electronics

MDPI

Article

Integrated AI Medical Emergency Diagnostics Advising System

Sergey K. Aityan [1], Abdolreza Mosaddegh [1], Rolando Herrero [2], Francesco Inchingolo [3], Kieu C. D. Nguyen [3], Mario Balzanelli [4], Rita Lazzaro [4], Nicola Iacovazzo [5], Angelo Cefalo [6], Lucia Carriero [7], Manuel Mersini [7], Jacopo M. Legramante [8], Marilena Minieri [9], Luigi Santacroce [3] and Ciro Gargiulo Isacco [3,*]

[1] Department of Multidisciplinary Engineering, Northeastern University, Oakland Campus, Oakland, CA 94613, USA; s.aityan@northeastern.edu (S.K.A.); a.mosaddegh@northeastern.edu (A.M.)
[2] Smart Everything Innovation Lab (SEIL), College of Engineering, Northeastern University, Boston, MA 02115, USA; r.herrero@northeastern.edu
[3] Department of Interdisciplinary Medicine (DIM), Aldo Moro University of Bari, 70121 Bari, Italy; francesco.inchingolo@uniba.it (F.I.); drkieukaren@gmail.com (K.C.D.N.); luigi.santacroce@uniba.it (L.S.)
[4] Territorial Emergency System SET 118, 74121 Taranto, Italy; mario.balzanelli@gmail.com (M.B.); rita-lazzaro@libero.it (R.L.)
[5] Territorial Center for Medical Assistance, 74121 Taranto, Italy; iacovazzonicola@gmail.com
[6] Department of Hygiene and Prevention, Federico II University of Naples, 80138 Napoli, Italy; ugemel@yahoo.it
[7] Sant'Andrea Longevity Center, 70125 Bari, Italy; lcarriero@gmail.com (L.C.); manuelmersini@gmail.com (M.M.)
[8] Department of Systems Medicine, University of Rome Tor Vergata, 00133 Rome, Italy; legraman@uniroma2.it
[9] Department of Experimental Medicine, University of Rome Tor Vergata, 00133 Rome, Italy; minieri@med.uniroma2.it
* Correspondence: drciroisacco@gmail.com

Abstract: The application of AI (Artificial Intelligence) in emergency medicine helps significantly improve the quality of diagnostics under limitations of resources and time constraints in emergency cases. We have designed a comprehensive AI-based diagnostic and treatment plan decision-support system for emergency medicine by integrating the available LLMs (Large Language Models), like ChatGPT, Gemini, Claude, and others, and tuning them up with additional training on actual emergency cases. There is a special focus on early detection of life-threatening and time-sensitive diseases like sepsis, stroke, and heart attack, which are the major causes of death in emergency medicine. Additional training was conducted on a total of 600 cases (300 sepsis; 300 non-sepsis). The collective capability of the integrated LLMs is much stronger than each individual engine. Emergency cases can be predicted based on information from multiple sensors and streaming sources combining traditional IT (Information Technology) infrastructure with Internet of Things (IoT) schemes. Medical personnel compare and validate the AI models used in this work.

Keywords: AI; artificial intelligence; IoT; Internet of Things; LLMs; large language models; medical emergency; REST; representational state transfer API; sepsis

check for updates

Citation: Aityan, S.K.; Mosaddegh, A.; Herrero, R.; Inchingolo, F.; Nguyen, K.C.D.; Balzanelli, M.; Lazzaro, R.; Iacovazzo, N.; Cefalo, A.; Carriero, L.; et al. Integrated AI Medical Emergency Diagnostics Advising System. *Electronics* **2024**, *13*, 4389. https://doi.org/10.3390/electronics13224389

Academic Editor: Pawel Strumillo

Received: 10 September 2024
Revised: 1 November 2024
Accepted: 2 November 2024
Published: 8 November 2024

1. Introduction

This research work deals with the need for a critical evaluation of the evidence supporting whether a clinical digital solution involving AI (Artificial Intelligence) in home, outpatient, and ambulance settings plays a key role in patient outcomes [1,2]. Specifically, we propose the collection of data from different contexts starting from the home up to the ICU department, both of a clinical and socio-economic nature. This evaluation would concern the living context and social status, age, sex, clinical history, and current condition along with the measurement of various parameters and markers for validation and impact on patient outcomes. To date, few system providers have questioned their products and services in terms of healthcare parameters at home, in the community, and in the ambulance [1].

