Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI and Social Good

AI Diagnostic Assistant (AIDA): A Predictive Model for Diagnoses from Health Records in Clinical Decision Support Systems

Dmitry Umerenkov[1], **Alexandr Nesterov**[1], **Vladimir Shaposhnikov**[1,6], **Ruslan Abramov**[2], **Nikolay Romanenko**[3], **Vladimir Kokh**[4], **Marina Kirina**[5], **Anton Abrosimov**[5], **Dmitry V. Dylov**[1,6], **Ivan Oseledets**[1,6]
[1] AIRI,
[2] SberMedAI,
[3] Sberbank Artificial Intelligence Laboratory,
[4] Sberbank,
[5] Moscow Center for Innovative Technologies in Healthcare,
[6] Skoltech

Abstract

Clinical Decision Support Systems (CDSS) play an increasingly important role in medical diagnostics. We present AI Diagnostic Assistant (AIDA), a real-time predictive model designed to assist doctors in interpreting patient conditions while working in a CDSS. AIDA analyzes electronic health records (EHR), including medical history, laboratory results, and complaints, to suggest potential diagnoses from 95 common conditions before doctor makes final decision. The model acts as a verification and backup tool, ensuring that no critical details are overlooked. Trained on 1.5 million patient records and validated on a dataset curated by a panel of experts, AIDA proves trustworthy as a diagnosis-making assistant (87.7% accuracy compared to 91.7% accuracy among doctors).

Integrated into a megapolis-wide CDSS, AIDA has assisted doctors in making over 3 million real-world diagnoses to date.

1 Introduction

Information technologies are transforming healthcare, with Clinical Decision Support Systems (CDSS) playing a key role in improving diagnostic accuracy and reducing errors. Despite their importance, real-world adoption faces challenges such as model interpretability, integration with diverse data sources, and even sheer trust to physicians [Ledley and Lusted, 1991; Kline *et al.*, 2022].

One of the key decisions doctors make is determining the correct diagnosis for a patient's condition. However, this is a complex task, with misdiagnosis rates estimated to be as high as 30% [Liberman and Newman-Toker, 2018]. The challenge arises from the vast amount of patient data available, including unstructured clinical notes, structured laboratory results, imaging reports, and prior medical history. Integrating and interpreting this information efficiently remains a bottleneck in clinical practice. To assist with this challenge, we developed AI Diagnostic Assistant (AIDA), a machine learning model designed to predict ICD-10 diagnoses based on the full medical history of a patient. We employed ICD-10 [Organization, 1992] as a list of possible diagnoses, consistent with the Electronic Health Record (EHR) system used as the data source and as the integration endpoint.

We trained AIDA on 1.5 million patient records to predict ICD-10 codes of diagnoses, confirmed by laboratory or instrumental results and flagged as "final". AIDA achieved target 75% accuracy for the 95 most frequent diagnoses. For these codes, we created a testing dataset from real cases imported from the EHR data. This dataset was annotated by a panel of three experienced doctors, allowing us to compare the panel results with both the model's predictions and the original doctor assessments imported from the EHR. Unlike previous studies that focus on retrospective evaluation, AIDA has been integrated into a real-world CDSS, where it actively assists doctors in diagnostic decision-making.

In the first year of deployment in Moscow City's CDSS, AIDA has processed over 5 million diagnostic queries and provided over 3 million diagnostic recommendations, achieving a doctor agreement rate of 84.3%. A crucial aspect of AIDA's integration into clinical workflows is its acceptance by medical professionals. Throughout development, we gathered qualitative feedback from doctors to assess both the model's practical utility and its usability in a real-world setting. While we could not interfere with real-time patient-doctor interactions, post-factum surveys indicate that AIDA was particularly beneficial for young specialists, helping them feel more confident in their diagnostic decisions. This aligns with our broader goal of using AI to enhance medical decision-making and reduce diagnostic uncertainty, contributing to improved patient outcomes.

Below, we describe two approaches to this multilabel target task in detail. The first one is a fully multimodal system that uses specialized encoders for each data type to generate embeddings that are then processed by a recurrent neural network. The second one is a text-only model where all modalities are represented as texts and used as inputs to a transformer model with a long-range attention. We provide results

on the validation sets for both approaches, including the studies of the data impact and the effects of different preprocessing methods. We also tested these approaches on a specially curated testing dataset, which enabled us to compare doctors' accuracy with the model's across different subsets of cases. Finally, we discuss the practical challenges of integrating our model into a large-scale CDSS.

2 Related Work

Over the last decade, the field of machine learning in medicine has significantly expanded, shifting its focus from model development to model deployment using real-world data. For a comprehensive review, refer to Zhang et al. [Zhang *et al.*, 2022], which covers recent advances in AI-driven clinical decision support and diagnostic prediction.

