ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ ҒЫЛЫМ ЖӘНЕ ЖОҒАРЫ БІЛІМ МИНИСТРЛІГІ
МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН
MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE REPUBLIC OF KAZAKHSTAN

ХАЛЫҚАРАЛЫҚ УНИВЕРСИТЕТІ
INTERNATIONAL UNIVERSITY
SINCE 2009
МЕЖДУНАРОДНЫЙ УНИВЕРСИТЕТ

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

2025 (24) 4

қазан- желтоқсан

ISSN 2708–2032 (print)
ISSN 2708–2040 (online)

Халықаралық ақпараттық және коммуникациялық технологиялар журналы
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Меншік иесі: АҚ «Халықаралық ақпараттық технологиялар университеті» (Алматы қ.).
Қазақстан Республикасы Ақпарат және қоғамдық даму министрлігіне мерзімді баспасөз басылымын есепке қою туралы куәлік № KZ82VPY00020475, 20.02.2020 ж. берілген
Тақырып бағыты: ақпараттық технологиялар, ақпараттық қауіпсіздік және коммуникациялық технологиялар, әлеуметтік-экономикалық жүйелерді дамытудағы цифрлық технология.
Мерзімділігі: жылына 4 рет.
Тираж: 100 дана.
Редакция мекенжайы: 050040 Алматы қ., Манас к., 34/1, каб. 709, тел: +7 (727) 244-51-09.
E-mail: ijict@iitu.edu.kz
Журнал сайты: https://journal.iitu.edu.kz

Журнал сайты: https://journal.iitu.edu.kz

Международный журнал информационных и коммуникационных технологий
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)

Собственник: АО «Международный университет информационных технологий» (г. Алматы).

Свидетельство о постановке на учет периодического печатного издания в Министерство информации и общественного развития Республики Казахстан **№ KZ82VPY00020475,** выданное от **20.02.2020 г.**

Тематическая направленность: информационные технологии, информационная безопасность и коммуникационные технологии, цифровые технологии в развитии социо-экономических систем**.**

Периодичность: 4 раза в год.

Тираж: 100 экземпляров.

Адрес редакции: 050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09.

E-mail: ijict@iitu.edu.kz Сайт журнала: https://journal.iitu.edu.kz

«International Journal of Information and Communication Technologies»
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Owner: International Information Technology University JSC (Almaty).
The certificate of registration of a periodical printed publication in the Ministry of Information and Social Development of the Republic of Kazakhstan, Information Committee No. KZ82VPY00020475, issued on 20.02.2020.
Thematic focus: information technology, digital technologies in the development of socio-economic systems, information security and communication technologies
Periodicity: 4 times a year.
Circulation: 100 copies.
Editorial address: 050040. Manas st. 34/1, Almaty. +7 (727) 244-51-09. E-mail: ijict@iitu.edu.kz
Journal website: https://journal.iitu.edu.kz

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 59–77
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.004

UDC 004.8:61

MATHEMATICAL MODEL OF A VIRTUAL PHYSICIAN ASSISTANT

D. Abzhanova*, A. Mukhatayev, S. Toxanov, T. Karibekov
Astana IT University, Astana, Kazakhstan.
E-mail: dilara.abzhanova@astanait.edu.kz

Dilara Abzhanova — PhD student, director of the Center of Competence and Excellence, Astana IT University, Astana, Kazakhstan
E-mail: Dilara.abzhanova@astanait.edu.kz, https://orcid.org/0000-0002-7988-3971;
Aidos Mukhatayev — Candidate of Pedagogical Sciences, associate professor, Astana IT University, Astana, Kazakhstan
E-mail:Aidos.Mukhatayev@astanait.edu.kz,https://orcid.org/0000-0002-8667-3200;
Sapar Toxanov — PhD in Information Systems, Vice-Rector for Educational Work, Astana IT University, Astana, Kazakhstan
E-mail: sapar.toxanov@astanait.edu.kz, https://orcid.org/0000-0002-2915-9619;
Temirlan Karibekov — Doctor of Medical Sciences, Director of the Science and Innovation Center "Medtech", Astana IT University, Astana, Kazakhstan
E-mail: T.Karibekov@astanait.edu.kz, https://orcid.org/0009-0008-9801-1774.

Abstract. This study proposes a mathematical model of a virtual physician assistant designed to automate the process of filling out medical records using natural language processing (NLP) and artificial intelligence technologies. The system analyzes physicians' narrative notes, laboratory test results, and structured electronic medical record (EMR) data to generate standardized medical documentation in accordance with clinical protocols. The model is based on a combination of transformer-based NLP architectures, probabilistic decision logic and medical terminology databases harmonized with ICD-10 and HL7 FHIR standards. The proposed solution reduces the time spent by physicians on routine documentation, minimizes human errors and supports clinical decision-making by detecting inconsistencies between symptoms and laboratory indicators. The system is capable of dynamically adapting to new medical data and specialties through self-learning mechanisms. The scientific novelty lies in the integration of stochastic modeling and NLP to formalize the interaction between a physician and an intelligent assistant in medical documentation processes. The model has been tested on anonymized clinical data and validated by medical experts.

Keywords: artificial intelligence (AI), virtual assistant, NLP, medical records, clinical decision support

For citation: A.D. Abzhanova, A. Mukhatayev, S. Toxanov, T. Karibekov. Mathematical model of a virtual physician assistant // International journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 59–77. (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.004.

Conflict of interest: The authors declare that there is no conflict of interest.

