УДК 004.8:004.9 DOi 10.30837/ bi.2025.2(103).14

С. Ф. Чалий[1], І. О. Лещинська[1]

[1]ХНУРЕ, м. Харків, Україна, serhii.chalyi@nure.ua, ORCID iD: 0000-0002-9982-9091

[1]ХНУРЕ, м. Харків, Україна, iryna.leshchynska@nure.ua, ORCID iD: 0000-0002-8737-4595

МЕТОД ПОБУДОВИ НЕЙРОСИМВОЛЬНОГО ПРЕДСТАВЛЕННЯ МЕНТАЛЬНОЇ МОДЕЛІ РІШЕННЯ ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ

Розглянуто методи побудови ментальних моделей рішень інтелектуальних систем на основі інтеграції нейромережевих і символьних компонентів. Розроблено метод побудови нейросимвольного представлення ментальної моделі, який базується на двошаровій нейросимвольній архітектурі з можливістю виявлення прихованих ознак, відбору значущих ознак та механізмом нейросимвольного перетворення для відображення скритих представлень у символьні концепції. Метод містить етапи вилучення прихованих ознак, відбору ознак на основі уваги, нейросимвольного перетворення, побудови орієнтованого ациклічного графа для каузальної структури та перевірки каузальності з використанням лінійної темпоральної логіки. Метод створює умови для автоматизованого виявлення індивідуальних ментальних моделей користувачів із можливостями їх інтерпретації згідно з особливостями предметної області, а також побудови персоналізованих пояснень у системах пояснювального штучного інтелекту.

МЕНТАЛЬНІ МОДЕЛІ, НЕЙРОСИМВОЛЬНИЙ ШТУЧНИЙ ІНТЕЛЕКТ, ПОЯСНЮВАНИЙ ШТУЧНИЙ ІНТЕЛЕКТ, ОРІЄНТОВАНІ АЦИКЛІЧНІ ГРАФИ, ЛІНІЙНА ТЕМПОРАЛЬНА ЛОГІКА, ПЕРСОНАЛІЗОВАНІ ПОЯСНЕННЯ

S. F. Chalyi, I. O. Leshchynska. Method for constructing neurosymbolic representation of mental model of intelligent system decision. Methods for constructing mental models of intelligent system decisions based on integration of neural network and symbolic components are considered. A method for constructing neurosymbolic representation of mental model has been developed, based on dual-layer neurosymbolic architecture with capability for latent feature identification, significant feature selection, and neural-symbolic transformation mechanism for mapping hidden representations to symbolic concepts. The method includes stages of latent feature extraction, attention-based feature selection, neural-symbolic transformation, directed acyclic graph construction for causal structure representation, and causality verification using linear temporal logic. The method creates conditions for automated identification of individual user mental models with capabilities for their interpretation according to domain specifics, as well as construction of personalized explanations in explainable artificial intelligence systems.

MENTAL MODELS, NEUROSYMBOLIC ARTIFICIAL INTELLIGENCE, EXPLAINABLE ARTIFICIAL INTELLIGENCE, DIRECTED ACYCLIC GRAPHS, LINEAR TEMPORAL LOGIC, PERSONALIZED EXPLANATIONS

Вступ

Інтелектуальні інформаційні системи поєднують переваги традиційних інформаційних систем та систем штучного інтелекту при вирішенні комплексних задач у фінансовій сфері, промисловості, транспортній галузі [1]. Такі системи використовують моделі машинного навчання, що утруднює розуміння логіки їх роботи для користувачів [2]. Для забезпечення прозорості інтелектуальних інформаційних систем на сьогодні використовуються методи пояснювального штучного інтелекту (XAI) [3]. Для побудови зрозумілих пояснень в рамках XAI можуть бути використані ментальні моделі, які є внутрішніми представленнями користувачів про те, як працює інтелектуальна система, які каузальні залежності вона використовує для прийняття рішень [4, 5]. Ментальна модель відображає розуміння користувачем логіки роботи системи та обумовлює інтерпретацію отриманих рішень [6]. Використання ментальних моделей дає можливість адаптувати пояснення рішень інтелектуальної системи відповідно до рівня знань користувача та до його очікувань щодо можливостей застосування цього рішення [7].

Існуючі підходи до побудови ментальних моделей базуються переважно на використанні нейромережних та символьних методів [8, 9]. Перші дають можливість персоналізувати модель для користувача, проте представлені у вигляді чорної скриньки, що утруднює пояснення представлених в моделі залежностей [10]. Символьні методи використовують каузальне міркування і тому дають можливість сформувати явні залежності, які можуть бути безпосередньо інтерпретовані користувачем [11]. Однак ці методи потребують додаткових експертних знань для адаптації правил до індивідуальних потреб користувачів.

Таким чином, поєднання переваг обох підходів в рамках нейросимвольних архітектур створює умови для побудови пояснювальних персоналізованих ментальних моделей, що і свідчить про актуальність теми даного дослідження.