The widespread adoption of electronic health records (EHRs) has enabled machine learning models to leverage large-scale multimodal patient data for various healthcare applications, including diagnosis prediction. Since Choi's pioneering work, which applied recurrent neural networks (RNNs) to process sequential EHR data for diagnosis prediction [Choi *et al.*, 2016], the field has evolved significantly. More recent approaches, such as BEHRT [Li *et al.*, 2020], introduced transformer architectures for this task, demonstrating improved performance over traditional sequence models. Additionally, SCOPE [Mukherjee *et al.*, 2023] challenged the assumption that deep learning always outperforms simpler models, demonstrating that logistic regression and random forests can achieve competitive results for diagnosis prediction.

A major challenge in leveraging EHRs for ML-based diagnosis is the large volume of unstructured clinical text. Extracting meaningful insights from such data requires advanced natural language processing (NLP) techniques. EHRs contain a substantial amount of unstructured clinical text, making natural language processing (NLP) a crucial component of AI-driven CDSS. Devlin et al. [Devlin *et al.*, 2018] demonstrated the power of transformer networks in NLP, particularly after masked language modeling (MLM) pretraining on large text corpora. However, the computational cost of transformers scales quadratically with sequence length, making them impractical for processing long-term patient histories in their original form. To mitigate this, Beltagy et al. [Beltagy *et al.*, 2020] introduced Longformer, a transformer variant that incorporates global attention and windowed local-context self-attention, significantly reducing computational complexity while maintaining long-range dependencies.

To leverage these advances, we explored two distinct modeling approaches:

1. **Longformer as a feature extractor:** We used Longformer embeddings for unstructured texts, which were then processed by an LSTM [Hochreiter and Schmidhuber, 1997] to sequentially model patient history.
2. **Longformer as an end-to-end model:** In this approach, the entire patient history was concatenated into a single text sequence and fed directly into the Longformer model.

Our work extends prior research by integrating AIDA into a real-world CDSS, differentiating it from retrospective studies. AIDA builds upon prior work, such as the TOP3 preliminary diagnosis prediction model [Blinov *et al.*, 2020], which provided top-3 probable diagnoses based on visit-time complaints. Unlike TOP3, which predicts only preliminary diagnoses, AIDA processes full patient history to determine the final confirmed diagnosis, making it clinically relevant.

3 Data

We used three datasets: pretraining, finetuning, and testing. The pretraining data were used to train a specialized Longformer model on a masked language modeling task. The finetuning dataset was used to train and validate the model for ICD-10 code prediction. The testing dataset, independently gathered from the same EHR system, was used to assess AIDA's performance against that of the medical experts.

3.1 Pretraining Dataset

The pretraining dataset was assembled using EHRs from several large medical clinics and one regional health records system. The dataset was used to compile a corpus that included patient visit information in the format `<visit text>:<ICD code and description, where available>`. Each patient's visits were concatenated in chronological order. Additional data, such as laboratory results or instrumental examination outcomes, were not included in the pretraining dataset.

3.2 Finetuning Dataset

The diagnostic model was trained on a finetuning dataset containing anonymized health records from 1.5 million patients in the Central Medical Information and Analysis System of Moscow City. Each record included demographics, doctor visits, laboratory results, various instrumental measurements. Patient visit records contained complaints, medical history, examination data, ICD-coded diagnoses, and recommendations. Laboratory results detailed test names, individual values, and reference ranges, while instrumental measurements included protocols and conclusions.

The dataset was split into 1.35 million patient histories for training and 133 thousands for validation. Multiple training records were extracted per patient for each final diagnosis, while validation patients had only one record based on their last diagnosis.

Historically, the target set of ICD diagnoses has evolved in the following three stages:

1. Initially, 256 most common diagnoses were selected.
2. Enlarged to 571 diagnoses, covering 95% of incoming cases, but excluding obstetrics, oncology, and checkups.
3. Filtered down to 363 diagnoses, removing those diagnoses not assigned by therapists, absent in clinical guidelines, or unsuitable as final diagnoses.

At each stage, the cases that did not belong to the selected subset of ICD-10 codes were removed from training and validation sets. To ensure comparability, we report performance separately for each target subset rather than across all stages.

Table 1 presents the distribution of final diagnoses in the finetuning dataset.

3.3 Testing Dataset

To evaluate AIDA, we constructed a specialized test dataset of 3,500 anonymized cases, independently exported from the same EHR system as the finetuning dataset. To prevent data leakage, cases were selected only after finalizing the training dataset. The test set maintained a natural distribution while ensuring that no cases overlapped with the training data, all diagnoses belonged to the predefined target set, and diagnoses were recorded after the training period. Medical experts further curated the set to reflect the real-world distribution.

Each test case was independently reviewed by three experienced medical experts, who provided final diagnoses based on full patient histories. Ground truth was established when at least two experts agreed. Their agreement could be summarized as follows:

- **Complete agreement** (3 experts): 67% (2,367 cases).
- **Partial agreement** (2/3 experts): 28% (965 cases).
- **No agreement** (all experts disagreed): 5% (168 cases).

Cases with partial agreement were considered more challenging than those with complete agreement. Below, we will show that both doctors and the model perform significantly worse in these cases.