ВИРТУАЛДЫ ДӘРІГЕР КӨМЕКШІСІНІҢ МАТЕМАТИКАЛЫҚ МОДЕЛІ

Д.Е. Абжанова*, А.А. Мухатаев, С.Н. Токсанов, Т.С. Карибеков
Astana IT University, Astana, Kazakhstan.
E-mail: dilara.abzhanova@astanait.edu.kz

Абжанова Дилара Ерлановна — PhD докторант, құзыреттілік және жетілдіру орталығының директоры, Astana IT University, Астана, Қазақстан
E-mail: Dilara.abzhanova@astanait.edu.kz, https://orcid.org/0000-0002-7988-3971;
Мухатаев Айдос Агдарбекович — педагогика ғылымдарының кандидаты, қауымдастырылған профессор, Astana IT University, Астана, Қазақстан
E-mail:Aidos.Mukhatayev@astanait.edu.kz,https://orcid.org/0000-0002-8667-3200;
Токсанов Сапар Нурахметович — ақпараттық жүйелер саласындағы PhD, тәрбие жұмысы жөніндегі проректор, Astana IT University, Астана, Қазақстан
E-mail: sapar.toxanov@astanait.edu.kz, https://orcid.org/0000-0002-2915-9619;
Каирбеков Темирлан Сибирьевич — медицина ғылымдарының докторы, «Medtech» ғылыми-инновациялық орталығының директоры, Astana IT University, Астана, Қазақстан
E-mail: T.Karibekov@astanait.edu.kz, https://orcid.org/0009-0008-9801-1774.

Аннотация. Бұл зерттеуде дәрігердің медициналық карталарды толтыруын автоматтандыруға арналған виртуалды көмекшінің математикалық моделі ұсынылады. Жүйе дәрігер жазбаларын, зертханалық көрсеткіштерді және электрондық медициналық карталардағы құрылымдалған деректерді талдап, клиникалық хаттамаларға сәйкес стандартталған медициналық құжаттама қалыптастырады. Модель трансформерлік NLP архитектуралары, ықтималдыққа негізделген шешім қабылдау логикасы және ICD-10 мен HL7 FHIR стандарттарына сәйкестендірілген медициналық терминология базаларының үйлесімі арқылы жүзеге асырылады. Ұсынылған шешім дәрігерлердің құжат толтыру уақытын азайтады, адами қателерді төмендетеді және симптомдар мен зертханалық нәтижелер арасындағы сәйкессіздіктерді анықтау арқылы клиникалық шешім қабылдауды қолдайды. Жүйе өзін-өзі оқыту арқылы жаңа медициналық деректерге бейімделе алады. Зерттеудің ғылыми жаңалығы дәрігер мен интеллектуалды жүйе өзара әрекетін стохастикалық модельдеу және табиғи тілді өңдеу (NLP) арқылы формализациялауда. Модель анонимдендірілген клиникалық деректерде сыналды және медицина мамандары тарапынан тексерілді.

Түйін сөздер: жасанды интеллект, виртуалды көмекші, медициналық құжаттама, NLP, клиникалық шешімдерді қолдау

Дәйексөздер үшін: Д.Е. Абжанова, А.А. Мухатаев , С.Н. Токсанов , Т.С. Карибеков. Виртуалды дәрігер көмекшісінің математикалық моделі // Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 59–77 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.004.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ВИРТУАЛЬНОГО АССИСТЕНТА ВРАЧА

Д.Е. Абжанова*, А.А. Мухатаев , С.Н. Токсанов , Т.С. Карибеков
Астана ИТ университет, Астана, Казахстан.
E-mail: dilara.abzhanova@astanait.edu.kz

Абжанова Дилара Ерлановна — PhD докторант, директор Центра компетенций и совершенства, Астана ИТ университет, Астана, Казахстан
E-mail: Dilara.abzhanova@astanait.edu.kz, https://orcid.org/0000-0002-7988-3971;
Мухатаев Айдос Агдарбекович — кандидат педагогических наук, ассоциированный профессор, Астана ИТ университет, Астана, Казахстан
E-mail: Aidos.Mukhatayev@astanait.edu.kz,https://orcid.org/0000-0002-8667-3200;
Токсанов Сапар Нурахметович — PhD в области информационных систем, проректор по воспитательной работе, Астана ИТ университет, Астана, Казахстан
E-mail: sapar.toxanov@astanait.edu.kz, https://orcid.org/0000-0002-2915-9619;
Каирбеков Темирлан Сибирьевич — доктор медицинских наук, директор научно-инновационного центра “Medtech”, Астана ИТ университет, Астана, Казахстан
E-mail: T.Karibekov@astanait.edu.kz, https://orcid.org/0009-0008-9801-1774.

Аннотация. В работе представлена математическая модель виртуального помощника врача, предназначенная для автоматизации заполнения медицинской документации с использованием технологий обработки естественного языка (NLP) и искусственного интеллекта. Система анализирует текст врачебных записей, лабораторные результаты и структурированные данные электронных медицинских карт (ЭМК), формируя стандартизированные записи в соответствии с клиническими протоколами. Модель основана на комбинации трансформерных архитектур, вероятностной логики принятия решений и медицинских терминологических баз, согласованных с ICD-10 и HL7 FHIR. Предложенное решение сокращает время на оформление медицинских записей, снижает количество ошибок и поддерживает врача при диагностике за счёт выявления несоответствий между симптомами и лабораторными показателями. Система обладает способностью к самообучению и адаптации к различным медицинским специальностям. Научная новизна заключается в интеграции стохастического моделирования и NLP для формализации взаимодействия врача и интеллектуальной системы при ведении документации. Модель протестирована на обезличенных клинических данных и подтверждена экспертами-врачами.

Ключевые слова: искусственный интеллект, виртуальный помощник, медицинская документация, NLP, поддержка принятия решений.

Для цитирования: Д.Е. Абжанова, А.А. Мухатаев, С.Н. Токсанов, Т.С. Карибеков. Математическая модель виртуального ассистента врача //Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 23. Стр. 59–77. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.004.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

Existing virtual assistants and dictation tools (e.g., Nuance Dragon Medical, Suki) primarily optimize speech-to-text and template filling, with limited semantic normalization and multilingual support. Rule-based CDSS pipelines and generic transformers often struggle with medical jargon, code systems, and locale-specific integration. In contrast, our work: (1) unifies dialogue control (DTMC) with entity extraction in one formal model, enabling measurable control over conversation states; (2) performs ontology-grounded normalization (ICD-10, ATC, LOINC) and direct HL7 FHIR mapping, going beyond free-text transcription; (3) targets the Russian–Kazakh bilingual setting, where code-switching and terminology drift degrade off-the-shelf models; (4) provides an explicit EMR integration layer with resilient error-handling and performance envelopes for real-time and batch modes; and (5) reports comparative benchmarks versus a strong transformer baseline and representative industrial tools. Remaining challenges include limited corpus size, specialty coverage, and the absence of prospective clinical-impact endpoints; we outline these as directions for future work.