However, we are well aware that the use of AI in healthcare could possibly raise some concerns, such as liability for errors and the potential to perpetuate and amplify existing biases, data privacy and security, but most importantly the transparency in doctors' decision-making process. We addressed all these issues and created robust data security measures, creating a transparent AI system that constantly upgrades data and procedures and is capable of identifying and correcting biases in AI data and algorithms [3,4].

Although some systems have already been implemented in pre-hospital, emergency room, and intensive care areas, in this article, we have attempted to describe the panorama of applications based on artificial intelligence with a focus on the field of daily emergencies. We processed a total of 600 cases (300 = sepsis; 300 = non-sepsis), obtained from our archive (the University of Medicine Aldo Moro of Bari-Italy and the University of Medicine Tor Vergata of Rome-Italy), Google-Scholar, and PubMed. We also proposed future directions and perspectives [4].

New studies have demonstrated the effectiveness of LLMs in aiding healthcare professionals in generating comprehensive lists of potential diagnoses based on patient symptoms, medical history, and other relevant data. However, our study shows that the collective capability of LLMs in diagnosis is much stronger than each LLM, which identifies the need for an integrated decision-support solution.

1.1. Uncertainty in Medical Diagnostics

In primary care patients, who often present with non-specific symptoms suggestive of an ongoing early disease process, diagnostic uncertainty is a pervasive issue. Often, the signs of the disease are not typical, either symptoms are too common or are not frequently associated with severe disease: Dizziness, for instance, occurs in the majority of cases of stroke, but only very few cases of dizziness result in stroke. Major misdiagnosis-related mistakes are often related to the so-called "Big Three" (major vascular events, infections, and cancers), in which diagnostic uncertainty, resulting in diagnostic errors, is of particular concern [5,6]. In addition, laboratory and diagnostic tests have significant limitations. For instance, laboratory biomarkers such as lactate, C-reactive protein (CRP), and procalcitonin (PCT), which can assist in diagnosis, are not definitive and can be influenced by other conditions, potentially leading to misdiagnosis [7].

Diagnostic errors in first care result from an inadequate amount of information among clinicians, patients, and families. Information certainty is crucial as it is involved in the diagnostic process and it is required for shared decision-making, thus impacting patients' disease evolving progress [5,6]. Effectively managing diagnostic uncertainty can be challenging for doctors given unknown competing priorities and expectations and wide variability exist in the degree to which clinicians engage in first assessment and therapy [6,8]. Although there are protocols for how to proceed and perform in such critical circumstances, few are evidence-based. Recent AI models have elucidated the main traits and the management of the implications of diagnostic uncertainty in primary care [5].

1.2. AI for Medical Diagnostics

The greatest achievement in the use of AI in the medical field is undoubtedly that of being able to carry out tasks that humans sometimes cannot perform with the pace, simplicity, reliability, and diligence that AI can provide, even at a lower cost [9,10].

Healthcare is just one of the many sectors where AI has made great progress in recent years. The potential of artificial intelligence in disease diagnosis has attracted much interest and study. The ability of integrative artificial intelligence to detect earlier signs or symptoms of a particular condition or pathology, for example, is greatly improved by using many algorithms instead of a single algorithm, resulting in better diagnostic accuracy [11,12]. Data have demonstrated how the development of different AI-based approaches helps predict breast cancer recurrence [12]. In this sense, artificial intelligence systems could be of considerable help to doctors on home visits, helping to monitor patients with insulin irregularities, and the first signs related to a possible infection [12] This article will cover the advantages, disadvantages, and ethical issues related to the current role of AI in achieving more specific and predictive diagnoses for disease detection, as well as its potential in the future. The influence of AI on medical diagnostics is expected to expand dramatically as healthcare systems adopt digital technology and AI algorithms continue to advance [12].

