BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Can general purpose large language models assist pediatricians in predicting infants with serious bacterial infection?

Ivan Šimunović[1], Klara Rezić[1], Nikola Franić[3], Gabrijel Boduljak[4], Marijan Batinić[2], Ivana Jukić[2], Ivana Jelovina[2], Jela Biočić[1,2], Zenon Pogorelić[1,5*] and Joško Markić[1,2*]

Abstract

Background Serious Bacterial Infection (SBI) in neonates and young infants often exhibit nonspecific symptoms and clinical signs in the early stages of illness, making early diagnosis challenging. Timely recognition and appropriate treatment are essential to prevent adverse outcomes. While several clinical algorithms are widely used for SBI risk stratification, these tools have limitations, particularly low positive predictive value. This study evaluates the diagnostic accuracy of general-purpose large language models (LLMs) in detecting SBI in neonates and infants under 90 days of age admitted to the emergency department. Our objective is to improve diagnostic precision, reduce unnecessary interventions, and enhance patient outcomes. LLM performance was compared against traditional machine learning models, state-of-the-art rule-based methods, and an ensemble of physicians to assess their potential as clinical decision-support tools in scenarios of diagnostic uncertainty.

Results On a dataset of 742 patients, LLMs demonstrated diagnostic accuracy comparable to traditional machine learning models and state-of-the-art rule-based methods. The optimized CatBoost (class-weighted) model achieved the best overall performance, with a PPV of 0.70, NPV of 0.90, sensitivity of 0.54, specificity of 0.95, F1-score of 0.60, and MCC of 0.54, outperforming the baseline CatBoost model and achieving results on par with large language models (LLMs) and physicians. When optimally prompted, LLMs performed on par with ensembles of experienced clinicians. Additionally, LLMs exhibited effective medical reasoning and provided credible diagnostic predictions, particularly valuable in cases of clinician uncertainty. The models achieved balanced performance across multiple evaluation metrics, including PPV, NPV, sensitivity, specificity, F1-score, and Matthew's correlation coefficient (MCC). ChatGPT-4o achieved a sensitivity of 0.65 and specificity of 0.83, with an MCC of 0.41. Claude Sonnet 3.5 reached a sensitivity of 0.60 and specificity of 0.86, MCC 0.42 and Google Gemini 2.0 Flash had lower sensitivity (0.43) but the highest specificity (0.94), with an MCC of 0.43. In comparison, the best-performing individual pediatrician achieved a higher sensitivity (0.74) but lower specificity (0.68), with an MCC of 0.33, while the pediatricians' majority vote yielded sensitivity of 0.69, specificity of 0.81, and MCC of 0.43 — comparable to the top-performing LLMs.

*Correspondence:
Zenon Pogorelić
zenon.pogorelic@mefst.hr
Joško Markić
jmarkic@mefst.hr

Full list of author information is available at the end of the article

Conclusions These Artificial intelligence tools offer a promising direction for SBI risk prediction, achieving performance comparable to that of experienced pediatric specialists, while maintaining simplicity of use/data-preprocessing for potential real-world applications.

Keywords Large language models, Machine learning, Serious bacterial infection, Pediatrics, Infectology, Diagnostics, Prediction

Background

Infants with Serious Bacterial Infection (SBI), such as pneumonia, urinary tract infection, meningitis, bacteremia, or bone and joint infections, often exhibit nonspecific symptoms and clinical signs, like fever, in the early stages of illness [1–3]. This makes it difficult to differentiate SBIs from viral infections [4]. Early recognition and treatment with appropriate antibiotic therapy are critical to avoiding the severe consequences of delayed or missed diagnoses [2]. Determining which infants require hospitalization and antibiotic treatment is a persistent challenge, complicated by changing clinical presentations and epidemiology [5].

Several algorithms are widely used to evaluate the risk of SBI in infants, including the Rochester, Boston, Philadelphia, Lab-Score, and Step-by-Step approach [6–9]. These tools are based on specific clinical and laboratory parameters that help classify infants into risk categories. For example, the Step-by-Step approach validated by Gomez et al. [9] showed higher sensitivity and specificity than the Rochester and Lab-Score criteria in identifying low-risk infants for invasive bacterial infections (IBI). However, significant limitations remain. The Rochester, Boston, and Philadelphia criteria have low positive predictive value (PPV), reducing their effectiveness in specifically identifying SBIs [10]. The Step-by-Step approach has been prospectively validated, showing a sensitivity of 0.92 and a negative predictive value (NPV) of 0.993 [9]. However, significant variability exists across institutions, as not all guidelines advocate for these strategies, contributing to inconsistencies in clinical practice [11]. However, the accuracy of these scores or rules relies heavily on the availability and reporting of test results. Predictions cannot be made in the presence of missing values, which represents a significant limitation [12]. As a result, in resource-limited settings or for patients without specific test results, these methods may not be applicable.

Machine learning (ML) models offer a promising alternative to traditional methods by leveraging large datasets and sophisticated algorithms to improve predictive accuracy. Unlike conventional tools, ML models can handle missing or incomplete data effectively through techniques like imputation or treating missing values as predictors. Studies by Ramgopal et al. [13] and Lee et al. [14] have shown the potential of ML algorithms, such as Random Forest and Support Vector Machines, in predicting SBI with high sensitivity and specificity. For instance, Ramgopal et al. [13] developed an ML model using features like urinalysis, white blood cell count, absolute neutrophil count, and procalcitonin levels. Their Random Forest model achieved a sensitivity of 0.986 and a specificity of 0.749, significantly reducing unnecessary procedures like lumbar punctures, hospitalizations, and antibiotic treatments. Similarly, Lee et al. [14] created an ML model that demonstrated robust performance even with incomplete datasets, achieving an area under the receiver operating characteristic (AUROC) curve of 0.964 in internal validation and 0.950 in external validation.

Large Language Models (LLMs) have further enhanced the capabilities of traditional ML approaches. These models can process and contextualize large amounts of structured and unstructured data, including clinical notes, medical literature, and electronic health records. This ability allows LLMs to uncover patterns and correlations that traditional models might miss. Fisch et al. [15], for example, found that LLMs like ChatGPT4o performed better than physicians in recommending diagnostic procedures and treatments for bacterial meningitis. However, the study also noted challenges, including the generation of misleading statements and inconsistencies in treatment recommendations. For SBI prediction, LLMs offer an attractive approach due to their minimal data preprocessing requirements and demonstrated success across various medical applications [15].