Нейросимвольні підходи до пояснювального штучного інтелекту реалізують інтеграцію нейромережних та символьних компонентів з використанням трьох основних парадигм [8, 9]. Перша парадигма, нейросимвольне перетворення, полягає у вилученні символьних правил з навчених нейронних мереж. Друга парадигма,

символьно-нейронне вбудовування, полягає у введенні символьних обмежень у процес навчання нейронних мереж [12]. Третя парадигма, використання гібридних архітектур, передбачає паралельну роботу нейромережевого та символьного компонентів [9].

Моделі концептуального вузького місця (Concept Bottleneck Models) використовують інтерпретовані концепти як проміжний шар між входом та виходом мережі [12]. Обмеженням даного підходу є статичність символьних компонентів та відсутність персоналізації під індивідуальні ментальні моделі користувачів [6].

Методи побудови ментальних моделей у когнітивній психології визначають ментальні моделі як внутрішні представлення зовнішнього світу, які користувачі використовують для міркування та прогнозування. Витягування ментальних моделей (mental model elicitation) виконується за допомогою структурованих інтерв'ю та побудови концептуальних схем моделей [4, 13]. Підходи на основі аналізу поведінки [14] аналізують патерни взаємодії користувачів з інтелектуальною системою для імпліцитного виявлення ментальних моделей. Обмеженням цього підходу виступає відсутність масштабованості ручного витягування та неможливість забезпечити пояснюваність каузальних зв'язків у виявлених моделях [6].

Двошарові архітектури в штучному інтелекті розділяють систему на нейромережний шар для адаптивного навчання та символьний шар для верифікованого міркування з двонаправленим потоком інформації [9, 10].

Фреймворк каузального міркування (causal reasoning framework) використовує каузальні байєсівські мережі у символьному шарі для верифікації каузальних залежностей. Обмеженням цієї архітектури є відсутність специфікації для побудови ментальних моделей у системах пояснювального штучного інтелекту [7].

Для визначення каузальних залежностей у ментальних моделях використовують темпоральні знання, які фіксують часові послідовності змін станів керованого об'єкта [15]. Автоматизоване керування базами знань забезпечують темпоральні правила у логіко-ймовірнісній формі, що використовують оператори лінійної темпоральної логіки (LTL): NeXt, Future, Until [16]. Модель представлення темпоральних знань містить множину фактів виникнення станів, темпоральні відношення між фактами, а також операції над фактами, що дає можливість відобразити багатоваріантність рішень із заданим ступенем деталізації для відповідного ієрархічного рівня організації. Темпоральні залежності дають можливість формалізувати послідовність керуючих дій у часі та верифікувати каузальні зв'язки через темпоральні обмеження [15, 16].

Таким чином, існуючі підходи окремо виявляють латентні ознаки, реалізують каузальне міркування, а також виконують темпоральну перевірку з використанням лінійної темпоральної логіки. Відповідно, задача розробки підходу, що поєднує виявлення ознак, каузальне міркування та темпоральну верифікацію для побудови нейросимвольного представлення ментальної моделі потребує свого вирішення.

1. Постановка задачі

Метою є розробка підходу до побудови двошарової нейросимвольної архітектури, яка забезпечує інтерпретованість залежностей в ментальній моделі, автоматизацію виявлення індивідуальних ментальних моделей та можливість пояснюваності моделі на основі інтеграції компонентів глибокого навчання й каузального графа з подальшою перевіркою узгодженості у часі з використанням темпоральної логіки.

Для досягнення поставленої мети вирішуються такі задачі:

– розробка підходу до побудови нейромережного та символьного шарів ментальної моделі рішення інтелектуальної системи з урахуванням можливості побудови персоналізованого пояснення;

– розробку методу побудови ментальної моделі на основі двошарової архітектури з варіаційним автокодувачем для вилучення прихованих ознак, механізмом уваги для відбору значущих ознак, нейросимвольним перетворенням для відображення прихованих ознак у символьне представлення, орієнтованим ациклічним графом для відображення каузальної структури моделі з використанням лінійної темпоральної логіки для перевірки каузальних залежностей.

2. Підхід до побудови нейромережного та символьного шарів ментальної моделі рішення інтелектуальної системи

Розроблений підхід орієнтований на формування двошарової архітектури, яка інтегрує нейромережний шар, призначений для адаптивного виявлення прихованих ознак з поведінкових даних користувачів, та символьний шар, призначений для реалізації причинно-наслідкового міркування з можливістю подальшої перевірки засобами темпоральної логіки.

Запропонована двошарова нейросимвольна архітектура включає такі основні компоненти: варіаційний автокодувач; багатоголовий механізм уваги; нейросимвольний перетворювач; орієнтований ациклічний граф. Варіаційний автокодувач використовується для ймовірнісного виявлення латентних ознак. Багатоголовий механізм уваги реалізує відбір значущих ознак на основі фільтрації поведінкового шуму. Нейросимвольний перетворювач відображує латентні представлення у символьні концепції. Орієнтований ациклічний граф призначений для побудови каузальної структури. Перевірка каузальних залежностей виконується з використання лінійної темпоральної логіки шляхом фільтрації хибних залежностей.

Розглянемо ключові особливості запропонованого підходу.

