Построение интеллектуальной системы дифференциальной диагностики неходжкинских лимфом

М. В. Седых*, Ю. А. Бродская‡

Предложена гибридная система дифференциальной диагностики неходжкинских лимфом, объединяющая статистические и нейросетевые модели с правилами Prolog и нечётким выводом Мамдани. Научная новизна заключается в двухконтурном согласовании, использующем семантические ограничения для фильтрации клинически некорректных гипотез. На тестовой выборке макроусреднённая метрика F_1 составила 0,75.

Ключевые слова: неходжкинские лимфомы, гибридные интеллектуальные системы, нечёткая логика, логическое программирование, машинное обучение, объяснимый ИИ.

1. Введение

Рост заболеваемости лимфомами и дефицит гематологов обусловливают потребность в СППР. Цель работы — разработка гибридной СППР для робастной классификации неходжкинских лимфом при неполных данных. Научная новизна — семантико-вероятностный консилиум на основе Prolog и нечёткого вывода, корректирующий вероятности с учётом клинических ограничений и ИГХ-маркеров при пропусках данных.

* *Седых Матвей Владимирович* — аспирант, Саратовский государственный технический университет имени Ю.А. Гагарина, Российская Федерация, 410054, г. Саратов, ул. Политехническая, 77, e-mail: ghetreek@bk.ru, ORCID: 0009-0008-0440-9715.

Sedykh Matvey Vladimirovich — Postgraduate Student, Yuri Gagarin State Technical University of Saratov, 77 Politekhnicheskaya St., Saratov, 410054, Russian Federation.

‡ *Бродская Юлия Анатольевна* — канд. физ.-мат. наук, доцент, Саратовский государственный технический университет имени Ю.А. Гагарина, Российская Федерация, 410054, г. Саратов, ул. Политехническая, 77, e-mail: gekata21@yandex.ru, ORCID: 0009-0004-7377-5488.

Brodskaya Yulia Anatolyevna — Candidate of Physical and Mathematical Sciences, Associate Professor, Yuri Gagarin State Technical University of Saratov, 77 Politekhnicheskaya St., Saratov, 410054, Russian Federation.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

2. Постановка задачи

Состояние пациента задано вектором из $n = 38$ признаков в пространстве $\mathcal{X} = \prod_{i=1}^{n} M_i$, разбитых на три группы: 1) клинико-анамнестические (12 признаков); 2) лабораторно-биохимические (8 непрерывных признаков, нормированных к $M_i = [0, 1]$ методом min-max с порогом неразличимости $\varepsilon_i > 0$); 3) ИГХ и молекулярные маркеры биоптата (18 бинарных признаков $M_i = \{0, 1\}$).

С учётом пропусков ($\perp$) введено пространство $\widetilde{\mathcal{X}} = \prod_{i=1}^{n}(M_i \cup \{\perp\})$. Множество диагнозов $D = \{d_1, \ldots, d_m\}$ ($m = 5$) включает: d_1 — ДВККЛ, d_2 — фолликулярную, d_3 — мантийноклеточную, d_4 — маргинальной зоны, d_5 — периферическую Т-клеточную лимфому. Пусть $\mathfrak{S}_D$ — множество перестановок D, $\mathcal{R}$ и $\mathcal{R}_{\text{ban}}$ — множества нечётких правил и жёстких запретов в Prolog, а $\mathcal{E} = 2^{\mathcal{R}} \times 2^{\mathcal{R}_{\text{ban}}} \times 2^{\{1,\ldots,n\}}$ — пространство объяснений. Ищется отображение $f\colon \widetilde{\mathcal{X}} \to D \times [0, 1] \times \mathfrak{S}_D \times \mathcal{E}$, возвращающее основной диагноз d^*, уверенность $C(d^* \mid \widetilde{x})$, ранжированный по убыванию уверенности список диагнозов $\pi(\widetilde{x}) \in \mathfrak{S}_D$ и объяснение $e(\widetilde{x}) \in \mathcal{E}$ (множества активированных нечётких правил, сработавших правил-запретов и ключевых признаков).

3. Архитектура и методы гибридной диагностики

Отображение $f(\widetilde{x}) = \Phi(g(\widetilde{x}), \widetilde{x})$ объединяет нечёткую импутацию $g\colon \widetilde{\mathcal{X}} \to \mathcal{X}$ и классификатор Φ. Эталоны классов $v^{(j)} \in \mathcal{X}$ ($j = 1, \ldots, m$) вычисляются исключительно по обучающей выборке $\mathcal{D}_{\text{train}}$ [1, 2]: медианы для непрерывных и моды для дискретных признаков. По наблюдаемым признакам $I_{\text{obs}}(\widetilde{x})$ с учётом весов ω_i вычисляются расстояния $\rho_j(\widetilde{x})$ и степени близости $u_j(\widetilde{x})$ при $\sigma = 0{,}5$, используемые для восстановления пропущенных значений взвешенным средним.

Восстановленный вектор $x = g(\widetilde{x})$ подаётся на четыре базовых классификатора, обученных на $\mathcal{D}_{\text{train}}$: KNN ($k = 1$, $K_{\text{nn}} = 7$), Random Forest ($B = 200$ деревьев), полносвязную нейронную сеть FCNN с архитектурой 38–64–32–5 и наивный байесовский классификатор. Коллективное решение формируется по алгебраическому подходу [5].