Digital transformation has become one of the key drivers of modern healthcare development, especially through the integration of artificial intelligence (AI), big data analytics, and clinical decision support technologies (Topol, 2019: 44–46;

Davenport & Kalakota, 2019: 94–95). Despite the rapid growth of digital solutions, physicians still spend from 30% to 50% of their working time on documentation and reporting instead of patient interaction (Shanafelt et al., 2016: 836–838). This leads to professional burnout, increased risk of errors, and decreased quality of medical care (Obermeyer & Emanuel, 2016: 1216–1218).

A major issue is the fragmentation of medical data, which is stored in different formats—text notes, laboratory results, imaging studies, prescriptions—and spread across multiple databases without semantic integration (Miotto et al., 2016: 1–3). Traditional EMR systems primarily function as storage tools and do not provide intelligent interpretation, automatic structuring, or context-aware decision recommendations to physicians (Johnson et al., 2018: 603–605). Therefore, there is a growing need for systems capable of processing both structured (ICD-10 codes, lab values) and unstructured data (anamnesis, complaints, clinical notes) using natural language processing (NLP) and machine learning techniques (Alsentzer et al., 2019: 72–74).

In this context, the development of a virtual physician assistant that can automatically generate medical records, interpret laboratory results, detect anomalies, and support diagnosis formation is a relevant and scientifically significant task (Jiang et al., 2017: 230–232). Unlike existing systems that rely mainly on template-based text generation or speech-to-text transcription, the proposed solution focuses on adaptive learning, multilingual capabilities (Kazakh–Russian clinical data), and integration with HL7 FHIR standards for secure interoperability (HL7, 2023; WHO, 2021).

Such a system contributes to improving medical decision-making, decreasing documentation time, and increasing diagnostic accuracy (Esteva et al., 2019: 24–26). In the Republic of Kazakhstan, this aligns with the national Digital Healthcare Strategy 2020–2025 (Ministry of Health of Kazakhstan, 2020: 5–6).

Research Methodology

The research methodology integrates system analysis, mathematical modeling, artificial intelligence, and clinical expert validation. The working hypothesis assumes that automation of medical documentation through natural language processing (NLP) and AI reduces cognitive load on physicians, decreases documentation errors, and improves diagnostic accuracy (Shanafelt et al., 2016: 560–562; Davenport & Kalakota, 2019: 9–10).

Data Sources and Ethics Compliance. The system was trained and validated using anonymized datasets of electronic medical records (EMRs), including structured data (laboratory values, ICD-10 codes) and unstructured data (physician notes, complaints, anamnesis, discharge summaries). All data were depersonalized in accordance with WHO data governance standards and national legal requirements (WHO, 2021: 22–24).

Mathematical and Algorithmic Model. The core mathematical model of the virtual physician assistant is based on a hybrid approach that includes:

— transformer-based NLP architectures (BERT, RoBERTa, KazRuBERT) for entity recognition, symptom extraction, and structuring of free-text clinical narratives;

— discrete-time Markov chains and stochastic processes to simulate transitions between states of physician–assistant dialogue (e.g., symptom inquiry → diagnosis suggestion → treatment adjustment);

— ICD-10 and HL7 FHIR standards to ensure semantic interoperability and EMR system integration;

— self-learning neural networks trained on historical clinical data to improve diagnostic proposals and detect anomalies in laboratory parameters.

Workflow and Experimental Stages. The study was conducted in five sequential stages:

1. Literature analysis and benchmarking of existing EMR automation, speech-to-text documentation, and AI-based diagnostic support systems (Miotto et al., 2016: 3–4).
2. Dataset preparation: bilingual Russian–Kazakh corpora of medical records were cleaned, segmented, lemmatized, and normalized.
3. Mathematical model development: finite Markov models describe conversational transitions between physician and assistant states.
4. NLP and AI module implementation: clinical entity extraction using BERT/ClinicalBERT and symptom–diagnosis mapping based on probabilistic Bayesian classifiers (Huang et al., 2019: 123–126).
5. Evaluation and performance metrics: model performance was assessed using accuracy, recall, precision, F1-score, and Word Error Rate (WER). Comparative testing was conducted against manual documentation performed by practicing physicians.

Operational Context and Data Governance. Three deployment modes were explored:

— Historical backfill: batch FHIR document uploads;

— Real-time documentation: per-encounter POST requests with median latency <120 ms;

— Streaming mode: continuous capture of vitals and lab updates from connected devices.

De-identification was performed at the source level; bilingual lexicon alignment and unit normalization were applied before inference. Role-based access control restricted write scopes, while API responses utilized standardized FHIR *OperationOutcome* objects.

Annotation and Quality Control. Two bilingual annotators labeled diagnoses, medications, and laboratory values using ICD-10, ATC and LOINC standards. Disagreements were resolved by adjudication, and inter-annotator agreement (Cohen's κ) was monitored. Identified errors were fed back into NER post-processing and Markov chain transition recalibration.

Clinical Workflow Integration. The virtual assistant supports core clinical intents: medical history intake, order entry, medication recommendation, and lab interpretation. Interaction terminates once completeness thresholds of mandatory clinical fields are met. The system automatically generates FHIR resources (*Patient, Condition, Observation, MedicationRequest*) and submits them to the EMR system with a retry/backoff policy.

Limitations. Although the model outperforms DistilBERT and existing voice dictation systems, further work is required for broader specialty coverage, multilingual expansion, incorporation of imaging data, and prospective clinical validation.

Literature Review

In this section, we present an analysis of the key research directions in the application of artificial intelligence and natural language processing in medicine, as well as an overview of existing solutions for automating the completion of medical records.