Нейромережевий шар приймає на вхід послідовність $\mathbf{X}$ векторів спостережень $\mathbf{x}_i$ у моменти часу i: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_i, ..., \mathbf{x}_I\}$. Наприклад, при вирішенні задачі підтримки прийняття рішень з медичної діагностики вектор $\mathbf{x}_i$ може містити тип дії (перегляд симптому, замовлення тесту, вибір діагнозу), темпоральну мітку, параметри взаємодії з інтерфейсом системи. Варіаційний автокодувач кодує поведінкову послідовність у латентний простір. Ймовірнісний характер латентного простору дає можливість розрізняти користувачів з різним рівнем впевненості: новачки мають високу невизначеність (різноманітна поведінка), експерти – низьку (представлену стабільними патернами).

Регуляризація виконується з використанням дивергенції Кульбака-Лейблера (KL-divergence) між апостеріорним розподілом $q_\phi(z \mid X)$ та апріорним розподілом $p(z)$, шо дає можливість виконати інтерполяцію між ментальними моделями різних користувачів. Функція втрат $\mathcal{L}_{VAE}$ варіаційного автокодувача має вигляд:

$$\mathcal{L}_{VAE} = \mathbb{E}_{q_\phi(z|X)}\left[\log p_\theta(X \mid z)\right] - \\ -\beta \cdot D_{KL}(q_\phi(z \mid X) \,\|\, p(z)). \tag{1}$$

Формула (1) складається з двох доданків. Перший доданок $\mathbb{E}_{q_\phi(z|X)}[logp_\theta(X \mid z)]$ – це математичне сподівання $\mathbb{E}_{q_\phi(z|X)}$ логарифма ймовірності відновлення вхідної послідовності $X \mid z$ з латентного коду z декодером з параметрами θ, що забезпечує точність кодування поведінкових патернів.

Високе значення першого доданку свідчить про точне відновлення поведінкових патернів. Наприклад, якщо користувач спочатку переглядає симптоми, потім замовляє тести, автокодувач має коректно відновити цю послідовність з латентного представлення.

Другий доданок $D_{KL}(q_\phi(z \mid X) \,\|\, p(z))$ – це дивергенція Кульбака-Лейблера між апостеріорним розподілом (навчений розподіл латентних кодів для даного користувача) та апріорним розподілом (стандартний нормальний розподіл). Дивергенція Кульбака-Лейблера визначає, наскільки розподіл $q_\phi(z \mid X)$ відрізняється від $p(z)$. Цей доданок виконує регуляризацію для отримання латентного простору без ізольованих кластерів. Параметр β контролює баланс між відновленням і регуляризацією. Типове значення $\beta = 1$. При $\beta > 1$ підвищується можливість інтерпретувати окремі розмірності латентного простору.

Багатоголовий механізм уваги обробляє латентні представлення $\mathbf{z}$ через H паралельних голів уваги для відбору значущих ознак. Для кожної голови уваги та латентного представлення $\mathbf{z}$ традиційно обчислюються матриці запитів $\mathbf{Q}$, ключів $\mathbf{K}$ та значень $\mathbf{V}$ розмірністю d_k.

Увага Attention($\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$) для кожної голови обчислюється таким чином:

$$\text{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}. \tag{2}$$

Вираз (2) відображає механізм вибіркової уваги. Функція softmax перетворює схожості у ваги уваги (сума ваг дорівнює 1). Скалярний добуток $\mathbf{Q}\mathbf{K}^\top$ відображає схожість між запитами та ключами. Цей скалярний добуток може приймати великі значення при збільшенні розмірності матриць. Результатом може бути насичення функції softmax та виникнення проблеми зникаючих градієнтів. Масштабування на $\sqrt{d_k}$ нормалізує ці значення, забезпечуючи стабільність градієнтів при великих розмірностях матриць. Множення на $\mathbf{V}$ дає можливість відібрати значущі ознаки. Так, одна голова уваги може фокусуватися на темпоральних патернах, визначаючи порядок дій, а інша – на типах дій, наприклад діагностичних тестах або терапевтичних рішеннях.

Після того, як кожна з голів уваги обробила латентне представлення згідно (2), виконується об'єднання (конкатенація) отриманих векторів h_j розміру d_k у один довгий вектор $Concat\left(h_1,\ldots,h_H\right)$ розмірності $H \cdot d_k$. Об'єднаний вектор перемножується з проекційною матрицею $\mathbf{W}^O$, в результаті чого виконується агрегація інформації від усіх голів уваги, а також відновлюється сумісність розмірностей з латентним представленням $\mathbf{z}$. Результуюче представлення $\mathbf{z}_{\text{salient}}$ має вигляд:

$$\mathbf{z}_{\text{salient}} = \text{Concat}\left(h_1,\ldots,h_H\right)\mathbf{W}^O. \tag{3}$$

Відфільтроване представлення $\mathbf{z}_{\text{salient}}$ містить лише значущі ознаки для символьного шару, що створює умови для інтерпретації рішення.

Нейросимвольне перетворення відображає латентні ознаки $\mathbf{z}_{\text{salient}}$ на множину символьних концептів $C = \{c_1, c_2, \ldots, c_L\}$. Попередньо визначається концептуальний словник $V = \{v_1, v_2, \ldots, v_R\}$ предметної області.