1.3. The Structure of This Paper

This paper is structured as follows. In Section 2, we address the distinctive specifics and challenges of emergency medicine that require special attention. Section 3 of this paper describes our vision of solving those challenges using the AI-based system. In Section 4, we present the high-level design of the proposed AI decision-support diagnostic and treatment plan system. The working prototype of the system is described in Section 5. Clinical assessment of the system is presented in Section 6, and Section 7 provides the conclusion and future directions of our research in this domain.

2. Diagnostic Challenges in Emergency Medicine

Artificial intelligence for medical diagnostics is an emerging application in the medical field that already covers various clinical aspects. Since there is no definitive test for sepsis, physicians must rely on a group of tests, experience, and clinical criteria, all with suboptimal performance [4,13]. Acute illness conditions such as infections and sepsis account for many millions of visits to the ED (Emergency Department) in the US every year [14]. The majority of these cases are handled by ED physicians, who must immediately recognize the patient's clinical presentation and interpret the results of diagnostics imaging and various clinical tests to reach the patient's diagnosis [14]. However, the weakest point is determined by the fact that current diagnosing procedures are largely inaccurate or too slow, resulting in dangerous delays or inappropriate treatment. Not only does this harm patients but it furthers the growing and costly global problems of useless therapies and inappropriate procedures [14,15].

Sepsis is a fast-progressing and life-threatening condition that the body's immune system faces while already fighting a severe infection, and it accounts for almost half of all in-hospital deaths [16]. A correct diagnosis requires the presence of an infection and a specific increase in organ dysfunction, all signs that are difficult to diagnose as these parameters are often imprecise with poor specificity resulting in unnecessary hospital resource utilization [14,17].

The history of medicine is deeply intertwined with technological knowledge, so the application of the current techniques is as old as medicine itself. Every era has seen the death of old technological applications for the birth of more accurate applications, so artificial intelligence is predominantly changing the scientific scenario, and medicine is one of those most affected, thanks to the enormous amount of information that is now easily available [18].

With medicine being one of the most appropriate applications of AI, the scientific literature is full of reviews on how the present and future use of tools based on artificial intelligence could positively influence the diagnostic process. Currently, there are more than 20,000 reviews on the use of associated AI data to support diagnostics in the scientific bibliography database, spanning almost 30 years of activity [18]. Diagnostic imaging, interpretation of test results, and clinical evaluation are some of the various aspects that artificial intelligence has already shown its value in [19].

Most international reviews are dedicated to covering the diagnostic process starting from patients in the hospital or clinic, with the unreasonable omission of disease prevention. It is true that AI diagnostic prediction is more accurate if the available data are large enough, but, thanks to the advancement of AI-related technologies, we believe that data collection could be sufficient to evaluate valid support at an early stage before hospital admission, such as at home or on an emergency stretcher in the ambulance [19].

Emergency Diagnostics Under Stress and Time and Resources Constraints

Emergency Medicine represents a crucial part of patient treatment, bridging from the prevention of individual life-threatening diseases to mass epidemic illness management. Whatever the enemy is to face, the emergency care practitioner must fight against time and resource constraints, and as the stake arises so does the stress response. The necessity of making an early diagnosis in these conditions and the resources of human capabilities seem irreconcilable [8].

Beyond the resource limitation, the diagnostic process in an emergency setting is made more difficult by the enormous mole of information that the emergency specialist needs to know. Stroke, myocardial infarction, and sepsis are just a few of the countless conditions that a patient could present during an urgent medical intervention. There is no room for error even if the variety of conditions is vast and the rapidity of action and accuracy of diagnosis must always be maximum.

As reported in the American Systematic Review, although the overall accuracy in Emergency medicine is high, the percentage of incorrect diagnoses is also concerningly elevated (~5.7%). Of consequences, an elevated number of patients will suffer adverse events caused by that (~2.0%), and some of them (~0.3%) are lethal [4]. Moreover, it is stated that clinical error covers all the parts of clinical decision-making, including imaging or test request, the result interpretation, clinical assessment, and illness treatment, particularly in atypical case presentations, or in uncommon pathological conditions [20]. So, in this way, medical malpractice leads to a serious waste of healthcare resources, in terms of hospital stay, cost of treatment, and patients' substandard treatment [20].