Prior studies provide strong evidence that medical large language models possess substantial clinical knowledge [16, 17]. However, these findings are based primarily on benchmark datasets like MedQA, MedMCQA, and PubMedQA, while the models are proprietary (e.g. no open access). While useful, such datasets consist of multiple-choice questions or unrealistic isolated scenarios that do not capture the complexity and ambiguity of real-world clinical practice. Moreover, they largely test general medical knowledge, leaving unclear whether strong benchmark performance translates into practical clinical utility.

In contrast, we investigate whether open-access, general large language models (LLMs) can address a highly specific and challenging real-world task in the pediatric emergency department: assessing the risk of SBI in infants during pediatric emergencies. We find that LLMs perform on par with—or better than—both traditional machine learning methods and expert pediatricians. Importantly, our approach requires minimal data

preprocessing (e.g. works on raw lab reports), helping to bridge the gap between academic benchmarks and actual clinical application.

Methods

Database and preprocessing

This study utilizes a dataset comprising infants under 90 days of age who visited the pediatric emergency department (PED) at the University hospital of Split, Croatia. The dataset includes only those infants who were hospitalized following their emergency department visit, covering the period from January 1st 2021 until August 30th 2024. The data was collected from hospital records and electronic health systems to ensure comprehensive and accurate documentation of patient demographics, clinical presentations, and hospitalization details. The dataset primarily comprises laboratory parameters obtained during the PED visit, as well as clinical and demographic data recorded by the attending pediatricians at admission to the hospital. The confirmation of SBI was obtained from the discharge letters of hospitalized infants, where diagnoses were documented using the ICD-10 classification system. The SBI group comprised patients with an ICD-10 code corresponding to an SBI diagnosis available in Additional file 1. The non-SBI group comprised patients without ICD-10 diagnosis indicating SBI, as stated in the discharge letter. This group was heterogeneous and included patients with viral infections, self-limiting febrile illnesses, and various other non-bacterial conditions presenting to the pediatric emergency department. The decision to build the dataset in an explained manner was made to ensure proper evaluation of all classifiers in the most realistic environment, while focusing on SBI. A total of 742 patients were included in this study. Among them, 126 had confirmed SBI based on discharge letters, while 616 were classified otherwise. The data used in this study was examined and anonymized by the IT department of University Hospital Split to ensure patient confidentiality and compliance with data protection regulations. No personally identifiable information was accessible to the researchers. Patient data were anonymized and handled in accordance with institutional guidelines to protect confidentiality and privacy.

For the initial comparison of LLMs with logistic regression, CatBoost (Categorical Boosting), LabScore, and FIRST algorithms, a set of 60 categorical and numerical features was selected by a medical expert. These features were deemed relevant for training models to predict SB, based on the medical expert's opinion. The complete list of features is provided in Additional file 2. In the subsequent analysis, we selected the best-performing algorithm and further tested it in a more realistic environment, where the model was provided with the full dataset from the PED. Its performance was then compared to that of medical doctors.

Implementation

The code was written in Python programming language (version 3.11.9, Python Software Foundation, Wilmington, DE, USA).

Algorithms

The **Febrile Infants Risk Score at Triage (FIRST) Algorithm** [18] is a scoring system designed to assess the risk of serious bacterial infections (SBIs) in febrile infants at triage. It assigns points based on four key clinical variables: age, temperature, sex, and duration of fever. For age, infants younger than 21 days receive 17 points, those between 21 to less than 28 days receive 0 points, and those 28 days or older receive 30 points. For temperature, infants with a temperature below 38.5 °C receive 0 points, those between 38.5 °C to less than 39.0 °C receive 4 points, those between 39.0 °C to less than 40.0 °C receive 13 points, and those with a temperature of 40.0 °C or higher receive 43 points. For sex, male infants receive 17 points while female infants receive 0 points. Finally, for the duration of fever, infants with less than 2 days of fever receive 0 points, whereas those with fever lasting 2 days or more receive 9 points. The total score is measured against the specified threshold and used to determine the risk level, guiding clinicians in deciding on further evaluation or treatment for febrile infants. *For the motivation behind thresholds see their original paper*.

Lab-Score Method [19, 20] is a scoring system used to predict the likelihood of serious bacterial infections (SBI) in febrile infants based on laboratory test results. It considers three key biomarkers: procalcitonin (PCT), C-reactive protein (CRP), and urine dipstick results. The score is assigned based on the presence and levels of these markers. A positive urine dipstick, defined as positive leukocyte esterase and/or positive nitrate, contributes 1 point. If PCT is ≥ 0.5 ng/mL or CRP is ≥ 40 mg/L, 2 points are assigned. If PCT is ≥ 2 ng/mL or CRP is ≥ 100 mg/L, 4 points are assigned. The total Lab-score ranges from 0 to 9 points. A score of 3 or higher is considered the optimal cutoff for predicting SBI, helping clinicians determine the need for further evaluation or treatment in febrile infants.

Logistic regression is a supervised machine learning algorithm commonly used for binary classification tasks. It models the relationship between one or more independent variables—typically numerical features—and a binary dependent variable. The algorithm estimates the probability of an instance belonging to a particular class using a logistic (sigmoid) function, producing outputs in the range [0, 1]. A discrete class label (e.g., 0 or 1) is then

assigned by applying a threshold to this predicted probability [21, 22].