Наприклад, для медичної системи підтримки прийняття рішень словник може містити концепти: «аналіз симптомів», «диференційна діагностика», «призначення тестів», «оцінка ризиків», «вибір терапії».

Нейромережний класифікатор обчислює розподіл ймовірностей над концептуальним словником:

$$p(c \mid \mathbf{z}_{\text{salient}}) = \text{softmax}\left(f\left(\mathbf{z}_{\text{salient}}\right)\right). \tag{4}$$

Зокрема, ймовірність $p(c_i \mid \mathbf{z}) = 0{,}85$ означає, що латентні ознаки з впевненістю 0,85 відповідають концепту c_l.

При побудові символьного шару відбираються концепти з найвищими ймовірностями. Кожному концепту присвоюється семантичне вбудовування e_{c_l} з попередньо навченого ембеддінгу для структурованого представлення у побудові графа.

Можливість двонаправленої інтеграції створює умови для узгодженості між нейромережним та сим-

вольним шарами. Така інтеграція полягає у зворотній перевірці, коли символьні обмеження з графа можуть бути використані для регуляризації навчання нейронної мережі.

Символьний шар призначений, щоб сформувати орієнтований ациклічний граф $G=(C,E)$ для множини символьних концептів C з тим, щоб явно представити ментальну модель за допомогою каузальних залежностей.

Для кожної пари концептів (c_l,c_m) обчислюється показник $s(c_l \to c_m)$ на основі перевірки можливої каузальної залежності між ними. В даному випадку оцінюється каузальна залежність між концептами на основі поведінкових даних, коли за концептом c_l безпосередньо (тобто з використанням оператора $Next$) слідує концепт c_m:

$$s(c_l \to c_m)=\frac{1}{N}\sum_{n=1}^{N}\left(1\ iff\ \left((c_l,n)\,Next(c_m,n)\right)\right). \qquad (5)$$

Згідно (5) каузальний зв'язок між концептами у метальній моделі оцінюється на основі послідовності виявлення концептів (c_l,c_m) у вхідних даних. Тобто, якщо у поведінковій послідовності концепт c_l (наприклад, «аналіз симптомів») передує в часі концепту c_m (наприклад, «призначення тестів»), то значення функції під знаком суми дорівнюватиме 1. В іншому випадку значення дорівнюватиме 0. Усереднення за всіма отриманими послідовностями визначає частоту спільної появи пар (c_l,c_m) за умови, що c_l передує c_m. Орієнтоване ребро $c_l \to c_m$ додається до графа G, якщо $s(c_l \to c_m)$ перевищує порогове значення, а також відсутня ациклічність, тобто не існує шляху по графу у зворотному напрямку. Такий підхід забезпечує причинно наслідкове упорядкування пар (c_l,c_m) у графі. Тобто за умови, що c_l передує c_m, c_l можна розглядати як причину для c_m.

Перевірка каузальності через лінійну темпоральну логіку дає можливість підтвердити коректність каузальних дуг графа G. Для кожної каузальної дуги $c_l \to c_m$ формується обмеження лінійної темпоральної логіки:

$$G(c_l\ F\ c_m), \qquad (6)$$

де темпоральний оператор G означає, що послідовність $c_l \to c_m$ має виконуватись для всіх траєкторій, а темпоральний оператор F задає послідовність «c_l передує c_m у деякій наступній точці у майбутньому».

Відповідно, формула (6) має значення «якщо відбувається c_l, то пізніше відбудеться c_m».

Також у процесі перевірки обчислюється відношення послідовностей, для яких виконується умова (6), до загальної кількості послідовностей. Ребра зі значенням відношення меншим за порогове видаляються з графа як хибні залежності, які були сформовані внаслідок випадкових збігів у даних.

3. Метод побудови нейросимвольного представлення ментальної моделі рішення інтелектуальної системи

Розроблений метод формує ментальну модель на основі представленої двошарової архітектури. Метод включає наступні етапи.

Етап 1. Вилучення прихованих ознак варіаційним автокодувачем.

Вхідна послідовність, що відображає дії користувача, кодується у латентний простір. Декодер реконструює вхідну послідовність з латентного представлення для навчання енкодера. Регуляризація з використанням дивергенції Кульбака-Лейблера (вираз 1) забезпечує відсутність ізольованих кластерів у латентному просторі і, відповідно, створює умови для плавної інтерполяції.

Етап 2. Відбір ознак на основі уваги для фільтрації поведінкового шуму.

Латентне представлення обробляється багатоголовим механізмом уваги з метою виявлення значущих ознак. Для кожної голови уваги обчислюються запити, ключі та значення. Ваги уваги для кожної голови обчислюються за формулою (2). Виходи голів уваги агрегуються через конкатенацію (3). Фільтрація поведінкового шуму відбувається на основі відкидання ознак з низькими вагами уваги. В результаті знижується розмірність входу для символьного шару.

Етап 3. Нейросимвольне перетворення через відображення прихованих ознак у символьне представлення.