4 Models

4.1 Longformer Models

Our first approach used a transformer-based method, leveraging pretrained large language models (LLMs) for EHR analysis. To ensure compatibility with the model, patient data was structured into text and processed using a Longformer model. Initially, we trained a BERT model on masked language modeling using unstructured clinical notes, a process taking two weeks on a Tesla K40 GPU. This model was then converted into a Longformer architecture with sliding window attention for efficient processing of sequences up to 8192 tokens. We further pretrained the Longformer on structured medical texts for three days on an NVIDIA V100 GPU before finetuning it on ICD-10 code prediction using patient histories. To construct patient histories, we manually selected relevant text fields from EHR forms for each patient visit in the training and validation sets. Instrumental reports were included in both short and long formats, while lab test results were added only if abnormal, as full lab data degraded performance. To improve representation, demographic data (age, gender) was appended at the beginning of each sequence and historical records were then concatenated chronologically (most recent first).

The maximum sequence length was an important hyperparameter, as it dictated how much historical patient data the model could process at once. We carefully tuned the token limit to retain the most relevant medical history while avoiding excessive truncation. Interestingly, including full laboratory results in textual form degraded model performance, as older medical history was truncated, reducing the model's ability to capture long-term patient trends.

AIDA was finetuned as a multi-class classifier, with a linear layer placed over the first token representation to produce the final diagnosis prediction. We also explored additional optimizations: adding time interval embeddings to account for gaps between visits, though this did not lead to measurable improvements. Additionally, physical observations such as height and weight were included only if explicitly recorded in text fields, as their overall impact on model accuracy remained negligible.

4.2 Multimodal Model

Besides the text-based approach, we developed a composite architecture that we refer to as Multimodal model that separately encodes medical events within a patient's history to address the challenge of representing diverse and complex EHR data. This model separately processes textual, laboratory, and categorical information before combining them into a unified patient history representation.

The textual modality includes patient complaints, medical history, clinical diagnoses, instrumental examination reports, and textual laboratory results. Because free-text fields contain rich clinical insights, each textual element is treated as an independent medical event, tokenized, and encoded using a pretrained medical transformer (as described in the Longformer model section). This approach ensures that semantic relationships in medical text are captured well, contributing to a more interpretable and structured input representation.

For laboratory data, we initially considered encoding test values as normalized real numbers and feeding them into an MLP encoder, but this approach failed to capture critical outliers, which often play a decisive role in clinical diagnostics. To address this, we applied structured discretization: a Histogram-based Outlier Score model [Goldstein and Dengel, 2012] identified anomalies, and a Birch clustering algorithm [Zhang *et al.*, 1997] segmented values into 3 to 30 sub-ranges, mapping results to discrete tokens that were combined into a one-hot vector for MLP processing. Categorical data, including demographics (age, gender), physician specialty, and ICD codes, were similarly encoded as one-hot vectors and processed via separate MLP encoders.

All encoded clinical events were then chronologically assembled into a structured sequence, with explicit positional encodings added to capture event timing and category type. The final sequence representation, containing up to 256 medical events per patient, was processed by a Bidirectional Long Short-Term Memory (BiLSTM) network for multi-class diagnosis classification.

As an additional improvement for the Multimodal model, we used a special asymmetric loss function [Ridnik *et al.*, 2021] designed to mitigate class imbalance in the training data (instead of binary cross-entropy). This function adjusts the weighting of positive and negative samples dynamically, reducing the impact of frequent diagnoses while ensuring better learning for rare cases.

$$L_+ = (1-p)^\gamma \log(p)$$

$$L_- = (p_m)^\gamma \log(1-p_m), \quad p_m = \max(p-m, 0),$$

$$L = -yL_+ - (1-y)L_-$$

ICD-10	Name	Total visits	Percentage	Cumulative
I11.9	Hypertensive heart disease without (congestive) heart failure	465,308	18.8%	18.8%
U07.1	COVID-19, virus identified	329,144	13.3%	32.1%
J06.9	Acute upper respiratory infection, unspecified	215,392	8.7%	40.8%
M42.1	Adult osteochondrosis of spine	183,511	7.41%	48.22%
I67.8	Other specified cerebrovascular diseases	83,534	3.38%	51.59%
I25.1	Atherosclerotic heart disease	55,703	2.25%	53.84%
E11.7	Non-insulin-dependent diabetes mellitus: With multiple complications	45,886	1.85%	55.7%
J45.8	Mixed asthma	43,162	1.74%	57.44%
J12.8	Other viral pneumonia	41,665	1.68%	59.13%
U07.2	COVID-19, virus not identified	36,171	1.46%	60.59%
I25.2	Old myocardial infarction	34,934	1.41%	62.0%
J04.1	Acute tracheitis	34,343	1.39%	63.39%
K29.3	Chronic superficial gastritis	29,472	1.19%	64.58%
I20.8	Other forms of angina pectoris	28,852	1.17%	65.74%
-	Other 348	847,839	34.26%	100.0%
	Total	2,474,916	100.0%	100.0%