The arising of old and new global medical challenges requires the most unshakable determination to guarantee the best clinical management for every patient and all human intellect should give their contribution [8]. As technology advances, the development of new medical tools and their application is taking a big leap, and the use of the most modern devices could fill some of the constraints of everyday Emergency Medicine practice [8].

3. The Vision

The vision of this study is to introduce an integrated AI decision-support system that serves as a real-time consultant for the emergency doctor to provide a second opinion. The integrated AI engines have been comprehensively trained on huge medical datasets and are additionally fine-tuned on real data of emergency cases. These systems support medical decisions by emergency doctors, who make the final decision. In other words, the AI decision-support system does not make final decisions but provides a second opinion and suggests options with the respective probabilities. Medical doctors make the final decision in each case and the final responsibility for their decision remains with them.

3.1. A Decision-Support System

Medical diagnostics is a complex process with multiple uncertainties, obscure symptoms, and observations that also significantly depend on the patient's individuality and health conditions. In simple cases, doctors may give a clear-cut diagnosis while in many moderate-to-hard cases, even experienced doctors may not be immediately clear about the diagnosis and hence about the treatment plan. Different experienced doctors may come up with different diagnoses and treatment plans individually. For this reason, having a second opinion is an important stage in diagnostics in complex cases.

The quality and adequacy of immediate diagnostics in emergency medicine play a crucial role in the recovery of patients, and sometimes it is a decisive factor for the patient's survival. Thus, having an immediate opportunity for a comprehensive second opinion would provide a very important contribution to the emergency doctors, who work under dramatic time constraints and under high psychological stress taking responsibility for the health and survival of their patients.

In our vision, during the patient's assessment, the doctor inputs the information on the patient's symptoms, vital readings, health conditions, general observations, and medical history if it is available. A patient’s medical history may also be retrieved from the integrated patient management system if such a system is available.

In this paper, we describe the medical diagnostics decision-support system that is being developed by the authors. This is an AI-driven diagnostic system that provides comprehensive advice on possible diagnoses in response to the input information provided by the doctor. Possible diagnoses are given together with their probabilities for the particular cases including the explanation of each diagnosis.

Health care is associated with comprehensive liability regulations. For this reason, our system is limited to decision support rather than decision-making. The final decision is made by the emergency doctor who uses some advice from our systems as information support rather than a legitimate diagnosis or treatment plan. This very important issue will be directly addressed in the service agreement with the emergency service that uses our system.

3.2. User Interface and Communication Platform

Computers and communication resources, which doctors typically work with in medical emergency cases, are limited to smartphones, laptops, or personal computers for clients. For this reason, the system is thought to operate on such platforms with the backend on the server or cloud.

Doctors enter input information by using text or voice-to-text capabilities. For some information such as blood pressure, heartbeat rate, and others, doctors may choose a preformatted input user interface and enter the rest of the information as a free-flowing unstructured text or voice. Upon the doctor’s choice, the information may be entered in an unstructured free-flowing way.

The entered information is used for generating the input prompts for the AI diagnostic advisor (decision-support system).

The system output is presented in written form with the list of possible diagnoses together with their probabilities and explanations. The emergency doctor is entitled to make the final determination based on the suggested diagnoses. The system output may be duplicated in voice upon the desire of the emergency doctor.

3.3. Training, Fine-Tuning, and Integration of the Existing AI Solutions

The system is specially trained in the cases supplied by the emergency medicine organization. For the sake of completeness, the diagnostic decision-support system uses the integrated knowledge of all available medical diagnostic AI systems integrated with our solution for knowledge transfer. At the present time, we use the knowledge available at ChatGPT, Claude, and Gemini. We will also integrate other AI diagnostic engines as they become available and integrated with our system.

A distinct feature of our decision-support system is that it is permanently learning from experience. The system training is an ongoing automatic activity. The more cases are run, the more knowledge the system is accumulated and used in future diagnostics.