Відфільтровані латентні ознаки відображаються на множину символьних концептів через класифікатор. Класифікатор обчислює розподіл ймовірностей над попередньо визначеним концептуальним словником предметної області. Відбираються концепти з найвищими ймовірностями. Кожному концепту присвоюється семантичне вбудовування з попередньо навченого ембеддінгу для структурованого представлення у побудові графа.

Етап 4. Побудова орієнтованого ациклічного графа для відображення каузальної структури ментальної моделі.

На базі множини символьних концептів будується орієнтований ациклічний граф, який представляє каузальну структуру ментальної моделі. Для кожної пари концептів обчислюється оцінка каузальності (5).

Орієнтоване ребро додається до орієнтованого графа, якщо оцінка каузальності перевищує порогове значення (наприклад, 0,5), а також шлях у зворотному напрямку відсутній.

Етап 5. Перевірка орієнтованого ациклічного графа з використанням лінійної темпоральної логіки.

Каузальні ребра графа перевіряються на відповідність обмеженням, представленим формулами лінійної темпоральної логіки. Для кожного каузального ребра формується обмеження у вигляді правила типу Future

лінійної темпоральної логіки: $G(c_l \ F \ c_m)$. Порушення темпоральних обмежень свідчать про некоректні каузальні залежності. При перевірці обмежень обчислюється відношення послідовностей, що задовольняють обмеженню, до загальної кількості послідовностей. Ребра графа, що мають значення відношення нижче, ніж порогове значення (наприклад, 0,8), вилучаються з графа, оскільки вони моделюють нестійкі каузальні залежності.

Отриманий в результаті імплементації методу направлений граф дає можливість пояснити зв'язки між елементами рішення у ментальній моделі.

4. Експериментальна перевірка розробленого методу

Експериментальна перевірка розробленого методу виконана з використанням синтетичних медичних даних, отриманих з використанням рушія Synthea (Synthetic Health Data Engine). Ці дані моделюють клінічні траєкторії пацієнтів на основі реальних епідеміологічних даних та клінічних протоколів США. Набір даних представлено в офіційному репозиторії Synthea за посиланням https://mitre.box.com/shared/static/aw9po06ypfb9hrau4jamtvtz0e5ziucz.zip. З повного набору даних для експерименту відібрано таблицю conditions.csv, яка містить часові послідовності станів пацієнтів (діагнози, соціальні та поведінкові фактори здоров'я) з повними часовими мітками початку (START) та завершення (STOP) кожного стану, ідентифікаторами пацієнта (PATIENT) та епізоду надання допомоги (ENCOUNTER), а також кодами та описами станів

Набір даних містить 38 094 записів станів для 1 147 пацієнтів з 26 904 унікальними епізодами надання допомоги та 202 унікальними кодами станів. Середня кількість станів на одного пацієнта дорівнює 33,2. Дані охоплюють період з 1944 до 2024 року, що забезпечує можливість перевірки темпоральних обмежень. Для кожного пацієнта задано впорядковану послідовність станів згідно з міткою START, що відповідає поведінковим послідовностям користувачів у постановці задачі методу.

При проведенні експерименту використано спрощену реалізацію запропонованої двошарової архітектури, яка забезпечує повну реалізацію методу, але дає можливість знизити обчислювальні витрати.

Для формування вхідних ознак нейромережевого шару кожний стан класифіковано за тривалістю у три категорії: гострі короткострокові стани (acute_short, тривалість менше 30 днів), середньострокові стани (medium_term, тривалість від 30 до 365 днів) та хронічні/тривалі стани (chronic_or_open, тривалість більше 365 днів або відсутність дати завершення STOP). Ця класифікація відповідає доменній інтерпретації у медичній практиці: гострі епізодічні захворювання (вірусні інфекції, травми), середньострокові курси лікування (реабілітація, терапія) та хронічні стани (гіпертензія, діабет, серцева недостатність). Для кожного пацієнта обчислено вектор з семи ознак: частки трьох типів станів у його послідовності (acute_prop, medium_prop, chronic_prop), середня та максимальна тривалість станів (avg_dur, max_dur), кількість унікальних епізодів надання допомоги (n_enc) та часовий розмах між першим та останнім станом (span_days).

У процесі експериментальної перевірки оцінювались можливість пояснити рішення з використанням нейросимвольного представлення ментальної моделі, а також точність персоналізації, темпоральна узгодженість та час виконання. При оцінці можливості пояснити рішення використана шкала від 1 до 5, де 5 відповідає повній прозорості каузальних залежностей без потреби додаткових пояснень. Точність персоналізації оцінювалась як якість кластеризації латентних представлень. Для оцінки використано коефіцієнт силуету у відсотковій шкалі, який визначає, наскільки кожен елемент даних підходить до свого кластера у порівнянні з іншими кластерами. Темпоральна узгодженість розраховується як відношення кількості шляхів лікування, для яких виконана формула лінійної темпоральної логіки, до загальної кількості шляхів лікування. Оскільки кожний шлях лікування у наборі даних пов'язаний із окремим пацієнтом, то розрахунок виконано по пацієнтам. Обчислювальні витрати при проведенні експерименту оцінювались через час виконання у секундах в розрахунку на одного користувача.