CatBoost (Categorical Boosting) is an efficient classification algorithm based on gradient boosting over decision trees, specifically designed to handle categorical features in the data. Unlike many other methods that require prior transformation of categorical variables, CatBoost incorporates statistical techniques to process them automatically. Furthermore, by optimizing a wide range of input parameters, CatBoost effectively reduces the risk of overfitting [23–25]. To evaluate model performance, we employed five-fold cross-validation. The dataset was split into five folds, and for each iteration, the model was trained on four folds and tested on the remaining fold. This process was repeated five times, producing five models trained on different subsets of the data. All models shared the same hyperparameters; the only difference between them was the training subset. We report the mean evaluation metrics of the five models as the final result for the CatBoost model. Hyperparameter optimization was performed by specifying a fixed set of hyperparameters for each trial, applied across all five folds. The outcome of a single optimization trial was the average evaluation metric across the five models, along with the corresponding hyperparameters. This process was repeated for 30 trials, resulting in the optimized hyperparameters presented in the study. We acknowledge that using the same five-fold cross-validation procedure both for hyperparameter optimization and for reporting performance may introduce a slight optimistic bias in the results, as the folds used for tuning are not fully independent of those used for evaluation. However, this approach was adopted for model comparison rather than for unbiased generalization estimation. All models used in this study were evaluated under exactly the same cross-validation scheme. Consequently, any potential overestimation affects all models equally, preserving the validity of relative comparisons. The reported values should therefore be interpreted as cross-validated performance estimates, not as performance on a completely unseen test set.

In addition to these traditional models, we evaluated some of the **best general-purpose LLMs** available at the time of writing: **Chat Generative Pre-trained Transformer 4o (ChatGPT4o) (Open AI)**, **Gemini 2.0 Flash (Google)**, **and Claude 3.5 Sonnet (Anthropic).** These models were tasked with predicting the likelihood of high-risk SBI cases while also providing a detailed explanation of their predictions, including the rationale behind their decision-making process and the confidence levels associated with each prediction. The prompt used for analysis is available in the Additional file 3. Different approaches to LLM inference are used to improve their performance in healthcare applications: **Zero-Shot learning** is a prompt-engineering approach in which a model is provided with a prompt without undergoing any task-specific training. It is commonly applied in transfer learning, where a pre-trained model is repurposed for a new task instead of fine-tuning a separate model, thereby substantially reducing the computational cost and resources required for training [26].

Few-Shot learning refers to a setting in which the model is provided with a small number of task demonstrations at inference time as conditioning, without any updates to the model's weights [27]. **Chain-of-Thought (CoT) prompting** is a technique designed to enhance reasoning in large language models by encouraging them to generate intermediate steps when solving complex problems. This approach enables models to decompose multi-step tasks into smaller reasoning units, thereby allocating additional computation to problems that require deeper inference. Importantly, CoT provides an interpretable trace of the model's reasoning process, offering insights into how answers are derived and allowing researchers to identify potential sources of error [28].

We also conducted a **human evaluation** to compare the performance of pediatric specialists with tested algorithms in predicting the risk of SBI. The five specialists were randomly selected from a pool of board-certified pediatricians currently working in the pediatric department. The specialists were presented with the same data as our tested algorithms. Laboratory parameters (available in Additional file 4) were taken at the PED along with basic patient information, including sex, age, days of fever, and body temperature at the time of the emergency department visit. Their task was to assess the risk of SBI by assigning a prediction of 1 (high risk, possible SBI) or 0 (no risk for SBI) based on the available data. Additionally, they were required to indicate their confidence level (confident or not confident) for each prediction. This evaluation protocol provides valuable insights into how closely ML and LLM models align with expert judgment in assessing SBI risk in infants, considering diagnosis uncertainty beyond just accuracy.

Model training and evaluation

For Logistic Regression (LR) and CatBoost, we employed a 5-fold cross-validation procedure. Specifically, models were trained on four folds and evaluated on the remaining fold, with the process repeated across all five folds. The reported performance metrics represent the average over the five runs. LLMs were evaluated on the same folds without additional training. All folds were generated using a fixed random seed to ensure reproducibility.

Evaluation metrics

Evaluating the performance of machine learning models in medical research requires a set of well-defined

metrics to assess their predictive ability. The choice of metric depends on the nature of the problem, particularly whether it involves balanced or imbalanced datasets and the importance of different types of errors. Several commonly used evaluation metrics include **PPV (positive predictive value)**, **NPV (negative predictive value)**, **sensitivity**, **specificity**, **F1-score**, and **Matthews Correlation Coefficient (MCC)**, each providing distinct insights into model performance. To better estimate model performance on unseen data, we used five-fold cross-validation and reported the mean values as the final performance metrics. Following performance evaluation was guided by prioritizing high specificity, followed by high sensitivity. This strategy was adopted to minimize false-positive predictions while maintaining robust detection capability, reflecting clinical priorities. The approach was applied consistently across all models to ensure comparability and interpretability of results.

Results

State-of-the-art rule-based methods are insufficient

The **FIRST** algorithm was tested against five threshold values (13, 21, 30, 38, and 43) to determine its effectiveness in identifying high-risk SBI cases. The best-performing threshold was t = 43, achieving the highest PPV 0.2202 and specificity 0.6251, thereby reducing false positives. Sensitivity at the highest threshold remained moderate at 0.5156 with MCC scores 0.1084 and F1 scores 0.3078. These findings suggest that higher threshold values provide an optimal trade-off between detecting SBI cases and avoiding excessive hospital admissions.

The **Lab-Score Method** demonstrated high specificity (0.9476), effectively reducing false positives and minimizing unnecessary hospitalizations. However, its low sensitivity (0.2308) indicated a higher likelihood of missed SBI cases. Despite achieving an overall PPV of 0.4832, the imbalance between sensitivity and specificity suggests that the model struggles to reliably detect true SBI cases. The MCC score (0.2427) and F1 score (0.3014) further reflects its moderate diagnosis ability. While the Lab Score Method is effective at ruling out non-SBI cases, its limited sensitivity requires complementary diagnostic procedures to prevent missed infections.

Traditional ML methods are much better than rule-based methods

Logistic Regression demonstrated considerably better performance than previous algorithms, achieving an F1 score of 0.488 and an MCC score of 0.369. The LR class-balanced model achieved a PPV of 0.39 and NPV of 0.92, showing improved precision. This improvement was expected, as logistic regression was the first algorithm to leverage the entire dataset of 60 preprocessed features. To address the imbalance between positive and negative cases, the class-balanced weighting strategy was applied during model training, ensuring more equitable representation of both classes. These enhancements contributed to the superior predictive performance of logistic regression compared to earlier methods.