Table 1: Distribution of diagnoses in the finetuning dataset.

excluded**Data type**	**Results ONLY on this data type**		**Results with this data type excluded**	
	top 1	**top 3**	**top 1**	**top 3**
Anamnesis	40.99%	60.88%	6.60%	5.50%
ICD code history	31.30%	49.41%	17.60%	12.30%
Complaints	26.80%	42.95%	4.75%	4.44%
Physical examination categorical data	23.56%	41.83%	2.27%	1.50%
Age and gender	22.22%	36.36%	2.89%	1.66%
"Full diagnosis" text field history	22.16%	37.69%	0.95%	0.58%
Physical examination numeric data	20.98%	31.99%	1.32%	0.81%
Doctor specialization	19.99%	34.34%	2.27%	1.21%
Laboratory results	19.23%	35.56%	1.96%	1.08%
gistology	14.75%	34.35%	0.71%	0.44%
Instrumental conclusions	12.58%	28.01%	0.75%	0.56%

Table 2: Multimodal model top 1 and top 3 accuracy results on different data types and on all the data except the select data types.

where γ and m are hyperparameters fine-tuned for optimal performance, and y represents the true class label, determining which component of the loss contributes to the final optimization objective. Additionally, at later experimental stages, we switched from Batch Normalization [Ioffe and Szegedy, 2015] to Layer Normalization [Ba *et al.*, 2016] to stabilize multimodal model training, leading to improved convergence and generalization across diverse patient records.

The Multimodal AIDA architecture integrates sequential and multimodal data, capturing a broader temporal context than traditional text-based methods. Time encoding further enhances learning of temporal relationships between medical events. Figure 1 shows the diagram of the multimodal model. Next, we primarily focus on the text-based version of AIDA, with comparisons to the multimodal approach to assess the impact of temporal modeling.

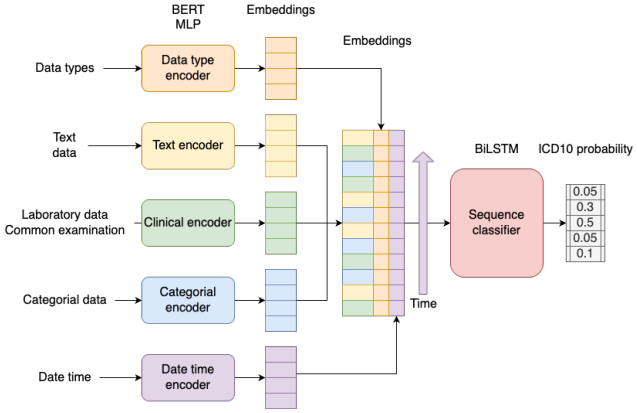

Figure 1: AIDA: Multimodal Model Architecture

5 Results

5.1 Feature Importance on Validation Set

During model development, we faced the challenge of processing redundant information in patient EHRs. Many critical details are manually entered by physicians into anamnesis and complaint fields, potentially making some structured data redundant. To assess the relative importance of different EHR modalities, we conducted ablation studies on the multimodal model, training it on specific data subsets and measuring performance changes when excluding different features.

Table 2 presents the results, showing that patient complaints, anamnesis, ICD code history, and demographics contribute most significantly to diagnostic accuracy. Surprisingly, laboratory results and instrumental conclusions had minimal impact, contradicting initial expectations that explicit diagnostic confirmations (*e.g.*, abnormal lab tests or imaging results) would strongly influence predictions. A possible explanation for this counterintuitive result is that physicians tend to manually summarize critical lab and imaging findings within the anamnesis fields. As a result, explicitly including this data as separate features provides little additional value to the model. This insight aligns with findings from the Longformer model experiments reported in Table 3, reinforcing the observation that structured lab and instrumental data may not significantly improve model performance when textual descriptions are already available.

5.2 AIDA Results on Validation Split of Finetuning Dataset

We conducted multiple experiments, summarized in Table 3, evaluating models on their ability to predict the correct diagnosis (hit@1) and whether it appeared in the top three predictions (hit@3). Experiments were performed on 265, 571, and 363 target classes, as detailed in the Data section. We performed three series of experiments with 265, 571 and 363 target classes as discussed in the Data section. The differing number of classes makes direct comparisons between series invalid, but comparisons within each series provide valuable insights. We established a logistic regression baseline, transforming textual data into numerical form using TF-IDF, which assigns importance-based weights to words for classification.

Our experiments indicate that both extending the maximum sequence length to provide additional context and incorporating additional data lead to improvements in metrics. However, it is important to note that the benefits from increased context diminish over time, and the gains from including explicit information about lab results outside reference values were minimal. These findings align closely with the results of additional data and lab results in SCOPE study [Mukherjee *et al.*, 2023] using logistic regression. This suggests that the findings are consistent with the task and the data rather than the specific model used for disease prediction. We experimented with two methods of fusing visit data in the multimodal model: transformers and recurrent neural networks. Early in the experiments, we decided to focus on bidirectional LSTMs. A general finding from our experiments is that both models perform equally well on validation data, with the multimodal model being marginally better but requiring much more complex data processing. Most importantly, the multimodal model requires retraining the embedding submodules. These considerations, along with potential implementation hurdles described further, led us to continue experiments on the testing dataset focusing on the Longformer model for CDSS integration.

