи із поясненням клінічних рекомендацій для пацієнтів, фінансових систем планування з поясненням інвестиційних рекомендацій для клієнтів тощо.

Список літератури

[1] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, 30* (pp. 4765–4774). Curran Associates. https://arxiv.org/abs/1705.07874

[2] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "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). ACM. https://doi.org/10.1145/2939672.2939778

[3] Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence, 298*, Article 103502. https://doi.org/10.1016/j.artint.2021.103502

[4] Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

[5] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[6] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[7] Ambalavanan, R., Snead, R. S., Marczika, J., Towett, G., Malioukis, A., & Mbogori-Kairichi, M. (2025). Ontologies as the semantic bridge between artificial intelligence and healthcare. *Frontiers in Digital Health, 7*, Article 1668385. https://doi.org/10.3389/fdgth.2025.1668385

[8] Guizzardi, G., Fonseca, C. M., Almeida, J. P. A., Sales, T. P., Benevides, A. B., & Porello, D. (2022). Ontology-driven conceptual modeling as a service: Surveys, blueprints, and roadmaps for composable modeling platform. In *Conceptual Modeling* (pp. 3–18). Springer. https://doi.org/10.1007/978-3-031-17995-2_1

[9] Hossain, D., & Chen, J. Y. (2025). A study on neuro-symbolic artificial intelligence: Healthcare perspectives. *arXiv preprint*. https://arxiv.org/abs/2503.18213

[10] Olivares-Alarcos, A., Beßler, D., Khamis, A., Goncalves, P., Habib, M. K., Bermejo-Alonso, J., Barreto, M., Diab, M., Rosell, J., Quintas, J., Olszewska, J., Nakawala, H., Pignaton, E., Gyrard, A., Borgo, S., Alenya, G., Beetz, M., & Li, H. (2019). A review and comparison of ontology-based approaches to robot autonomy. *The Knowledge Engineering Review, 34*, Article e29. https://doi.org/10.1017/S0269888919000237

[11] Bampi, D., Miranda, W. K. de M., & Almeida, J. L. V. (2025). Ontology-driven monitoring system for ambient assisted living. *The Knowledge Engineering Review, 40*, Article e3. https://doi.org/10.1017/S0269888924000250

[12] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint*. https://arxiv.org/abs/1812.04608

[13] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[14] Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 295–305). ACM. https://doi.org/10.1145/3351095.3372852

[15] Szymanski, M., Millecamp, M., & Verbert, K. (2021). Visual, textual or hybrid: The effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (pp. 109–119). ACM. https://doi.org/10.1145/3397481.3450662

[16] Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (pp. 318–328). ACM. https://doi.org/10.1145/3397481.3450650

[17] Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2021). "Let me explain!": Exploring the potential of virtual agents in explainable AI interaction design. *Journal on Multimodal User Interfaces, 15*(2), 87–98. https://doi.org/10.1007/s12193-020-00332-0

[18] Chazette, L., Brunotte, W., & Speith, T. (2021). Exploring explainability: A definition, a model, and a knowledge catalogue. In *2021 IEEE 29th International Requirements Engineering Conference (RE)* (pp. 197–208). IEEE. https://doi.org/10.1109/RE51729.2021.00025

Received (Надійшла) 16.01.2026

Accepted for publication (Прийнята до друку) 23.02.2026

Publication date (Дата публікації) 27.03.2026