The **CatBoost Algorithm** was tested in two scenarios. In the first, class weights were assigned based on data distribution, while default hyperparameter values were used. In the second, in addition to class weighting, hyperparameter optimization was performed by defining a custom range around the default values for specific parameters, running 30 optimization trials. The **CatBoost (default and class-weighted)** model achieved a sensitivity of 0.5544, improving SBI detection compared to previous algorithms, while maintaining a high specificity of 0.9416, effectively reducing false positives. With an F1-score of 0.6 and an MCC score of 0.533, this model demonstrated a strong balance between detecting true SBI cases and minimizing false positives. We tested the performance of CatBoost algorithms on different thresholds, and the result is that the default threshold of 0.5 gives the best performance. CatBoost (Class-Weighted) showed a substantial increase in PPV to 0.66 with an NPV of 0.91, demonstrating strong precision and maintained negative predictive accuracy.

Further hyperparameter optimization slightly improved sensitivity (0.5913), but at the cost of reduced specificity (0.7929). The optimized version achieved an F1-score of 0.54 and an MCC score of 0.37, indicating comparable performance but with a shift toward improved recall at the expense of more false positives but Optimized CatBoost (Class-Weighted) achieved the highest PPV of 0.70 and an NPV of 0.90, indicating the best balance between correctly identifying positive cases and maintaining high negative predictive performance across all models.

Details of the hyperparameter optimization process and specific parameter values are provided in the Additional file 5. To interpret the predictions of the CatBoost algorithm, we used the SHAP (SHapley Additive exPlanations) method [29]. By visualizing SHAP feature importance scores, we identify the most relevant features and present them in Fig. 1. Figure 1 confirms that CatBoost relies on medically relevant features for its predictions.

LLMs benefit from raw, unprocessed data

As shown in Table 1, the **optimized CatBoost** model demonstrated the best overall performance compared to other algorithms considering F1 and MCC metrics. In addition to its strong predictive ability, this algorithm offers the advantage of prediction interpretability. It was also observed that LLM models performed well when tested on the same preprocessed dataset used for training traditional machine learning algorithms. The results, presented in Table 2, indicate that LLM models performed

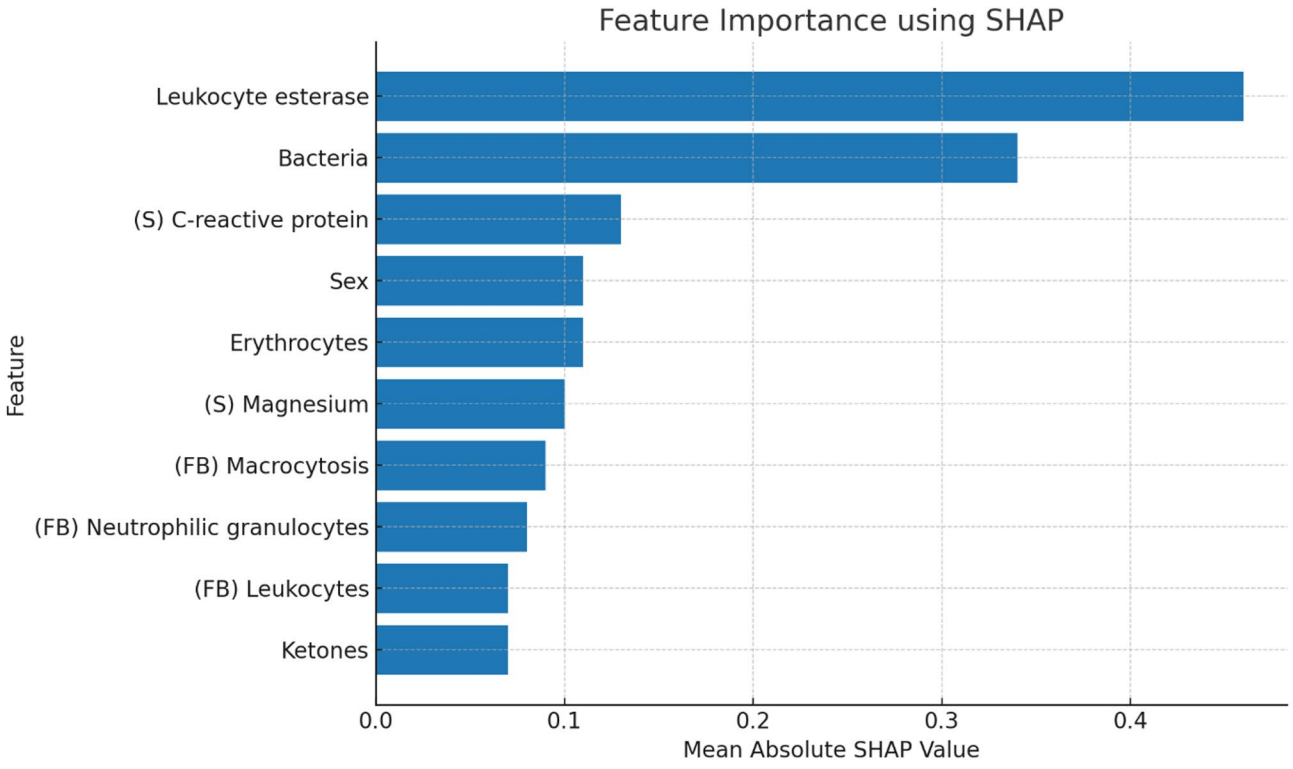

Fig. 1 SHAP feature importance of laboratory and demographic variables. SHAP – SHapley Additive exPlanations; FB – Full Blood; S – serum; features are ranked by mean absolute SHAP value indicating their impact on model predictions

Table 1 Performance metrics comparison of machine learning models

Model	PPV	NPV	Sensitivity	Specificity	F1 score	MCC
FIRST(t = 13)	0.1742	0.9158	0.9672	0.0550	0.2955	0.0553
FIRST(t = 21)	0.1750	0.8540	0.8384	0.1898	0.2895	0.0301
FIRST(t = 30)	0.1739	0.8487	0.8234	0.1996	0.2872	0.0235
FIRST(t = 38)	0.2126	0.8640	0.5478	0.5861	0.3059	0.1015
FIRST(t = 43)	0.2202	0.8635	0.5156	0.6252	0.3078	0.1085
LR class-balanced	0.3925	0.9168	0.6505	0.7906	0.4883	0.3691
Lab-Score algorithm	0.4832	0.8579	0.5203	0.9477	0.4015	0.2427
Catboost (Class-Weighted)	0.6636	0.9121	0.5545	0.9416	0.5107	0.5339
Optimized CatBoost (Class-Weighted)	0.6976	0.9021	0.5377	0.9498	0.6009	0.5379

PPV- positive predictive value; NPV- negative predictive value; Specificity – True Negative Rate; F1 – F1-Score (harmonic mean of Precision and Recall); MCC – Matthews Correlation Coefficient; "LR"-Logistic Regression; "ChatGPT"- Chat Generative Pre-trained Transformer; "FIRST"- The Febrile Infants Risk Score at Triage; values are presented as mean across five folds (rounded to four decimal places)

better when given access to the full, unprocessed patient data compared to the preprocessed dataset.