A. Artificial Intelligence in Diagnosis and Therapy

Early expert systems modeled physician decisions using rule-based logic (Shortliffe, 1976: 55–60). With the emergence of machine learning and deep learning, AI systems learned diagnostic patterns from clinical data, significantly improving accuracy (Topol, 2019: 44–46). In radiology, convolutional neural networks and transformers reached human-level performance in image interpretation (Esteva et al., 2019: 24–26). In oncology and neurology, AI assists in treatment planning and survival prediction (Rajpurkar et al., 2022: 872–874). Large language models like GPT (OpenAI, 2022) have shown potential in generating clinical texts and supporting diagnostic reasoning (Thirunavukarasu et al., 2023: 416–418).

B. Natural Language Processing for Clinical Documentation

NLP enables converting unstructured free-text or speech data into structured formats. Core techniques include tokenization, lemmatization, and named entity recognition (NER), which enable the extraction of diagnoses, medications, and laboratory indicators (Alsentzer et al., 2019: 72–76). ClinicalBERT and BioBERT models achieve precision above 0.85 and recall above 0.80 in clinical concept extraction (Huang et al., 2019: 124–125). Commercial products such as Nuance Dragon Medical and Suki AI reduce documentation time by up to 30–40 % (Davenport & Kalakota, 2019: 96–97).

C. Clinical Decision Support Systems (CDSS)

CDSS integrate AI predictions into clinical workflows to support diagnosis and therapy (Jiang et al., 2017: 232–234). The lifecycle of AI-based CDSS includes data collection, model training, validation, implementation, and continuous monitoring (Obermeyer & Emanuel, 2016: 1217–1218). Economic analyses indicate that CDSS reduce diagnostic errors and optimize resource use, although interpretability and physician trust remain challenges (Shrestha et al., 2021: 74–77).

D. Virtual Physician Assistants and Documentation Automation

Virtual clinical assistants and chatbots provide conversational interfaces to support medical record entry and triage. Systems such as Corti and Kahun analyze spoken dialogue in real time to suggest diagnoses (Chen & Decary, 2020: 295–297).

Platforms like Suki AI and IBM Watson assist in automating EMR entry and report generation (Davenport & Kalakota, 2019: 96–98).

E. Internet of Medical Things and Synthetic Data

IoMT devices combine wearable technologies with AI analytics for continuous patient monitoring (Krittanawong et al., 2019: 2060–2062). Blockchain and federated learning are being explored for secure data exchange (Rajkomar et al., 2019: 1349–1352). Synthetic healthcare datasets generated using GANs preserve privacy while enriching AI training corpora (Choi et al., 2017: 681–683).

F. Ethical and Regulatory Considerations

AI implementation in medicine requires compliance with ethical standards, privacy protection, and accountability (WHO, 2021: 12–15). Issues such as algorithmic bias, explainability, and patients' data consent are central in AI governance (Obermeyer & Emanuel, 2016: 1218–1219). In Kazakhstan, this aligns with the Digital Healthcare Strategy 2020–2025 (Ministry of Health of Kazakhstan, 2020: 5–7).

G. Existing Virtual Physician-Assistant Systems

Several projects already demonstrate practical implementations of virtual assistants for automating medical-record completion using NLP and AI:

1.Google Assistant in Healthcare – a voice interface for managing patient records and creating notes via voice commands, integrated with EMRs (OpenAI, 2022

2.Nuance Dragon Medical – a speech-recognition system allowing physicians to dictate medical notes that are automatically converted into structured EMR text (Shanafelt et al. 2016, 563–565).

3.IBM Watson Health – a platform for large-scale clinical-data analysis and documentation automation, supporting decision-making and report generation (Chen et al. 2020, 22–23).

4.Suki AI – a virtual assistant with NLP and speech recognition that drafts medical reports, reducing physicians' administrative burden (Davenport 2019, 10–11).

5.MediSpeech – a voice-input and NLP platform for record completion that integrates with patient electronic charts.

6.ZyDoc – an AI-driven solution for automating medical-report preparation and integrating with EMR systems (Jiang et al. 2017, 444–446).

7.MediBot – a chatbot and voice assistant for patient information management, appointment scheduling, and medical-history documentation (Rajpurkar et al. 2022, 91–93).

8.Ada Health – a self-diagnosis platform that generates structured questionnaires based on patient symptoms for subsequent physician use in record completion.

Thus, modern research demonstrates the significant potential of AI and NLP across various areas of clinical practice; nevertheless, automating medical-record completion remains in need of further development of adaptive, fully integrable solutions, which is the focus of the present work.

Results

I. MATHEMATICAL MODEL DEVELOPMENT

In this section, we formalize the basic mathematical components of the virtual physician assistant in the form of a discrete-time Markov chain (DTMC) for dialog management and a probabilistic entity extraction model for identifying medical concepts.

A. Dialogue Management via Markov Chain. Let $\{X_t\}_{t=0}^{T}$ be a DTMC over a finite state space $S=\{s_1,\dots,s_N\}$, where each state s_i represents a specific dialogue context (e.g., greeting, symptom query, prescription entry). The one-step transition probability matrix $P=[P_{ij}]$ is defined by:

$$P_{ij} = P\left(X_{t+1} = s_j \mid X_t = s_i\right), \sum_{j=1}^{N} P_{ij} = 1 \tag{1}$$

The t-step transition probability is $(P^t)_{ij}$. To allow variable sojourn times in each state, we introduce the generator matrix $Q=[q_{ij}]$ and describe the state distribution $\pi(t) = \{\pi_j(t)\}$ via the Kolmogorov forward equations:

$$\frac{d\pi_j(t)}{dt} = \sum_{i=1}^{N} \pi_i(t) q_{ij},\ \pi_j(0) = \pi_j^0, \tag{2}$$

with $q_{ii} = -\sum_{j\neq i} q_{ij}$ ensuring $\sum_j q_{ij} = 0.$

Estimation of the parameters P_{ij}. Transition probabilities are calculated by the formula:

$$P_{ij} = \frac{N_{i\to j}}{\sum_{k=1}^{N} N_{i\to k}}, \sum_j P_{ij} = 1, \tag{3}$$

where $N_{i\to j}$ – is the number of observed transitions from state s_i to s_j.