Порівняння розробленого методу виконано для трьох базових підходів. В рамках першого підходу використана нейромережна архітектура без символьного шару та без темпоральної перевірки. Другий базовий підхід базується на використання лише правил виду: «якщо chronic_prop > 0,6, то кластер «хронічний»; «якщо acute_prop > 0,4, то кластер «гострий»; інакше кластер «змішаний». Тобто даний підхід є типовим для класичних експертних систем із апріорно заданими правилами. В рамках третього базового підходу використана гібридна архітектура із типовою кластеризацією, без механізму уваги, без проекції у латентний простір та без фільтрації значущих ознак. Тобто дана архітектура не містить механізму інтеграції шарів.

Результати експериментальної перевірки наведено у табл. 1.

Оцінка пояснення для розробленого методу становить 4,3 з 5,0, що суттєво перевищує можливості чистої нейромережної архітектури (3,2 з 5,0) завдяки явним символьним концептам та каузальному графу переходів між типами станів. Чиста символьна архітектура має найвищий рівень зрозумілості пояснень (4,5 з 5,0) внаслідок використання детермінованих правил, але поступається розробленому методу за точністю персоналізації. Нейромережна архітектура показує таку саму точність персоналізації (77,6%), але не надає можливості побудови інтерпретованих ментальних

моделей внаслідок відсутності символьного шару. Гібридна архітектура без механізму уваги демонструє нижчий рівень персоналізації внаслідок шуму у вхідних ознаках, оскільки цей шум не було відфільтровано з використанням механізму уваги.

Таблиця 1

Порівняння методів побудови ментальних моделей на даних Synthea

Метод	Оцінка пояснюваності (1–5)	Точність персоналізації (%)	Темпоральна узгодженість (%)	Час виконання (сек/користувач)
Нейромережна архітектура	3,2	77,6	–	3,1
Символьна архітектура	4,5	78,2	80,4	12,3
Гібридна архітектура без уваги	3,6	75,2	81,2	8,9
Розроблена двошарова архітектура	**4,3**	**77,6**	**82,8**	**6,8**

Темпоральна узгодженість розробленого методу досягає 82,8%, що відповідає шляхам лікування з періодичними гострими епізодами на фоні хронічних станів. Зокрема, кластер «глибоко хронічні» має найвище значення темпоральної узгодженості на рівні 89,0%, кластер «стабільні хронічні» має значення 80,8%, а кластер «епізодичні» – 64,7%. Середньозважена темпоральна узгодженість 82,8% забезпечується внаслідок фільтрації хибних залежностей, лише пацієнти зі стійкими патернами переходів між станами задовольняють темпоральному правилу (6). Символьна архітектура має нижчу темпоральну узгодженість внаслідок того, що правила не є персоналізованими. Гібридна архітектура без механізму уваги має проміжне значення узгодженості у порівнянні з розробленим методом та підходом на основі правил. Темпоральна узгодженість для нейромережної архітектури не розраховувалась внаслідок відсутності символьних концептів й відповідних формальних залежностей, представлених засобами темпоральної логіки.

Обчислювальна ефективність розробленого методу має середнє значення, становить 6,8 секунд на одного користувача. Такі витрати часу є допустимими для пакетної обробки даних. Нейромережна архітектура має найменші витрати часу на користувача, через те, що не формувався направлений ациклічний граф й не виконувалась темпоральна перевірка. Використання правил пов'язано з найбільшими витратами часу внаслідок перевірки всіх правил на всіх послідовностях без фільтрації цих правил. Гібридна архітектура без уваги показує проміжний час виконання (8,9 секунди на користувача).

Порівняльний аналіз по кластерам показує, що кластер «глибоко хронічні пацієнти», який включає 484 особи, має найвищу темпоральну узгодженість 89%. Причина такого рівня узгодженості полягає в тому, що з 484 пацієнтів з хронічними станами 431 особа не має трьох і більше послідовних гострих епізодів після появи хронічного стану. Кластер 0 «стабільні хронічні пацієнти» включає 417 осіб та має рівень темпоральної узгодженість 80,8%. Тобто гострі епізоди виникають, але не формують довгі кластери, ймовірно завдяки менеджменту. Кластер «епізодичні пацієнти» включає 246 осіб та має найнижчу темпоральну узгодженість 64,7%, що свідчить про відсутність постійного контролю хронічних станів.

Таким чином, розроблений метод забезпечує можливість формування пояснень за рахунок символьного шару з трьома інтерпретованими ментальними моделями: «глибоко хронічні пацієнти», «стабільні хронічні пацієнти», «епізодичні пацієнти». Перехід між типами станів представляється у вигляді каузального графа. Програмно згенерований при проведенні експерименту граф ментальної моделі наведено на рис. 1. Використано мову програмування Python.