Zero-shot LLMs are strong, but slightly worse than CatBoost

ChatGPT4o achieved a sensitivity of 0.7040 with a specificity of 0.7792, indicating improved precision in identifying non-SBI cases. **Gemini 2.0 Flash** exhibited the highest specificity (0.9442), but had the lowest sensitivity (0.3789), suggesting a more conservative threshold that may lead to missed diagnoses. **Claude 3.5 Sonnet** demonstrated a more balanced performance, with a sensitivity of 0.6644 and a specificity of 0.8276. These results indicate that **ChatGPT4o** provides the best trade-off between sensitivity and specificity, making it the most reliable model for accurate SBI detection while minimizing unnecessary hospitalizations. It is worth noting that these models are general purpose language models, not specific to medical use cases.

Advanced prompting strategies can considerably improve LLMs diagnostic ability

Additionally, we explored the few-shot and chain-of-thought prompting methods to further improve LLM performance. For this step we chose only GPT4o because

Table 2 Performance metrics comparison of ML models and Human-Evaluators

Model	PPV	NPV	Sensitivity	Specificity	F1 score	MCC
ChatGPT4o AR	0.4339	0.9221	0.6476	0.8261	0.5177	0.4098
Google Gemini 2.0 Flash AR	0.6097	0.8895	0.4287	0.9413	0.5005	0.4289
Claude Sonnet 3.5 AR	0.4658	0.9140	0.6006	0.8587	0.5225	0.4173
ChatGPT4o AR Few-Shot	0.4514	0.9175	0.6237	0.8457	0.5231	0.4161
ChatGPT4o AR CoT	0.3451	0.9155	0.6536	0.7466	0.4505	0.3229
Best performing pediatrician	0.3247	0.9291	0.7445	0.6801	0.4510	0.3282
Pediatricians' majority voting	0.4406	0.9278	0.6881	0.8146	0.5345	0.4300
Pediatrician + ChatGPT4o AR Few-Shot	0.3661	0.9292	0.7198	0.7433	0.4845	0.3697
ChatGPT4o	0.3945	0.9292	0.7041	0.7792	0.5043	0.3953
Google Gemini 2.0 Flash	0.6004	0.8819	0.3790	0.9442	0.4586	0.3362
Claude 3.5 Sonnet	0.4453	0.9243	0.6645	0.8277	0.5305	0.4261

"AR" – all relevant data; "CoT" – chain-of-thought prompting; "ChatGPT"- Chat Generative Pre-trained Transformer; CoT- Chain of Thought; "Few-Shot"- Few-Shot learning; PPV – positive predictive value; NPV negative predictive value; Sensitivity – True Positive Rate; Specificity – True Negative Rate; F1 Score- (harmonic mean of Precision and Recall); MCC – Matthews Correlation Coefficient; values are presented as mean across five folds

it had the lowest numbers of false negative predictions which is in a scope of medical diagnosis the worst case scenario. The results, summarized in Table 2, show that the few-shot method had better, while chain-of-thought worse performance compared to zero-shot predictions, using all available patient data. Specifically, few-shot prompting resulted in a sensitivity of 0.6237, specificity of 0.8456, F1 of 0.5231 and MCC of 0.4160, while the chain-of-thought achieved a sensitivity of 0.6536, specificity of 0.7466, F1 of 0.4505 and MCC of 0.3229. Although chain-of-thought prompting typically outperforms few-shot prompting, in our experiments the few-shot approach yielded significantly better results. We hypothesize that this is because our task relies more on knowledge of laboratory reference values and pattern recognition within these results, rather than on advanced reasoning about general medical knowledge.

LLM's diagnostic ability is comparable to or better than human experts

For a final evaluation of LLM models, human specialist assessments were included in the analysis. Five medical specialists were provided with the same patient data as GPT4o. Their predictions were recorded, and Table 2 presents performance metrics for both the best-performing individual specialist and the majority voting outcome across all five experts. Additionally we tested the performance of the best performing doctor combined with predictions of LLM, which is in detail explained in the *Discussion* section. When these human evaluation metrics were compared, it was observed that *GPT4o and the optimized CatBoost model achieved comparable or superior performance to the best human evaluators.*

Discussion

The goal of this research was to explore the potential of AI-assisted models in reducing unnecessary hospitalizations while ensuring the early and accurate detection of high-risk SBI cases. Missing an SBI diagnosis poses a significant risk of mortality, making it critical to balance sensitivity and specificity in AI-driven decision support systems. Given that the consequences of false-negative diagnoses far outweigh the costs of false positives, we carefully balanced high specificity with high sensitivity during model development and performance analysis. Our objective was to minimize unnecessary hospitalizations while ensuring the accurate detection of high-risk SBI cases. Although SBI cases are rare, misclassification can have severe consequences, reinforcing the need for models that minimize both false negatives and false positives [30]. As demonstrated in the results, ML models, particularly LLMs and CatBoost, exhibited performance comparable to or better than pediatric specialists when evaluated using F1-score and MCC metrics. However, to fully understand their classification performance, additional metrics that include confidence were analyzed beyond just standard accuracy measurements.