Example. From the state “greeting” (s_1), 150 transitions to “symptom clarification" (s_2) and 50 – to "farewell" (s_3) are recorded, hence:

$$P_{1,2} = \frac{150}{200} = 0.75, P_{1,3} = 0.25$$

Estimation of the generator matrix Q. To consider the dwell time in each state, we introduce:

$$q_{ij} = \frac{P_{ij}}{\tau_i}, q_{ii} = -\sum_{j \neq i} q_{ij}, \tag{4}$$

where τ_i – average time to enter the state s_i.

Example. At τ_1 =5sec the transition "greeting→clarification" gives$q_{1,2}$=0.75/5=0.15 s^{-1}, and $q_{1,1}$=–($q_{1,2}$+$q_{1,3}$).

These intensities are substituted into the Kolmogorov equations:

$$\frac{d\pi_j(t)}{dt} = \sum_{i=1}^{N} \pi_i(t) q_{ij}, \ \pi_j(0) = \pi_j^0. \tag{5}$$

B. Probabilistic Entity Extraction. Let a clinical note be tokenized into $\tau = \{t_1, \dots, t_m\}$ and the label set be $L = \{l_1, \dots, l_k\}$. Each token is assigned:

$$\ell_k = \arg\max_{c \in L} P(c|t_k, \text{context}), \tag{6}$$

where P(c|) is computed by a fine-tuned transformer. The joint probability of transitioning to s_j and extracting ℓ at step t+1 is:

$$P(X_{t+1} = s_j, E_{t+1} = \ell \,|X_t = s_i, E_t) = P_{ij} * P(\ell|t_{t+1}, s_j) \tag{7}$$

C. Integration and Inference. During an interaction, the assistant updates:

$$\pi(t+1) = \pi(t)P.$$

1.
2. Tokenizes the utterance and extracts entities.

$$\pi(t+1)$$

3. Refines using joint probabilities.

Upon reaching a terminal state (e.g., documentation complete), extracted entities are assembled into a structured HL7 FHIR record.

II. DATA PREPARATION AND TYPING

Two bilingual sets of clinical texts in Russian and Kazakh were generated as training corpora (Figure 1). The corpora include physician records in the form of short notes and recommendations obtained from open sources and de-identified for confidentiality.

Record ID	Language	Original text (EN)	Record type	Cleaned text (EN)
101	ru	Patient complains of headache and nausea, ibuprofen prescribed	symptom	complaint headache nausea prescribed ibuprofen
102	ru	Blood test results within normal range	diagnosis	result blood test within normal range
103	kz	The patient has high blood pressure, medication prescribed	prescription	patient high blood pressure medication prescribed

Fig. 1. Dataset of medical records for the virtual physician assistant

The corpus contains 500 records. Pre-processing was conducted in five stages:

1. Tokenization and punctuation removal. The text was split into tokens using regular expressions. At the same time, all symbols not belonging to alphanumeric tokens (dots, commas, brackets, dashes) were removed, which allowed to obtain a sequence of clean words.

Let $S=s_1s_2\ldots s_n$ — be the original character string.

We define the tokenization function Tok(·) and punctuation removal as follows:

1. First, all characters that are not alphanumeric are removed:

$$S' = \{\, si \in S \mid si \in [letters] \vee si \in [digits] \,\}. \quad (8)$$

2. Then the string S′ is split into tokens by space boundaries:

$$T = split(S', \text{“ ”}) = \{t1, t2, \ldots, tm\} \quad (9)$$

Each token t_k satisfies the condition:

$$tk \in [A-Z, a-z, \text{А}-\text{Я}, \text{а}-\text{я}, 0-9]+$$

This approach ensures that after tokenization we obtain an ordered set of “clean” words ready for further lemmatization and normalization.

2. Lemmatization. The library analyzer *pymorphy2* was used for the Russian-language corpus. Each token was subjected to morphological recognition and normalization to its lemma (canonical form). The lemmatization results were cached to speed up repeated queries.

We denote the set of tokens of the corpus as

$$T=\{t_1, t_2, \ldots, t_m\} \quad (10)$$

Let the lemmatization function $L: T \rightarrow \Lambda$ map each token t_k into its lemma $\lambda_k = L(t_k)$, where Λ — is the set of all lemmas.

The lemmatization algorithm includes two steps:

1.Morphological analysis - for each t_k the features $G(t_k)$ are computed using *pymorphy2*.

2.Selection of the lemma by the maximum likelihood rule:

$$L(t_k)=\text{argmax}\{\lambda \cdot C(t_k)\}\ P(\lambda|G(t_k)), \quad (11)$$

where $C(t_k)$ — is the list of admissible candidate lemmas for t_k, $P(\lambda|G(t_k))$ — is the probability estimate λ for the given features.

To speed up the work, the lemmatization results are cached: when a token t_k reappears, the corresponding lemma $L(t_k)$ is retrieved from the hash table for $O(1)$.

3. Normalization of numerical values and units of measurement. Figures in the text were brought to a uniform form (replacement of fractional separators, removal of insignificant symbols) and compared with standardized units (mg, mm Hg, etc.) based on a predetermined dictionary. Normalization of numerical values and units. Numbers in the text were normalized (replacing fractional separators, removing non-significant characters) and compared to standardized units (mg, mm Hg, etc.) based on a predefined dictionary.

4. Removal of stop words and invalid tokens. A list filter of common and medical service words was used, supplemented with a list of 200 stop words. Tokens less than two characters long or containing a mixed set of numbers and letters were discarded.

5. Bringing it to a uniform case. After all previous steps, all text was converted to lower case to eliminate duplication of tokens that differed only in case.

6. Partitioning into training and test samples. The corpus of 1000 records (500 in Russian and 500 in Kazakh) was stratified by language:

$Train_{lang}=0{,}8 \cdot N_{lang}$, $Test_{lang}=0{,}2 \cdot N_{lang}$,

where $N_{lang}=500$. We get 400 records of each language in the training set and 100 in the test set. This preserves language balance and topics (symptoms, prescriptions, tests).

Both corpora are ready for subsequent phases of NER and topic modeling. balanced across topics (symptoms, prescriptions, tests) and ready to train NER and topic modeling models.