4. High-Level Design

4.1. An LLM Solution for Medical Emergency Decision-Support System

Large Language Models can be used for the diagnosis of disorders, prescribing treatments, and prediction of future states of patients with greater accuracy than traditional machine learning models, thanks to their broader general knowledge and advanced reasoning capabilities. Timely medical intervention is very important in emergency medicine. Integrating fine-tuned LLMs into medical decision-support systems can accelerate medical decisions by providing immediate decision support and improving diagnoses and treatments by suggesting evidence-based medical advice [21]. Integrating LLMs into clinical decision-support systems can also accelerate diagnosis by providing real-time interpretations of medical images [19]. This helps in identifying critical cases that require immediate attention. Note that this system provides an ongoing assessment using comparative and correlation analysis of the results as well as a confusion matrix. To satisfy the requirements of a decision-support system in emergency medicine, the LLM solution needs to include the following functionalities:

4.1.1. Medical Diagnosis and Treatment

Advanced LLMs are capable of processing vast amounts of structured and unstructured medical data. Some studies have demonstrated the effectiveness of these models in extracting valuable insights from clinical notes and electronic health records for diagnostic purposes. In this regard, LLMs have been applied to provide comprehensive lists of potential diagnoses based on patient symptoms, medical history, and other relevant data. The utilization of LLMs in medical diagnosis enhances diagnostic accuracy, accelerates decision-making processes, and improves treatment outcomes. A study indicates that GPT responses to patient queries are comparable to those provided by doctors [3]. In another study, Med-PaLM 2, a version of PaLM 2 fine-tuned on medical data, has achieved a high accuracy rate in medical diagnosis close to the human clinician level [21].

In our study, GPT, Claude, and Gemini were used as the base models. The models presented reasonable results for diagnosis and primary treatments of emergency patients based on feedback from a group of experienced subject matter experts (Table 1). There are also other pre-trained general and ad hoc medicine-specific models that can be identified and selected as base models during the next phases of this study.

Table 1. Comparing accuracy of responses from AI engines.

Diagnostic Responses	ChatGPT	Gemini	Claude
Correct Responses	209	190	216
Wrong Responses/Need Modifications	91	110	84
Accuracy	69.66%	63.33%	72.00%

We processed a total of 600 cases (300 sepsis; 300 non-sepsis), obtained from our archive (the University of Medicine Aldo Moro of Bari-Italy and the University of Medicine Tor Vergata of Rome-Italy), Google-Scholar, and PubMed.

4.1.2. Interpretation of Medical Images

The interpretation of medical images has a critical role in diagnosing various health conditions and making medical decisions. Using large language models for interpreting medical images is an emerging field that combines advanced image analysis techniques with the language processing and reasoning capabilities of LLMs. Some general-purpose LLMs handle multimodal inputs, including medical images, by integrating them with image-processing models such as convolutional neural networks and transformers. These models are pre-trained on large datasets of medical images such as X-rays and MRIs with their corresponding descriptions. Multimodal LLMs combine text and image processing, enabling the interpretation of medical images using natural language descriptions. LLMs can be fine-tuned for specific tasks like identifying disorders in radiographs or MRI scans, with the textual descriptions providing context and explanations [22]. They can assist in diagnosing health conditions by interpreting images or detecting anomalies in medical images [23]. Some studies show that integrating LLMs with image processing models can produce accurate and clinically relevant results, reducing the workload of medical staff [19].

In our study, GPT, Claude, and Gemini were used for interpretation of images. The evaluation of interpretations by medical doctors shows that GPT outperforms other AI engines (Table 2). However, there are also other pre-trained models on medical images that can be identified and selected for interpretation of images during the next phases of this study.

Table 2. Engine Metrics.

Feature	ChatGPT	Gemini	Claude
Diagnostic			
Accuracy	●	◑	◕
Treatment advice	◕	◔	◑
Imaging	◕	◑	◔
Advice explanation	●	◑	◕
Urgency detection	◕	◑	●
Alternative diagnosis	◕	◕	◕
Performance			
User friendly	●	◑	●
Text entry	●	●	●
Voice entry	◑	◑	◑
Output convenience	●	◑	◕

Table 2 above indicates the degree to which a particular item meets a particular criterion. The comparative metrics of AI engines such as ChatGPT, Gemini, and Claude are shown. The visual representation by means of Harvey balls maps the scores: 0 (0—feature is not available), 1 (1/4 full—poor), 2 (half full—some), 3 (3/4—satisfactory), and 4 (full—complete).