Before interpreting the human evaluation metrics, it is important to acknowledge that the specialists' performance was constrained by the data provided for classification. In real clinical settings, physicians have additional context from patient observations, which helps inform their decisions beyond laboratory data alone. Consequently, while AI models were evaluated on the same structured data as the specialists, real-world physician decision-making benefits from a broader clinical picture [31]. In clinical practice, an effective classification model must achieve a balance between high sensitivity (to minimize missed high-risk cases) and high specificity (to prevent unnecessary hospitalizations). Previous algorithms were developed with a focus on high sensitivity and high negative predictive values, making them effective for ruling out serious infections [32]. However, their relatively modest specificity reduced the positive predictive value, frequently leading to unnecessary hospitalizations and overtreatment. In contrast, our study demonstrates that

LLM and ML-based models exhibit a more balanced performance profile. The results show that pediatric specialists achieved the highest sensitivity (0.74), indicating a strong ability to identify SBI cases and minimize the risk of missed diagnoses. However, their lower specificity (0.68) compared to most tested LLM models suggests a greater tendency to hospitalize patients without SBI. This pattern reflects a more conservative clinical approach, prioritizing patient safety and minimizing false negatives, even at the expense of increased false positives.

This study highlights both the potential benefits and limitations of integrating AI into medical decision-making in hospital settings. Among the evaluated models, GPT4o (using all available lab data) and CatBoost demonstrated the best overall performance. Despite achieving human-equivalent performance, the key limitation of LLMs is their overconfidence. Previous studies [33, 34] have shown that when these models are integrated into medical environments, their overconfidence can bias specialists toward the model's predictions.

However, a key challenge in achieving perfect predictions both for doctors and algorithms was the use of real-world hospital data. This data inherently contains irregularities that, while challenging model performance improvements, also provide the most realistic testing environment. Notably, most failure cases for both the best-performing models and doctors occurred in borderline cases, where even after detailed examination, classifying the risk of SBI remained nontrivial. However, it is essential to recognize that this comparison does not imply that LLMs are on parto human specialists, as doctors in real world scenarios have access to contextual and observational information beyond laboratory data. Nonetheless, these results, together with other examples from literature, strongly support the role of ML models in diagnostics, demonstrating their potential to enhance and assist physicians in hospital settings [35, 36].

Both LLMs and CatBoost have unique strengths and limitations. LLMs require minimal data preprocessing and can generate natural, human-like explanations, making them well-suited for integration into physicians' workflows [37, 38]. With proper fine-tuning, their predictive accuracy could be further improved [39]. Interpretability remains a challenge, as these models function as black boxes, making it difficult to understand the exact reasoning behind their predictions [40]. However, by prompting a LLM model to give the explanation for its classification we can get a better insight in how the decisions are made. However, obtaining ground truth labels for a dataset of this size is not feasible due to the substantial time and cost required for expert physician annotation. In this study, we focused on a limited subset of cases where the explanations generated by large language models (LLMs) were reviewed and confirmed by physicians as medically accurate and clinically sound. While these findings are encouraging and suggest the potential of LLMs for reliable medical reasoning, further investigation on a larger, systematically annotated dataset is necessary to validate and generalize these observations. Despite showing highest MCC and F1 values, CatBoost showed lowest sensitivity compared to LLM and medical specialists, which means that it detects fewer true positive cases which is one of our main metrics that we try to minimize. But on the other hand it offers several advantages. It can be easily trained, making it a cost-effective and accessible solution that can be tailored to specific hospital needs [41]. Unlike LLMs, CatBoost models are highly interpretable, allowing for a clearer understanding of how predictions are made. The primary drawback, however, is the need for the intensive data preprocessing, as the model requires properly formatted input for effective training and inference (predictions) [42].

These findings underscore the potential of AI-driven approaches in supporting clinical decision-making, particularly in optimizing triage processes and reducing unnecessary hospitalizations [43]. However, careful validation is necessary before full clinical implementation to ensure patient safety, mitigate biases, and maximize the practical benefits of AI-assisted diagnostics in real-world hospital settings [44, 45]. Ultimately, this analysis demonstrates that ML models have the potential to assist physicians in their daily practice, contributing to more efficient and accessible healthcare [46]. But many of the limitations observed in our study align with findings from other domains beyond medicine, where LLMs have been shown to be easily influenced by extraneous information. Prior research has demonstrated that optimizing the instructions alone can lead to performance variations of 8% to 50%, further highlighting the instability of these models in high-stakes decision-making [47]. With rapid advancements in AI, improved iterations of these models may soon reach expert-level performance in specific medical applications, provided they undergo rigorous validation and refinement. For this reason, clinicians, especially those in infectious disease and emergency settings, must remain actively engaged in AI advancements to ensure that future implementations prioritize patient safety and clinical efficacy [48].

Conclusion

This study highlights the promising role of machine learning models, particularly LLMs and CatBoost, in improving diagnostic accuracy and clinical decision-making for pediatric serious bacterial infection (SBI). These AI tools offer a promising direction for SBI risk prediction, achieving performance comparable to that of experienced pediatric specialists, while maintaining simplicity of use/data-preprocessing for potential real-world

applications. We show preliminary evidence that these methods can help ensure timely intervention for high-risk infants, while reducing unnecessary hospitalizations.

While LLMs demonstrate strong potential, their "overconfidence" and "black box" nature require further investigation to ensure safe clinical integration. Future research should focus on developing methods to improve LLM calibration and interpretability, potentially through advanced prompting techniques or hybrid models that combine LLM strengths with more transparent algorithms like CatBoost. Additionally, future work should explore strategies to enhance model robustness across diverse clinical settings and patient populations, addressing limitations posed by real-world hospital data irregularities. Furthermore, the clinical workflow integration of these AI tools needs careful consideration. Studies should evaluate how LLMs and CatBoost can best assist physicians, optimizing their efficiency and decision-making.

Ultimately, while AI holds great potential to improve pediatric emergency care, ongoing research is essential to address current limitations, validate findings, and ensure the responsible and effective translation of these technologies into clinical practice.

Abbreviations

SBI	Serious Bacterial Infection
LLM	Large language models
MCC	Matthew's correlation coefficient
IBI	Invasive bacterial infections
NPV	Negative predictive value
PPV	Positive predictive value
ML	Machine learning
AUROC	Area under the receiver operating characteristic
PED	Pediatric emergency department
FIRST	The Febrile Infants Risk Score at Triage
CatBoost	Categorical Boosting
ChatGPT4o	Chat Generative Pre-trained Transformer 4o
CoT	Chain-of-Thought
SHAP	SHapley Additive exPlanations

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03258-3.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Acknowledgements
None.