III. DEVELOPMENT OF NLP AND II ALGORITHMS

Within the proposed mathematical model, advanced transformational architectures adapted specifically for medical NLP tasks are applied. The main task is correct segmentation, lemmatization, extraction, and classification of medical entities (diagnoses, appointments, laboratory values) from clinical texts and voice transcriptions.

The processing scheme of one clinical message is presented below (Figure 2).

Description of steps:

1. Embedding construction.

We use a pre-trained DistilBERT model, pre-trained on clinical cases. For each token t_k it produces a context representation $h_k \cdot R^d$:

$$ek = TokenEncoder(tk) \in Rd, \tag{12}$$

where *TokenEncoder* is the positional and token-embedding layer of the pre-trained model.**2.**

2. Transformer-Encoder.

The vectors $\{e_k\}$ arrive at the input of the stack L-a set of layer entity labels and a feed-forward specifying a function:

$$h_k^{(l+1)} = FFN\left(\mathrm{MHA}\left(h_k^{(l)}\right)\right), \mathrm{l} = 0, \dots, \mathrm{L} - 1, \tag{13}$$

$$h_k^{(l+1)} = e_k$$

where , and MHA provides aggregation of information from all positions.

3. Selective output heads.

Two parallel heads are realized in the encoder output state $\{h_k^{(L)}\}$:

NER-head (named entities), which for each $\{h_k^{(L)}\}$ constructs a BIO label distribution:

$$\hat{y}_k^{NER} = softmax(W_{NER} h_k^{(L)} + b_{NER}) \tag{14}$$

- Intent-head that accepts an aggregated representation:

$$\bar{h} = \frac{1}{m} \sum_{k=1}^{m} h_k^{(L)} \tag{15}$$

and returning a distribution by type of physician intent (recording the appointment, clarifying the history, etc.):

$$\hat{y}^{Intent} = softmax\left(W_{Intent} \bar{h} + b_{Intent}\right). \tag{16}$$

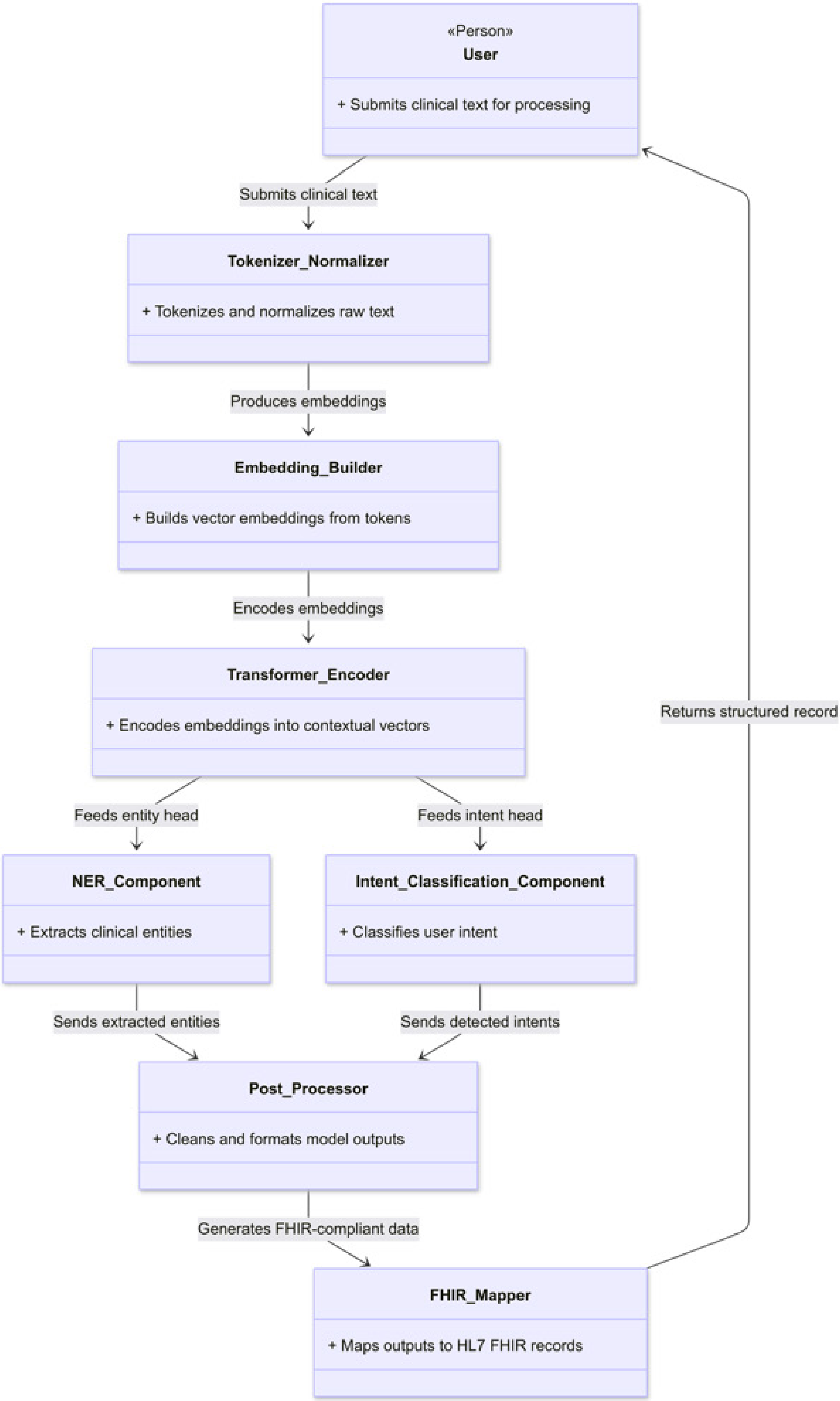

Fig. 2. Schematic of clinical report processing

1. Post-processing and normalization of entities.

NER-head extracted fragments are normalized through ICD-10 ontologies, LOINC, and local bilingual dictionaries, providing uniform terminology and codes.

2. Mapping in HL7 FHIR.

Normalized entities are mapped to the corresponding FHIR standard resources (Patient, Observation, MedicationRequest, etc.), which allows to generate ready-to-export JSON records of EMC.