Rapid advances in artificial intelligence based on LLMs have occurred in recent years. Thus, the integration of the existing AI engines with a comprehensive solution for emergency medicine is not yet known in the literature. We utilize the extraordinary efforts made by the developers of such solutions as GPT, Claude, and Gemini to build up an AI-based intelligent solution on top of the existing solutions.

4.1.3. Transfer Learning in Emergency Medicine Domain

Transfer learning leverages the general knowledge embedded in pre-trained LLMs using domain-specific data. This process allows models to improve accuracy in medical contexts. Fine-tuning large language models can adapt general-purpose models trained on large datasets to specific tasks in the medicine domain by further training them on medical data. Reference [24] evaluated four LLMs with and without fine-tuning for diagnosis of disorders. The results highlight that fine-tuned LLMs significantly outperform general LLMs in medical diagnosis. Fine-tuned LLMs presented significant results in analyzing patient symptoms and medical data. A study shows that these models can achieve diagnostic accuracy comparable to that of human doctors in some contexts [23].

Although LLMs were fine-tuned for various medical fields, there is no study in the existing literature on the application of fine-tuned LLMs in emergency medicine. In our study, a transfer learning approach on pre-trained LLMs is used by employing domain-specific medical emergency data. This process allows general-purpose LLMs to specialize in diagnosis and treatments in emergency medicine.

4.1.4. Human-Like Interaction with Medical Staff

One of the most useful applications of AI in medical decision-support systems is in providing human-like interactions with medical staff, which improves operational efficiencies. AI can simulate human-like interactions with medical staff primarily by utilizing natural language processing (NLP), and speech recognition techniques. These capabilities enable interactive systems that can understand and respond to queries and facilitate interactions with medical staff. LLMs can be used to process unstructured text data such as clinical notes besides health records and other relevant data and generate human-like text. They can process and respond to natural language queries from healthcare providers, helping to clarify symptoms, diagnose disorders, and recommend tests or primary treatments. LLMs such as GPT have been employed to interpret medical inquiries and generate clinically relevant responses [25]. Some studies show that conversational AI platforms present reasonable performance in understanding medical terminology and provide accurate medical information based on queries [26].

Technologies like Google's Speech-to-Text and Microsoft's Azure Cognitive Services enable AI systems to process spoken language. This capability is essential for voice-enabled virtual assistants that can interact with medical staff in real time [27].

In our solution, LLMs have been employed for interaction with medical staff using free format text. In addition, speech-to-text capability is used to facilitate interaction with medical staff in emergency situations.

4.1.5. Data Management

High-quality, domain-specific datasets are essential for effective fine-tuning and data preprocessing is often required to ensure high- quality and consistency in data. Effective data management provides accurate and high-quality data required for evidence-based recommendations for diagnosis and treatment. Data warehouses allow for efficient maintenance and preprocessing of large volumes of structured and unstructured data. In this regard, cloud-based data warehouses offer scalable and secure options for managing medical data as they provide robust storage, computing power, and accessibility [28].

In our solution, all structured and unstructured data that will be used for fine-tuning and production phases should be maintained in a cloud-based data warehouse. Information on each case, corresponding predictions from fine-tuned models, and the real outcome of the case will be recorded. This information can be used in different phases of the study and for future research and publications.

The diagnostic responses of various AI engines (LLMs such as ChatGPT, Claude, and Gemini) in emergency health conditions were evaluated as significant outcomes by medical doctors in emergency medicine. From 300 cases of medical emergency, only 43 cases (13.66%) were diagnosed wrong by all AI engines, however; in 159 cases (53%) some AI engines provide wrong diagnosis. Claude offers the best performance of an individual engine with 84 wrong diagnoses (28%). The results indicated that none of the LLMs is capable of providing a high accuracy rate suitable for emergency medicine, however, the collective capability of LLMs in diagnosis is much stronger than each LLM. This identifies the need for integrated decision-support solutions that can take advantage of the collective capabilities of LLMs (Table 1).