Author contributions
I.Š. (Writing – original draft, Data curation, Investigation, Formal analysis, Methodology, Visualization), G.B. (Writing – review & editing, Conceptualization, Formal analysis, Methodology, Software, Validation), N.F. (Writing – original draft, Data curation, Formal analysis, Methodology, Software, Validation), K.R. (Writing – original draft, Investigation, Resources), M.B. (Resources, Validation), I.J. (Resources, Validation), J.B. (Resources, Conceptualization), Z.P. (Project administration, Resources, Validation, Supervision) and J.M. (Project administration, Writing – original draft, Resources, Validation, Supervision)

Funding
This research received no external funding.

Data availability
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
The study was conducted in full compliance with the ethical principles set forth in the Declaration of Helsinki, the foundational guideline of the World Medical Association for research involving human participants. Written consent to participate in the study was obtained from all parents or legal guardians of the patients. This research was approved by the Ethical Committee of the University Hospital of Split, Croatia, ensuring compliance with ethical standards and regulations for medical research (approval number: 2181 – 147/01–06/LJ.Z.-24-02, Date of Approval: 30 August 2024).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]School of Medicine, University of Split, Split 21000, Croatia
[2]Department of Pediatrics, University Hospital of Split, Split 21000, Croatia
[3]Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Split 21000, Croatia
[4]Department of Computer Science, University of Oxford, Oxford, UK
[5]Department of Pediatric Surgery, University Hospital of Split, Split 21000, Croatia

Received: 24 June 2025 / Accepted: 22 October 2025
Published online: 14 November 2025

References

1. Nelson DS, Walsh K, Fleisher GR. Spectrum and frequency of pediatric illness presenting to a general community hospital emergency department. Pediatrics. 1992;90:5–10. https://pubmed.ncbi.nlm.nih.gov/1614779/.
2. Baraff LJ, Bass JW, Fleisher GR, et al. Practice guideline for the management of infants and children 0 to 36 months of age with fever without source. Agency for health care policy and research. Ann Emerg Med. 1993;22:1198–210. https://doi.org/10.1016/s0196-0644(05)80991-6.
3. Hamilton JL, John SP. Evaluation of fever in infants and young children. Am Fam Physician. 2013;87:254–60. https://pubmed.ncbi.nlm.nih.gov/23418797/.
4. DePorre AG, Aronson PL, McCulloh RJ. Facing the ongoing challenge of the febrile young infant. Crit Care. 2017;21:68. https://doi.org/10.1186/s13054-017-1646-9.
5. Greenhow TL, Hung YY, Herz AM, Losada E, Pantell RH. The changing epidemiology of serious bacterial infections in young infants. Pediatr Infect Dis J. 2014;33:595–9. https://doi.org/10.1097/INF.0000000000000225.
6. Jaskiewicz JA, McCarthy CA, Richardson AC, et al. Febrile infants at low risk for serious bacterial infection–an appraisal of the Rochester criteria and implications for management. Febrile Infant Collaborative Study Group Pediatr. 1994;94:390–6. https://doi.org/10.1542/peds.94.3.390.
7. Bachur RG, Harper MB. Predictive model for serious bacterial infections among infants younger than 3 months of age. Pediatrics. 2001;108:311–6. https://doi.org/10.1542/peds.108.2.311.
8. Markic J, Kovacevic T, Krzelj V, Bosnjak N, Sapunar A. Lab-score is a valuable predictor of serious bacterial infection in infants admitted to hospital. Wien Klin Wochenschr. 2015;127:942–7. https://doi.org/10.1007/s00508-015-0831-6.