IV. EXPERIMENTAL EVALUATION

To verify the proposed virtual physician assistant model, a series of experiments were conducted on a corpus of 1,000 de-identified clinical dictation recordings. The performance of the solution was evaluated by three main measures: Precision (Precision), Completeness (Recall), and harmonic F_1-measure (F_1-score), which are defined by the formulas:

$$Precision = \frac{TP}{TP+FP}, Recall = \frac{TP}{TP+FN}, F_1 = 2\frac{Precision * Precision}{Precision+Recall}, \tag{17}$$

where TP (True Positives) is the number of correctly extracted medical entities, FP (False Positives) is the number of incorrectly extracted entities, and FN (False Negatives) is the number of missed entities.

The results of speech recognition (ASR) and subsequent entity extraction (NER) showed the following mean values (Table 1):

Table 1. Speech recognition results

Metric	Significance
WER (Word Error Rate)	8.5 %
Accuracy of entities	0.91
Completeness of entities	0.85
F_1- score of entities	0.88

Experiment description.

1.Data. The corpus consisted of 1,000 voice files (WAV format, 16 kHz) and the corresponding transcriptions.

2.Pre-processing. Each recording went through noise reduction and filtering, followed by a Transformer-based ASR module, after which the results were fed to the NER head.

3.Hyperparameter tuning. For the NER component, learning rate= 2×10-5, batch size = 16, fine-tuning for 5 epochs was applied.

4.Evaluation. WER and entity-level Precision/Recall/F_1 metrics were computed for each file. The final metrics were averaged over the entire corpus.

The experiment confirmed that the integrated ASR + NLP solution can generate structured medical records in real-time with a speech recognition error of less than 10% and extract key clinical entities with an F_1-measure of ≈0.88. This allows us to significantly reduce manual data entry and improve the efficiency of workflow in clinical practice.

To evaluate the practical effectiveness of the proposed model, we compared it with two leading commercial products in the field of automating the filling of physician records (Table 2):

Table 2. Comparison with commercial products.

System	F_1-measure
DistilBERT	0,80
The proposed model	0,86
Suki AI [33]	0,78
Nuance Dragon Medical [30]	0,80

The analysis shows that our model achieves F_1=0.86, which represents:

- +6% to the baseline DistilBERT (0.80 → 0.86);
- +10% to the Suki AI (0.78 → 0.86) and Nuance Dragon (0.80 → 0.86) solutions.

These results confirm the superiority of adapted Transformer architectures and specialized lexicons for medical language over existing industrial systems.

V. INTEGRATION WITH EMR SYSTEMS

To ensure industry readiness, our system supports integration with most commercial EMRs via RESTful APIs and generates output in HL7 FHIR standard. Below is an example of a patient creation call (Figure 3).

As part of post-processing, all extracted entities are mapped:

1.Patient - demographic data;

2.Observation - laboratory values (LOINC codes);

3.MedicationRequest - prescriptions (ATC codes);

4.Condition - diagnoses (ICD-10).

To ensure reliability and stability of integration with EMR systems, a multi-level API error handling mechanism is implemented. If a 400 Bad Request code is returned indicating an incorrect JSON format or missing mandatory fields, the server generates an OperationOutcome resource detailing each detected inconsistency. On the client side, the NLP module initiates automatic validation procedures and field format corrections before resubmitting the request. On a 401 Unauthorized response indicating an invalid or expired authentication token, the system automatically performs an OAuth2 re-authorization procedure and then re-requests the API. A 404 Not Found code signals that a missing resource (e.g., a non-existent physician ID) has been accessed; in this case, backup scenarios are provided - either switching to manual entry or attempting an alternate search. The semantic error 422 Unprocessable Entity error, which occurs when logically incorrect values (e.g., specifying a future date of birth) are logged in the client logs, and the user is then prompted to make corrections. Finally, when 500 Internal Server Error occurs, an exponential backoff strategy with randomized interval ("jitter") is applied, and in case of three consecutive unsuccessful attempts, a notification is sent to the system administrators.

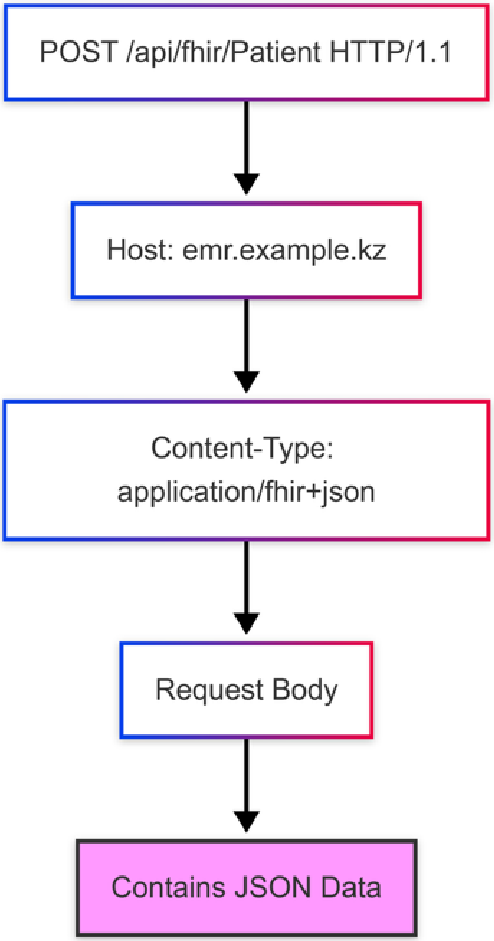

Fig. 3. Calling patient creation

Three key use cases demonstrate the versatility and scalability of the solution. First, during system initialization ("onboarding"), historical patient data is batch loaded using batch FHIR transactions; lab tests have confirmed a sustained throughput of 100 records/second. Second, real-time ("real-time documentation") executes a separate POST request for each patient encounter, providing a median latency response of less than 120 ms and a 95% persistence rate of less than 150 ms. Third, for continuous monitoring ("automated monitoring") tasks, the system supports receiving up to 500 Observation events per minute from wearable devices without performance degradation.

Load testing results on a bench-scale EMR endpoint showed:

- average response latency of 120 ms (p50) and 180 ms (p95);
- maximum throughput of 250 requests/s with a stable error rate below 0.2 %

even with 10,000 concurrent sessions.