For further testing and comparison with doctors, 95 ICD-10 codes where selected where the model accuracy was greater than 75%.

5.3 Results on Testing Dataset

To evaluate AIDA's performance, we compared its diagnostic accuracy with that of doctors on 3,332 clinical cases where the ground truth was established (95% of the dataset). Table 5 summarizes the results, with all accuracy differences statistically significant ($p < 0.05$, chi-square test). We also accounted for ICD-10 synonyms, where assigning a synonymous code was not considered an error (*e.g.*, I48.1 "Persistent atrial fibrillation" and I48.2 "Chronic atrial fibrillation" were treated as correct).

Overall, doctors achieved 90% accuracy (91.7% with synonyms), while AIDA reached 86.1% (87.7% with synonyms). In cases with complete expert agreement, accuracy was higher (doctors: 96.3%, AIDA: 93.5%, increasing to 97.1% and 94.1% with synonyms), indicating these were likely easier diagnoses. Conversely, in cases with partial expert agreement, accuracy dropped significantly (doctors: 78.6%, AIDA: 68.1%; with synonyms: 78.2% and 72%), reflecting higher diagnostic complexity.

When doctor and model diagnoses matched, accuracy was 98.5% (99% with synonyms), evidently suggesting straightforward cases. However, when they differed, doctor accuracy fell to 57.8% (64% with synonyms), while AIDA's was 39.3% (45.1% with synonyms), indicating these were some challenging clinical scenarios where model suggestions could warn for a more cautious decision-making.

Distribution of correct answers by doctors and AIDA:

- Both doctor and AIDA are correct: 77.9%
- Only doctor is correct: 12.1%
- Only AIDA is correct: 8.2%
- Both are incorrect: 1.8%

The accuracy lag between AIDA and doctors ranged from 2.8% to 18.9%, but the inclusion of synonyms notably improved the results, particularly in complex cases. The alignment between doctors and AIDA highlights its ability to identify cases with high diagnostic uncertainty. Despite slightly lower accuracy, AIDA adds clinical value by flagging difficult cases where human accuracy also drops (57.8% when disagreeing with AIDA). When AIDA suggests an alternative diagnosis, it is correct 39.3% of the time, meaning its input could boost overall diagnostic accuracy to 98.2%, if taken into account.

Rather than replacing doctors, AIDA serves as a "second opinion" system, helping to prevent errors in ambiguous cases. Its ability to highlight complex diagnoses further positions it as a decision-supporting tool. Despite its slightly

Model type	Input length	epochs	lr	BS	Data	hit@1	hit@3
256 classes							
Logistic regression	-	-	-	-	Visit, Report	53.92%	69.36%
Logistic regression	-	-	-	-	Visit, Report, Lab	55.28%	70.16%
Longformer	128	4	3e-5	96	Visit, Report	64.75%	82.73%
Longformer	256	4	3e-5	64	Visit, Report	67.72%	84.79%
Longformer	384	3	3e-5	48	Visit, Report	68.64%	85.52%
Longformer	512	3	3e-5	32	Visit, Report	68.90%	85.79%
Longformer	512	3	3e-5	32	Visit, Report, Lab	69.04%	85.93%
Longformer	512	3	3e-5	32	Visit, Report, Lab	69.03%	85.92%
Longformer	1024	3	3e-5	32	Visit, Report, Lab	69.27%	86.27%
Multimodal (Transformer)	-	5	5e-5	20	Visit, Report, Lab	70.36%	86.09%
Multimodal (BiLSTM)	-	7	5e-5	20	Visit, Report, Lab	72.92%	87.77%
Multimodal (BiLSTM)	-	9	3e-5	20	Visit, Report, Lab	73.56%	87.95%
571 classes							
Longformer	256	3	3e-5	64	Visit, Report, Lab	67.60%	84.24%
Longformer	256	3	3e-5	64	Visit, Report, Lab, Doctor	69.21%	86.14%
Longformer	256	3	3e-5	64	All data	69.45%	86.51%
Multimodal (BiLSTM)	-	2	3e-5	16	All data	68.50%	84.64%
Multimodal (BiLSTM), ASL	-	2	3e-5	16	All data	71.02%	86.77%
363 classes							
Longformer	512	3	3e-5	32	All data	72.45%	88.10%
Longformer	1024	3	3e-5	16	All data	73.70%	88.90%
Multimodal (BiLSTM), ASL	-	8	1e-5	36	All data	73.89%	88.88%

Table 3: AIDA results on the validation set, shown separately across three stages of ICD targets.

lower accuracy, AIDA met the CDSS integration threshold and has been widely in use since its deployment.

6 Deployment Challenges

The deployment of AIDA into the Clinical Decision Support System (CDSS) presented several challenges, primarily related to target class selection, data structure differences, system latency constraints, and model stability monitoring.