9. Gomez B, Mintegi S, Bressan S, Da Dalt L, Gervaix A, Lacroix L, et al. Validation of the Step-by-Step. Approach Manage Young Febrile Infants Pediatr. 2016;138:e20154381. https://doi.org/10.1542/peds.2015-4381.
10. Esposito S, Rinaldi VE, Argentiero A, et al. Approach to neonates and young infants with fever without a source who are at risk for severe bacterial infection. Mediators Inflamm. 2018;2018:1–11. https://doi.org/10.1155/2018/4869329.
11. Aronson PL, Thurm C, Williams DJ, et al. Association of clinical practice guidelines with emergency department management of febrile infants ≤ 56 days of age. J Hosp Med. 2015;10(6):358–65. https://doi.org/10.1002/jhm.2329.
12. Keitel K, Kilowoko M, Kyungu E, Genton B, D'Acremont V. Performance of prediction rules and guidelines in detecting serious bacterial infections among Tanzanian febrile children. BMC Infect Dis. 2019;19:769. https://doi.org/10.1186/s12879-019-4371-y.
13. Ramgopal S, Horvat CM, Yanamala N, Alpern ER. Machine learning to predict serious bacterial infections in young febrile infants. Pediatrics. 2020;146:e20194096. https://doi.org/10.1542/peds.2019-4096.
14. Lee B, Chung HJ, Kang HM, et al. Development and validation of machine learning-driven prediction model for serious bacterial infection among febrile children in emergency departments. Kakulapati V, editor. PLOS ONE. 2022;17:e0265500. https://doi.org/10.1371/journal.pone.0265500.
15. Fisch U, Kliem P, Grzonka P, et al. Performance of large Language models on advocating the management of meningitis: a comparative qualitative study. BMJ Health Care Inf. 2024;31:e100978. https://doi.org/10.1136/bmjhci-2023-100978.
16. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023;620:172–80. https://doi.org/10.1038/s41586-023-06291-2.
17. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. Nat Med. 31;943–950. https://doi.org/10.1038/s41591-024-03423-7.
18. Chong SL, Niu C, Ong GYK, et al. Febrile infants risk score at triage (FIRST) for the early identification of serious bacterial infections. Sci Rep. 2023;13:15845. https://doi.org/10.1038/s41598-023-42854-z.
19. Lacour AG, Zamora SA, Gervaix A. A score identifying serious bacterial infections in children with fever without source. Pediatr Infect Dis J. 2008;27:654–6. https://doi.org/10.1097/INF.0b013e318168d2b4.
20. Galetto-Lacour A, Zamora SA, Andreola B, et al. Validation of a laboratory risk index score for the identification of severe bacterial infection in children with fever without source. Arch Dis Child. 2010;95:968–73. https://doi.org/10.1136/adc.2009.176800.
21. Khairunnahar L, Hasib MA, Rezanur RHB, et al. Classification of malignant and benign tissue with logistic regression. Inf Med Unlocked. 2019;16:100189. https://doi.org/10.1016/j.imu.2019.100189.
22. Ganesh GA, Ganesh B, Srinivas A. Logistic regression technique for prediction of cardiovascular disease. Glob Transit Proc. 2022;3:127–30. https://doi.org/10.1016/j.gltp.2022.04.008.
23. Kumar PS, K AK, Mohapatra S, Naik B, Nayak J, Mishra M. CatBoost ensemble approach for diabetes risk prediction at early stages. In: 2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON) [Internet]. Bhubaneswar (India): IEEE; 2021 p. 1–6. https://doi.org/10.1109/ODICON50556.2021.9428943.
24. Baghdadi NA, Farghaly Abdelaliem SM, Malki A, et al. Advanced machine learning techniques for cardiovascular disease early detection and diagnosis. J Big Data. 2023;10(1):144. https://doi.org/10.1186/s40537-023-00817-1.
25. Shen Z, Chen H, Wang W, Xu W, Zhou Y, Weng Y, et al. Machine learning algorithms as early diagnostic tools for pancreatic fistula following pancreaticoduodenectomy and guide drain removal: A retrospective cohort study. Int J Surg. 2022;102:106638. https://doi.org/10.1016/j.ijsu.2022.106638.
26. Ghaffarzadeh-Esfahani M, Ghaffarzadeh-Esfahani M, Salahi-Niri A, et al. Large language models versus classical machine learning: performance in COVID-19 mortality prediction using high-dimensional tabular data. arXiv. 2024. https://doi.org/10.48550/arXiv.2409.02136.
27. Zhao X, Wang T, Rios A. Improving expert radiology report summarization by prompting large language models with a layperson summary. arXiv. 2024. https://doi.org/10.48550/arXiv.2409.02136.
28. Nachane SS, Gramopadhye O, Chanda P, Ramakrishnan G, Jadhav KS, Nandwani Y, et al. Few shot chain-of-thought driven reasoning to prompt LLMs for open ended medical question answering. ArXiv. 2024. https://doi.org/10.48550/ArXiv.2403.04890.
29. Louhichi M, Nesmaoui R, Mbarek M, et al. Shapley values for explaining the black box nature of machine learning model clustering. Procedia Comput Sci. 2023;220:806–11. https://doi.org/10.1016/j.procs.2023.03.107.
30. Pantell RH, Roberts KB, Adams WG, et al. Evaluation and management of Well-Appearing febrile infants 8 to 60 days old. Pediatrics. 2021;148:e2021052228. https://doi.org/10.1542/peds.2021-052228.
31. Baker A, Perov Y, Middleton K, et al. A comparison of artificial intelligence and human Doctors for the purpose of triage and diagnosis. Front Artif Intell. 2020;3:543405. https://doi.org/10.3389/frai.2020.543405.
32. Dagan R, Powell KR, Hall CB, et al. Identification of infants unlikely to have serious bacterial infection although hospitalized for suspected sepsis. J Pediatr. 1985;107(6):855–60. https://doi.org/10.1016/s0022-3476(85)80175-x.
33. Ayoub NF, Balakrishnan K, Ayoub MS, et al. Inherent bias in large Language models: A random sampling analysis. Mayo Clin Proc Digit Health. 2024;2(2):186–91. https://doi.org/10.1016/j.mcpdig.2024.03.003.
34. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inf Assoc. 2012;19:121–7. https://doi.org/10.1136/amiajnl-2011-000089.
35. Alowais SA, Alghamdi SS, Alsuhebany N, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23:689. https://doi.org/10.1186/s12909-023-04698-z.
36. Göndöcs D, Dörfler V. AI in medical diagnosis: AI prediction & human judgment. Artif Intell Med. 2024;149:102769. https://doi.org/10.1016/j.artmed.2024.102769.
37. Meng X, Yan X, Zhang K, et al. The application of large Language models in medicine: A scoping review. iScience. 2024;27:109713. https://doi.org/10.1016/j.isci.2024.109713.
38. Clusmann J, Kolbinger FR, Muti HS, et al. The future landscape of large Language models in medicine. Commun Med. 2023;3:141. https://doi.org/10.1038/s43856-023-00370-1.
39. Anisuzzaman DM, Malins JG, Friedman PA, Attia ZI. Fine-Tuning large Language models for specialized use cases. Mayo Clin Proc Digit Health. 2025;3:100184. https://doi.org/10.1016/j.mcpdig.2024.11.005.
40. Schwartz IS, Link KE, Daneshjou R, et al. Black box warning: large Language models and the future of infectious diseases consultation. Clin Infect Dis. 2024;78(4):860–6. https://doi.org/10.1093/cid/ciad633.
41. Lai CH, Mok PKL, Chau WW, et al. Application of machine learning models on predicting the length of hospital stay in fragility fracture patients. BMC Med Inf Decis Mak. 2024;24:26. https://doi.org/10.1186/s12911-024-02417-2.
42. Lundberg S, Lee SI. A unified approach to interpreting model predictions. arXiv. 2017. https://doi.org/10.48550/arXiv.1705.07874.
43. Williams CYK, Zack T, Miao BY, et al. Use of a large Language model to assess clinical acuity of adults in the emergency department. JAMA Netw Open. 2024;7:e248895. https://doi.org/10.1001/jamanetworkopen.2024.8895.
44. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349:255–60. https://doi.org/10.1126/science.aaa8415.
45. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7.
46. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018;29:1836–42. https://doi.org/10.1093/annonc/mdy166.
47. Yang C, Wang X, Lu Y, Liu H, Le QV, Zhou D, et al. Large Language models as optimizers. ArXiv. 2023. https://doi.org/10.48550/ArXiv.2309.03409.
48. Sezgin E. Artificial intelligence in healthcare: Complementing, not replacing, Doctors and healthcare providers. Digit Health. 2023;9:20552076231186520. https://doi.org/10.1177/20552076231186520.

Publisher's note