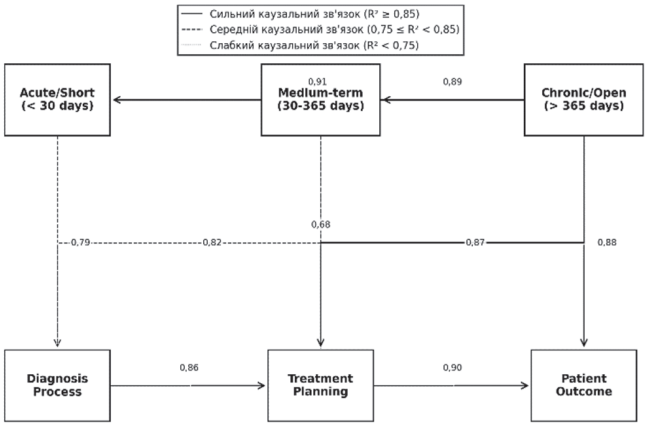

Рис. 1. Граф ментальної моделі

Представлений граф на верхньому рівні містить каузальні залежності між шістьма концептами ментальної моделі, які представляють типи захворювань – Acute/Short, Medium-term, Chronic/Open. Нижній рівень граф відображає клінічні процеси – Diagnosis Process, Treatment Planning, а також та результат лікування – Patient Outcome.

Граф включає 8 залежностей, виділених за силою зв'язку. Суцільна лінія позначає сильні каузальні зв'язки. Ми виявили п'ять сильних каузальних залежностей.

Залежність Chronic/Open – Acute/Short свідчить, що загальний стан приводить в майбутньому до виникнення загострень у захворюванні. Схожа каузальна залежність Chronic/Open – Medium-term демонструє перехід від хронічного стану до середньострокових епізодів. Найсильніша каузальна залежність Chronic/Open – Treatment підтверджує, що хронічні стани

приводять до ініціювання лікування, що відповідає клінічним протоколам.

Процес із залежністю Chronic/Open – Medium-term представляє собою традиційний клінічний потік робіт, в якому діагностика є передумовою для планування лікування.

Зв'язок Treatment Planning – Patient Outcome характеризує вплив лікування на результат. Велика вага правила підтверджує ефективність терапевтичного втручання.

Штрихова лінія маркує середні каузальні зв'язки. Зв'язок Acute/Short – Diagnosis (кореляція = 0,78) між гострими станами пацієнтів та діагностуванням описує типову ситуацію, коли пацієнти звертаються за невідкладною допомогою. Середня сила каузального зв'язку пояснюється випадками самолікування або відкладеного звернення.

Зв'язок Medium-term – Diagnosis (кореляція = 0,82) для середньострокових захворювань вказує на системний підхід до обстеження пацієнтів при тривалих симптомах хвороби.

Слабкий каузальний зв'язок Acute/Short – Patient Outcome вказує на високу варіативність результатів лікування при гострих станах. Такий зв'язок узгоджується з клінічною практикою, де прогноз залежить від таких чинників, як своєчасність звернення та стан здоров'я.

Таким чином, на основі аналізу графа ментальної моделі експерт може безпосередньо оцінити коректність віднесення пацієнта до кластера «глибоко хронічні» на основі частки його хронічних станів та протяжності клінічного шляху в часі. Персоналізація результатів реалізується через адаптивне виявлення прихованих ознак у латентному просторі, що дозволяє виявляти поведінкові патерни кожного пацієнта без додаткового налаштування правил. Кластеризація на латентних координатах виявила три стани пацієнта без попереднього знання про їх існування. Темпоральна узгодженість реалізується шляхом перевірки на відповідність формулам лінійної темпоральної логіки. Незалежність від предметної області забезпечується шляхом заміни концептуального словника. Наприклад, замість медичних станів можна використовувати відомі типи дій з підтримки прийняття рішень.

При проведенні експерименту були використані синтетичні медичні дані, які моделюють життєві траєкторії пацієнтів з високою варіативністю, в тому числі враховують випадкові травми, інфекції, а також соціальні фактори. Ця особливість темпоральних даних знижує темпоральну узгодженості порівняно з поведінковими логами медичної системи підтримки рішень, оскільки реальні логи медичних систем з фіксованими протоколами прийняття рішень мають більш стабільні темпоральні патерни без випадкових зовнішніх подій. Тобто метод демонструє темпоральну узгодженість 82,8 з використанням шляхів лікування з високою варіативністю. Проте на структурованих поведінкових логах медичних систем така узгодженість може бути суттєво вищою.

Метод потребує вхідних даних з темпоральними мітками при вирішенні задачі побудови персоналізованих пояснень.

Перспективи подальших досліджень включають адаптацію розробленого методу для інкрементного навчання в режимі онлайн при побудові ментальних моделей, а також інтеграцію з методами активного навчання для інтерактивного уточнення ментальних моделей в процесі діалога з користувачем.

5. Висновки

Запропоновано підхід до інтеграції нейромережного та символьного шарів для побудови ментальних моделей рішень з урахуванням вимог формування пояснень та адаптивності. Пояснення з використанням ментальної моделі базується на каузальних ациклічних графах з перевіркою отриманих казуальних залежностей за допомогою лінійної темпоральної логіки за умови збереження нейромережної адаптації до індивідуальних шаблонів поведінки користувача.