Changes in target diagnoses over time. The final diagnosis set was refined throughout the project. Initially based on prior research, it was later finalized by medical experts. Each update required retraining the model and revalidating results, adding complexity.

Differences in data storage structures. During training, AIDA used data from an analytical subsystem, but real-time querying was impractical. Instead, direct EHR retrieval was required, introducing challenges in aligning data formats to maintain consistency.

Infrastructure for timely processing. AIDA requires access to up to two years of patient history, making real-time retrieval from the EHR infeasible. To address this, we implemented a preloaded database storing patient records in a model-ready format. This significantly improved response times, though the initial data preparation process took several months due to EHR complexity.

Continuous monitoring. To maintain accuracy, we track two key metrics: (1) the proportion of diagnoses with high uncertainty and (2) the doctor agreement rate. A decline in either signals potential data drift or systemic changes, prompting retraining. This ensures AIDA remains reliable in evolving clinical environments.

7 Implementation Details

Figure 2 illustrates AIDA's conceptual role within the Clinical Decision Support System (CDSS), where it processes patient data from the last 1.5 years alongside real-time examination data. Within a second, the model generates a diagnosis, incorporating a confidence estimation module that indicates alignment with the doctor's assessment or highlights insufficient EHR data. This design keeps doctors in control while leveraging AI-driven insights.

Interaction with doctors. During a patient's visit, the doctor fills out an EHR form, documenting medical history, symptoms, and observations. When reaching the diagnosis field, a textual prediction from the model is displayed if the model's confidence is high enough, occurring in about 90% of relevant visits 2. Both the model's and the doctor's results are logged into the EHR's analytics module to analyze their agreement (currently 84.3%).

Preloading data. At launch, the system preloaded two years of electronic health records for all patients. Data was reformatted for model compatibility and stored in a specialized database to ensure rapid access during inference.

Continuous background data loading. To maintain up-to-date patient records, the system continuously ingests new medical documents as they are created. These include patient complaints, history, and other relevant information. Incoming data undergoes the same preprocessing steps as during the initial preloading phase, ensuring consistency in storage and retrieval.

Model prediction retrieval. During appointment, as the examination protocol is completed and the diagnosis field is reached, the CDSS queries the AIDA service. The re-

ICD-10	ICD Name	Count	Complexity (%)			Accuracy	F1
			Simple	Complex	Extra complex		
I11.9	Hypertensive heart disease	1445	78.34	19.65	2.01	0.936	0.931
E11.7	Type 2 diabetes mellitus w/ complications	277	57.76	34.66	7.58	0.832	0.845
E11.9	Type 2 diabetes mellitus w/o complications	167	47.90	36.53	15.57	0.823	0.806
I25.2	Old myocardial infarction	164	76.83	21.34	1.83	0.814	0.841
I48.0	Atrial fibrillation and flutter	163	69.94	25.77	4.29	0.891	0.861
I20.8	Other forms of angina pectoris	109	54.13	41.28	4.59	0.712	0.725
J45.8	Asthma	107	79.44	18.69	1.87	0.904	0.876
E11.6	Type 2 diabetes mellitus w/ manifestations	101	51.49	41.58	6.93	0.734	0.740
E89.0	Postprocedural endocrine disorders	74	91.89	8.11	0.00	0.973	0.973
I48.1	Persistent atrial fibrillation	70	65.71	31.43	2.86	0.824	0.828
Other	Various	823	50.91	42.40	6.69	0.774	0.88
Total		3500	55.20	40.80	5.00	0.861	0.859

Table 4: AIDA results on the test set for individual ICD codes, with proposed division of cases according to their complexity.

	No Synonym Accuracy		Synonym Accuracy		Cases	
	Doctor	AIDA	Doctor	AIDA	Number	% Total
Established Ground Truth	90	86.1	91.7	87.7	3332	95.2%
Full Expert Agreement	96.3	93.5	97.1	94.1	2367	67.6%
Partial Expert Agreement	74.6	68.1	78.2	72	965	27.6%
Doctor/AIDA Agree		98.5		99	2635	75.3%
Doctor/AIDA Do Not Agree	57.8	39.3	64	45.1	697	19.9%

Table 5: Comparison of doctor and AIDA accuracy across different conditions and the effect of employing synonymous terminology.

quest contains technical metadata alongside the patient's complaints and medical notes provided by the doctor. This information is supplemented with preloaded medical records from the database, forming a complete input sequence.

AIDA generates a predicted diagnosis with an uncertainty score. Initially measured by the highest softmax logit, uncertainty estimation was later refined using the HUQ-RDE model [Vazhentsev *et al.*, 2023] for greater reliability. If uncertainty is below a set threshold, AIDA provides a diagnosis; otherwise, no prediction is made. This filtering improves accuracy and naturally fosters trust to the tool among doctors.