Дополнительно используется база знаний на Prolog [4], включающая $|\mathcal{R}| = 45$ нечётких правил Мамдани и $|\mathcal{R}_{\text{ban}}| = 12$ правил-запретов. Оценка $p_5(d_j \mid x)$ определяется на основе нечёткого вывода и дефаззификации методом центра тяжести.

Свёртка объединяет модели с учётом фильтра $\gamma(d_j, \widetilde{x}) \in \{0, 1\}$ (где $\gamma = 0$ при срабатывании жёстких запретов из $\mathcal{R}_{\text{ban}}$):

$$S(d_j \mid \widetilde{x}) = \gamma(d_j, \widetilde{x}) \sum_{k=1}^{K} w_k p_k(d_j \mid x), \quad C(d_j \mid \widetilde{x}) = \frac{S(d_j \mid \widetilde{x})}{\sum_{l=1}^{m} S(d_l \mid \widetilde{x})}. \quad (1)$$

Основной диагноз d^* определяется как наиболее вероятный класс по итоговой оценке уверенности:

$$d^* = \arg\max_{d_j \in D} C(d_j \mid \widetilde{x}), \quad (2)$$

а ранжированный список $\pi(\widetilde{x})$ формируется упорядочиванием элементов множества D по невозрастанию $C(d_j \mid \widetilde{x})$. Веса w_k найдены методом SLSQP ($w_1 = 0{,}12$, $w_2 = 0{,}15$, $w_3 = 0{,}21$, $w_4 = 0{,}14$, $w_5 = 0{,}38$).

4. Экспериментальные результаты

Выборка включала 750 случаев пяти типов лимфом, верифицированных по критериям ВОЗ [1, 3]; исходная доля пропусков — 8,4 %. Разбиение train/val/test — 70/15/15 %. На тесте гибридная модель достигла Accuracy = 74,8 % (превысив лучший базовый метод на 6,1 п.п.) и макро-$F_1 = 0{,}75$. При MCAR-пропусках 10/20/30 % её F_1 составил 0,74/0,73/0,72 против 0,59 у Prolog при 30 % пропусков. Снижение F_1 на 4 % подтверждает устойчивость импутации и семантического фильтра.

5. Заключение

Гибридная СППР сочетает машинное обучение с прозрачностью правил Prolog и нечёткого вывода Мамдани. Двухконтурное согласование сохраняет $F_1 = 0{,}72$ при 30 % пропусков, выдавая врачу обоснованный диагноз, альтернативы и объяснение.

Список литературы

[1] М. В. Седых, “Формирование и валидация обучающих данных в задачах медицинской диагностики”, *Проблемы управления в социально-экономических и технических системах: Материалы XXI Международной научно-практической конференции. В 2-х томах*, Издательский центр «Наука», Саратов, 2025, 165–169.

[2] Ю. А. Бродская, “Повышение надежности распознавания образов при использовании рабочих словарей признаков с минимальными затратами ресурсов и времени”, *Устойчивость и процессы управления: Сб. трудов*, **2**, ред. Д. А. Овсянников, Л. А. Петросян, СПб., 2005, 751–759.

[3] Н. И. Стуклов, Г. И. Козинец, Н. Г. Тюрина, *Учебник по гематологии*, Практическая медицина, М., 2018, 336 с.

[4] М. Дженесерет, В. К. Чаудри, *Введение в логическое программирование*, пер. с англ. С. В. Минц, ДМК Пресс, М., 2022, 192 с.

[5] Ю. И. Журавлёв, В. В. Рязанов, О. В. Сенько, *Распознавание. Математические методы. Программная система. Практические применения*, Фазис, М., 2006, 176 с.

Статья поступила 9 октября 2026 г.

Hybrid Intelligent System for Differential Diagnosis of Non-Hodgkin Lymphomas

M. V. Sedykh, Y. A. Brodskaya

A hybrid system for non-Hodgkin lymphoma differential diagnosis is proposed, combining statistical and neural models with Prolog rules and Mamdani fuzzy inference. The novelty lies in a dual-loop consensus mechanism that uses semantic constraints to filter clinically invalid hypotheses. The macro-averaged F_1 score on the test set was 0.75.

Keywords: non-Hodgkin lymphoma, hybrid intelligent systems, fuzzy logic, logic programming, machine learning, explainable AI.

References

[1] M. V. Sedykh, "Formation and validation of training data in medical diagnostic problems", *Control Problems in Socio-Economic and Technical Systems: Proc. of the XXI Int. Scientific and Practical Conf. In 2 vols.*, Publishing Center "Nauka", Saratov, 2025, 165–169 (in Russian).

[2] Yu. A. Brodskaya, "Improving the reliability of pattern recognition when using working feature dictionaries with minimal resource and time costs", *Stability and Control Processes: Proc.*, **2**, eds. D. A. Ovsyannikov, L. A. Petrosyan, St. Petersburg, 2005, 751–759 (in Russian).

[3] N. I. Stuklov, G. I. Kozinets, N. G. Tyurina, *Textbook of Hematology*, Practical Medicine, Moscow, 2018 (in Russian), 336 pp.

[4] M. Genesereth, V. K. Chaudhri, *Introduction to Logic Programming*, translated from English by S. V. Mints, DMK Press, Moscow, 2022 (in Russian), 192 pp.

[5] Yu. I. Zhuravlev, V. V. Ryazanov, O. V. Senko, *Recognition. Mathematical Methods. Software System. Practical Applications*, Phasis, Moscow, 2006 (in Russian), 176 pp.

Received on October 9, 2026