Розроблено метод побудови ментальної моделі на основі двошарової архітектури. Метод містить п’ять етапів: вилучення прихованих ознак варіаційним автокодувачем; відбір ознак на основі уваги для фільтрації поведінкового шуму; нейросимвольне перетворення через відображення прихованих ознак у символьне представлення; побудова орієнтованого ациклічного графа для відображення каузальної структури моделі; валідація з використанням лінійної темпоральної логіки. Метод створює умови для автоматизованого виявлення індивідуальних ментальних моделей користувачів із можливістю сформувати пояснення згідно особливостей предметної області.

Експериментальна перевірка на синтетичних медичних даних Synthea показала, що розроблений метод забезпечує можливість побудови персоналізованих пояснень на основі інтерпретації ментальної моделі користувача інтелектуальної системи.

Список літератури:

[1] Kautz H. The third AI summer: AAAI Robert S. Engelmore memorial lecture / H. Kautz // AI Magazine. – 2022. – Vol. 43, No. 1. – P. 93–104. DOI: https://doi.org/10.1002/aaai.12036.

[2] Gunning D. DARPA’s explainable artificial intelligence (XAI) program / D. Gunning, D. Aha // AI Magazine. – 2019. – Vol. 40, No. 2. – P. 44–58. DOI: https://doi.org/10.1609/aimag.v40i2.2850.

[3] Chalyi S. Externalization of tacit knowledge in the mental model of a user of an artificial intelligence system / S. Chalyi, I. Leshchynska // Bulletin of National Technical University "KhPI". Series: System Analysis, Control and Information Technologies. – 2024. – Vol. 1. – P. 91–96. https://doi.org/10.20998/2079-0023.2024.01.15.

[4] Johnson-Laird P. N. Mental models and human reasoning / P. N. Johnson-Laird // Proceedings of the National Academy of Sciences. – 2010. – Vol. 107, No. 43. – P. 18243–18250. DOI: https://doi.org/10.1073/pnas.1012933107.

[5] Hoefer M. Designing AI systems for mental model development / M. Hoefer, A. Felfernig // CEUR Workshop Proceedings. – 2025. – Vol. 3957. – P. 9–14.

[6] Bansal G. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff / G. Bansal, T. Wu, J. Zhou, et al. // Proceedings of the AAAI Conference on Artificial Intelligence. – 2019. – Vol. 33, No. 1. – P. 2429–2437. DOI: doi: 10.1609/aaai.v33i01.33012429.

[7] Чалий С.Ф., Лещинська І.О. Уточнення ментальної моделі рішення на основі доповнення вхідних даних в задачі формування пояснень в інтелектуальній системі. АСУ та прилади автоматики. – 2024. – Вип. 182. – С. 66-72. https://doi.org/ 10.30837/0135-1710.2024.182.066

[8] Sarker M. K. Neuro-symbolic artificial intelligence: Current trends / M. K. Sarker, L. Zhou, A. Eberhart, et al. // 10.48550/arXiv.2105.05330.

[9] Hitzler P. Neuro-symbolic integration for AI / P. Hitzler, A. Eberhart, M. Ebrahimi, et al. // Neuro-Symbolic Artificial Intelligence: The State of the Art. – National Science Review. 9. 10.1093/nsr/nwac035.

[10] Mao J. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision / J. Mao, C. Gan, P. Kohli, et al. // 10.48550/arXiv.1904.12584. – 2019.

[11] Verma S. Counterfactual Explanations for Machine Learning: A Review/ Verma, Sahil & Dickerson, John & Hines, Keegan. // 10.48550/arXiv.2010.10596. – 2020.

[12] Koh P. W. Concept bottleneck models / P. W. Koh, T. Nguyen, Y. S. Tang, et al. // Proceedings of the 37th International Conference on Machine Learning (ICML). – 2020. – P. 5338–5348.

[13] Чалий С. Ф. Концептуальна ментальна модель пояснення в системі штучного інтелекту / С. Ф. Чалий, І. О. Лещинська // Вісник Національного технічного університету «ХПІ». Серія: Системний аналіз, управління та інформаційні технології. – 2023. – № 1(9). – С. 70–75. DOI: https://doi.org/10.20998/2079-0023.2023.01.11

[14] Chalyi S. A dynamic explanation model for human-computer interaction in an artificial intelligence system / S. Chalyi, I. Leshchynska // Advanced Information Systems. – 2020. – Vol. 4(4) 1. – P. 114–119. https://doi.org/10.20998/2522-9052.2020.4.16

[15] Чала О. В. Модель узагальненого представлення темпоральних знань для інтелектуальних систем підтримки прийняття рішень / О. В. Чала // Вісник Національного технічного університету «ХПІ». Системний аналіз, управління та інформаційні технології. – 2020. – Вип. 1(3). – С. 14–18. DOI: 10.20998/2079-0023.2020.01.03

[16] Levykin V. Development of a method of probabilistic inference of sequences of business process activities to support business process management / V. Levykin, O. Chala // Eastern-European Journal of Enterprise Technologies. – 2018. – Vol. 5, No. 3(95). – P. 16–24. DOI: https://doi.org/10.15587/1729-4061.2018.142664.

Надійшла до редколегії 30.10.2025