8 Conclusion

This study highlights the potential of AI-driven models, such as AIDA, in enhancing clinical decision-making. Trained on 1.5 million patient records, our model reached 87.7% accuracy, supporting over 3 million diagnostic decisions in the real-world clinical setting.

Our experiments comparing Longformer-based and multimodal models showed that while the multimodal approach had a slight performance advantage, the Longformer-based model was more practical for large-scale deployment due to its lower computational complexity. Feature importance analysis revealed that patient complaints and medical history contributed the most to diagnostic accuracy, while laboratory and instrumental results added value but were less impactful.

AIDA has been well received by healthcare professionals, particularly for supporting less experienced doctors. However, some concerns about workflow integration remain. To address this, AIDA's recommendations are optional, ensuring that AI enhances decision-making without overriding clinical judgment. It is important to note that AIDA does not attempt to replace doctors but rather serves as an AI-powered assistant aimed at enhancing the accuracy and efficiency of medical diagnoses while ensuring that no critical details are overlooked.

Future work will focus on expanding diagnostic coverage, improving interpretability to increase clinician trust, and optimizing integration into diverse healthcare systems. By addressing these areas, we aim to further bridge AI-driven diagnostics with real-world clinical practice.

9 Limitations

While this study contributes to the integration of AI-driven clinical decision support systems, several limitations merit discussion. Specifically, the model's performance is currently tied to the specific regional system where it was trained and deployed. Generalization to diverse healthcare environments, with varying patient populations and data structures, requires further investigation and validation.

Additionally, the model's reliance on Longformer and BiLSTM architectures, while effective, represents a relatively established approach within the field of diagnostic prediction. While we explored multimodal models, the focus remained on the Longformer-based model due to its practicality for large-scale deployment.

Finally, though EHR data cannot be shared publicly due to

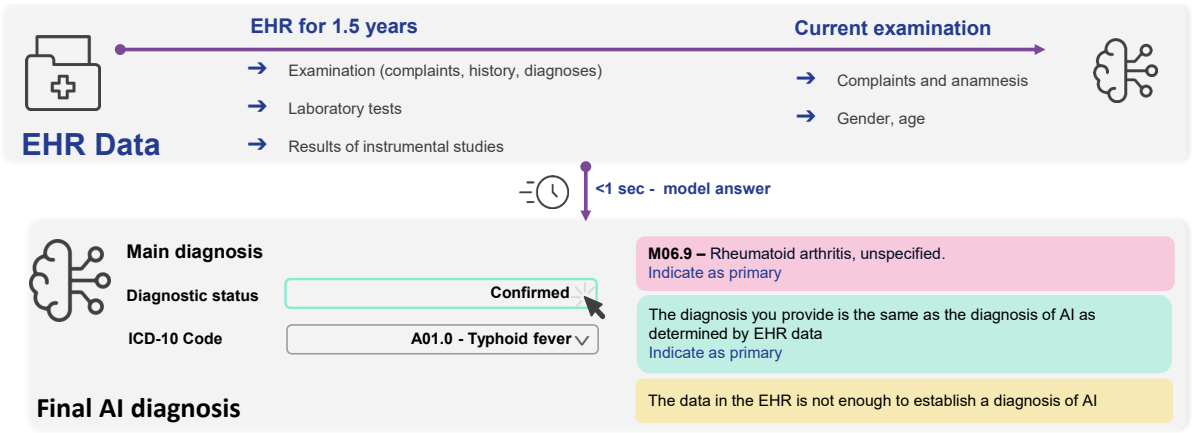

Figure 2: Schematic diagram of AIDA deployed into a CDSS with its 'Final AI diagnosis' recommendation appearing right on top of the user interface (translated to English for demonstration). The system retrieves past and current patient data to generate a real-time second opinion on probable diagnosis, incorporating confidence estimation and accepting immediate feedback from the doctor. If the model's confidence is high, an ICD-10 code is suggested. Otherwise, a notification appears indicating that the available EHR data is insufficient to establish a diagnosis. A warning about 'extra complex case' is issued when appropriate.

privacy regulations, we provide detailed preprocessing steps, and evaluation protocols to enable replication in comparable environments.

These limitations reflect trade-offs inherent to applied AI research but do not diminish AIDA's demonstrated success in enhancing diagnostic accuracy while preserving clinician autonomy through optional recommendations.

Appendices

A Processing an Incoming Patient Request

1. A patient or a medical professional initiates a request.
2. The **Diagnostic Assistant service** tokenizes the incoming request.
3. It then queries the patient's medical history from the **Data Process service**, which retrieves data from the **Historic Document Database**.
4. After obtaining the history, the **Diagnostic Assistant service** formats the data and forwards it to the **AIDA** for analysis and diagnosis prediction.
5. The predicted results are stored in the **Prediction Database**.

B Updating the Medical History

1. When a patient undergoes an examination or any medical event occurs, the data is recorded in the **MIS** (Medical Information System).
2. **MIS** sends specific documents to **Kafka** based on its internal logic.
3. **Converter SEMD** reads the data from **Kafka** and transforms it into a standardized format.
4. The processed data is then re-queued in **Kafka**.
5. **Data Process Service Adapter** extracts the data from **Kafka** and forwards it to the **Data Process service** for storage in the **Historic Document Database**.