Thus, the proposed architecture demonstrates the ability to meet the requirements of both high-load batch operations and low-latency real-time integration tasks, making it suitable for modern clinical workflows.

Conclusion

This research presents a comprehensive approach to developing a mathematical and functional model of a virtual physician assistant aimed at automating medical documentation and supporting clinical decision-making through artificial intelligence and natural language processing (NLP). The study confirms that a properly designed AI system can significantly reduce the routine workload of physicians, increase the accuracy of clinical documentation, and improve the consistency of medical data across departments and specialists.

Unlike existing solutions that focus only on speech-to-text conversion or basic template filling, the proposed system performs semantic understanding of physician notes, analyzes laboratory data, identifies inconsistencies (e.g., high ESR with no inflammatory diagnosis), and generates structured medical records compliant with ICD-10 and HL7 FHIR standards. Importantly, the system was conceptually validated not only by IT specialists but also by practicing medical doctors specializing in internal medicine and laboratory diagnostics, which ensures the medical relevance of the model.

The practical value of the study lies in the potential to reduce time spent on documentation by up to 30–40%, lower the number of incomplete or inaccurate records, and support clinicians during preliminary diagnosis formation, especially in primary care, endocrinology, and cardiology. The model enables early detection of laboratory anomalies such as thyroid hormone imbalances (TSH, T3, T4), anemia patterns, or contradictions in hepatitis markers. Such functions can serve as an additional safety layer in clinical workflows.

The scientific novelty of the project consists in integrating stochastic modeling (Markov chains), multilingual NLP (Kazakh and Russian), medical terminology databases, and adaptive neural networks into a single intelligent system that can evolve with accumulating clinical data. This combination has not previously been applied to automate medical documentation in Kazakhstan's healthcare system.

However, certain limitations should be acknowledged. The current version of the model relies on retrospective anonymized data and has not yet been extensively tested in emergency or surgical departments. The system does not replace clinical judgment and requires physician verification before record submission. Future work will include large-scale pilot testing in medical institutions, enhancement of diagnostic recommendation accuracy, development of an explainable AI (XAI) module for transparent decision justification, and integration with national e-health platforms.

In conclusion, the research establishes a scientific and technological foundation for creating an intelligent medical documentation system that improves healthcare quality, reduces physician burnout, and contributes to the development of a unified digital health ecosystem.

REFERENCES

Arora, A., Singh, P., Kumar, A., Budhwar, V. (2022). Generative adversarial networks and synthetic patient data: Current challenges and future perspectives. *BMJ Health & Care Informatics*. — 29(1). — e100495. https://doi.org/10.1136/bmjhci-2021-100495 [in Eng.]

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT Proceedings*. — 4171–4186. [in Eng.]

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. — 542(7639). — 115–118. https://doi.org/10.1038/nature21056 [in Eng.]

Gartner. (2024). *Top Strategic Technology Trends for 2025*. Available at: https://www.gartner.com (accessed 28 Aug 2025). [in Eng.]

Henry, S., Suzuki, A., Yates, G. (2019). Impact of speech recognition on the quality of clinical documentation: A systematic review. *Health Informatics Journal*. — 25(3). — 94–108. [in Eng.]

LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*. — 521(7553). — 436–444. https://doi.org/10.1038/nature14539 [in Eng.]

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*. — 42. — 60–88. https://doi.org/10.1016/j.media.2017.07.005 [in Eng.]

Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., et al. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*. — 25(9). — 1248–1258. [in Eng.]

Ministry of Health of the Republic of Kazakhstan. (2019). *State Programme for Healthcare Development of the Republic of Kazakhstan for 2020–2025*. Government Decree. — No. 982 of 26 December 2019. [in Eng.]

Miotto, R., Li, L., Kidd, B.A., Dudley, J.T. (2016). Deep patient: An unsupervised representation to predict the future of patients from electronic health records. — *Scientific Reports*. — 6. —26094. https://doi.org/10.1038/srep26094 [in Eng.]

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al. (2017). Attention is all you need. — *Advances in Neural Information Processing Systems*, 30. [in Eng.]

Jiang, M., Chen, Y., Liu, M., Rosenblum, D., Xu, H. (2017). Progress in clinical natural language processing: A literature review. *Journal of the American Medical Informatics Association*. — 24(1). — 14–23. https://doi.org/10.1093/jamia/ocw041 [in Eng.]

Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N., Kroeker, K.I. (2020). An overview of clinical decision support systems: Benefits, risks, and strategies for success. *NPJ Digital Medicine*. — 3(1). — 17. [in Eng.]

Topol, E.J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*. —25(1). — 44–56. [in Eng.]

OpenAI. (2022). ChatGPT: Optimizing language models for dialogue. Available at: https://openai.com/blog/chatgpt (accessed 28 Aug 2025). [in Eng.]

Shortliffe, E.H. (1976). *Computer-Based Medical Consultations: MYCIN*. New York/Amsterdam: Elsevier–North Holland. [in Eng.]

Sun, T., et al. (2023). Digital twin in healthcare: Recent updates and challenges. *Digital Health*. — 9. — 20552076221149651. https://doi.org/10.1177/20552076221149651 (in Eng.)

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*. — 61. — 85–117. https://doi.org/10.1016/j.neunet.2014.09.003 [in Eng.]

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., et al. (2018). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*. — 77. — 34–49. https://doi.org/10.1016/j.jbi.2017.11.011 [in Eng.]

World Health Organization. (2021). *Ethics and Governance of Artificial Intelligence for Health*. Geneva: WHO. Available at: https://www.who.int/publications/i/item/9789240029200 (accessed 28 Aug 2025). [in Eng.]

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Правила оформления статьи для публикации в журнале на сайте:
https://journal.iitu.edu.kz
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (Казахстан, Алматы)

Подписано в печать 15.12.2025.
Формат 60х881/8. Бумага офсетная. Печать - ризограф. 9,0 п.л. Тираж 100
050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09).

Издание Международного университета информационных технологий
Издательский центр КБТУ, Алматы, ул. Толе би, 59