In this system, the **Diagnostic Assistant service** acts as the orchestrator, ensuring patient data analysis and interaction with artificial intelligence, while **MIS**, **Kafka**, **Converter SEMD**, and **Data Process Service Adapter** handle updating and structuring the patient's medical history.

The system is designed as a structured pipeline that ensures efficient handling of patient data. The **Diagnostic Assistant service** manages patient data flow, coordinating data retrieval from the **Historic Document Database** via the **Data Process service** and leveraging artificial intelligence for predictive analysis. Simultaneously, patient examination records and related medical events are processed by **MIS** and forwarded to **Kafka** for structured transformation by the **Converter SEMD**. Once standardized, the data is reintroduced into **Kafka** and subsequently processed by the **Data Process Service Adapter**, which updates the **Historic Document Database**. This end-to-end pipeline ensures seamless integration of historical data retrieval, AI-driven diagnosis prediction, and medical history updates.

C Data example

Complaints: persistent moderate pain in the area of the left elbow joint. Diagnosis confirmed: M19.8 — Other specified osteoarthritis. **History**: According to the patient, they have been experiencing discomfort for about 1.5 months. Reports worsening. Self-treated with NSAIDs, no improvement noted. Visited for a nerve block procedure. Complaints: pain in the left heel with minimal load. Diagnosis confirmed: M77.3 — Heel spur. No other complaints; denies medication withdrawal symptoms. Diagnosis confirmed: E11.7 — Type 2 diabetes mellitus with multiple complications. History: X-ray of the right elbow joint (21.02.22) revealed radiological signs of stage 1-2 osteoarthritis.

Ethical Statement

This study was conducted in accordance with ethical guidelines for AI research in healthcare. All patient data were anonymized and aggregated, with identifiers removed to comply with privacy regulations. Informed consent was waived for this retrospective analysis of de-identified electronic health records. The authors declare no conflicts of interest related to AIDA's development or deployment.

Acknowledgements

We are grateful to the Central Medical Information and Analytical System of the city of Moscow for providing access to the anonymized electronic medical records that formed the basis of this paper. We are especially grateful to Yaroslav Bespalov for his guidance in the implementation process and significant contributions to the project. Finally, we express our appreciation to all participating physicians whose experience and feedback were instrumental in the development and implementation of AIDA in real-world practice.

Contribution Statement

Dmitry Umerenkov (D.U.), Alexandr Nesterov (A.N.) and Vladimir Shaposhnikov (V.S.) contributed equally: they designed and ran the experiments, computed the results and drafted the main manuscript. Vladimir Kokh (V.K.), Marina Kirina (M.K.) and Anton Abrosimov (A.A.) coordinated the physician-led data annotation, while V.K., Nikolay Romanenko (N.R.) and Ruslan Abramov (R.A.) oversaw the deployment process; A.N. and V.S. also implemented the production-level integration. Dmitry V. Dylov and Ivan Oseledets are co-senior authors who supervised the project and analysed the results. All authors took part in preparing the final manuscript.

References

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Beltagy *et al.*, 2020] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[Blinov *et al.*, 2020] Pavel Blinov, Manvel Avetisian, Vladimir Kokh, Dmitry Umerenkov, and Alexander Tuzhilin. Predicting clinical diagnosis from patients electronic health records using bert-based neural networks. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*, pages 111–121. Springer, 2020.

[Choi *et al.*, 2016] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Goldstein and Dengel, 2012] Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 1:59–63, 2012.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015.

[Kline *et al.*, 2022] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022.

[Ledley and Lusted, 1991] Robert S. Ledley and Lee B. Lusted. Reasoning foundations of medical diagnosis. *M.D. computing : computers in medical practice*, 8 5:300–15, 1991.

[Li *et al.*, 2020] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

[Liberman and Newman-Toker, 2018] Ava L Liberman and David E Newman-Toker. Symptom-disease pair analysis of diagnostic error (spade): a conceptual framework and methodological approach for unearthing misdiagnosis-related harms using big data. *BMJ quality & safety*, 27(7):557–566, 2018.

[Mukherjee *et al.*, 2023] Pritam Mukherjee, Marie Humbert-Droz, Jonathan H Chen, and Olivier Gevaert. Scope: predicting future diagnoses in office visits using electronic health records. *Scientific Reports*, 13(1):11005, 2023.

[Organization, 1992] World Health Organization. *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*, volume 1. World Health Organization, 1992.

[Ridnik *et al.*, 2021] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 82–91, 2021.

[Vazhentsev *et al.*, 2023] Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, 2023.

[Zhang *et al.*, 1997] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: A new data clustering algorithm and its applications. *Data mining and knowledge discovery*, 1:141–182, 1997.

[Zhang *et al.*, 2022] Angela Zhang, Lei Xing, James Zou, and Joseph C Wu. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering*, 6(12):1330–1345, 2022.