4.2. Training from Scratch vs. Fine-Tuning

Training LLMs from scratch requires huge computational resources and time to process large datasets and adjust millions of parameters. One of the primary advantages of fine-tuning over training from scratch is the significant reduction in computational resources. Pre-trained LLMs, such as GPT, have already undergone extensive training on large datasets, involving substantial computational resources. Fine-tuning these models for specific tasks requires adjusting only a subset of the parameters, which is computationally less expensive. Training advanced LLMs from scratch is typically feasible only for large organizations with significant financial and computational capabilities [29]. On the other hand, the risk associated with training LLMs from scratch is significantly higher compared to fine-tuning since it needs considerable resources, with no guarantee of achieving the desired outcomes [30]. Fine-tuning reduces this risk by employing pre-trained models, which have already demonstrated their efficiency across various tasks.

Data requirement is another important aspect of training. Training an LLM from scratch needs vast amounts of data to achieve the required level of performance and generalization. It can be difficult to achieve high performance on specialized tasks without considerable domain-specific data. In contrast, fine-tuning requires smaller domain-specific datasets. The pre-trained model inherits training on general patterns which reduces the need for huge training data. This is more important for applications in domains where large datasets are not available.

Finally, fine-tuning often results in superior performance compared to models trained from scratch even on a specific task. This is because fine-tuned models benefit from the vast amounts of knowledge embedded in their pre-trained parent models. This avoids overfitting since transferring learned knowledge from the pre-trained model enhances the model's generalization ability when comparing models trained on limited task-specific data [31].

Regarding the advantages of fine-tuning over training from scratch, we used a fine-tuning method in our solution to align LLMs with the requirements of a medical emergency decision-support system.

4.3. Fine-Tuning Methods

Although pre-trained LLMs take advantage of vast knowledge bases, they lack specialization in specific domains. Fine-tuning addresses this limitation by allowing the model to learn from domain-specific data to make it more accurate and effective for target applications. Emergency medicine has its own unique patterns, terminologies, and context. Fine-tuning pre-trained LLMs allows these models to align with the specific requirements of emergency medicine.

In our study, real-world emergency data and knowledge from healthcare professionals are used to fine-tune LLMs. Unlike general-purpose LLMs trained on available information on the internet and feedback from users, including many incomplete and fake information, our fine-tuned models will be trained on emergency-specific data and scientific decisions made by knowledgeable doctors to provide maximum accuracy in the emergency medicine field.

There are different approaches to fine-tuning LLMs. Full fine-tuning involves updating all the parameters of a pre-trained model on a new task-specific dataset. The authors of [32] demonstrated this approach with BERT where the model is pre-trained on a large dataset and then fine-tuned on a specific task by adjusting all its weights. This method usually provides a high level of alignment but can be computationally expensive. In this regard, some studies focused on reducing computational resources. Feature-based fine-tuning uses the pre-trained model as a fixed feature extractor without updating parameters and the output is fed into a task-specific model [31]. This method is less resource-intensive; however, it might not fully leverage the model's capabilities for the new task. A few studies focused on parameter-efficient fine-tuning methods that aim to reduce the number of parameters updated during fine-tuning. The authors of [33] proposed fine-tuning only on a small subset of parameters, such as bias terms, which can significantly lower the computational overhead.

One major problem of fine-tuning is overfitting to task-specific datasets especially if the dataset is small. Multi-task fine-tuning can avoid overfitting by fine-tuning the pre-trained model on multiple tasks simultaneously. Some study [34] shows that training on a mixture of tasks improves generalization and robustness. For general LLMs trained on diverse domains and tasks, prompt-based fine-tuning can also be a good option. Prefix-tuning and prompt-based fine-tuning adjust the input prompts or prefixes fed to the model rather than modifying the model parameters directly. Prefix-tuning by adding prefixes to the input sequence, allowed the model to adapt to specific tasks [35]. Unlike traditional fine-tuning where the model's parameters are adjusted, prompt-based fine-tuning involves minimal or no changes to the model's weights. Instead, it relies on designing effective queries that result in desired outputs from the model. This method leverages the pre-trained model's existing knowledge and can be efficient for LLMs trained on various tasks and knowledgebases; However, designing effective prompts and queries can be challenging and often requires domain knowledge. Also, some domains may require additional parameter adjustments to achieve optimal performance.

In our solution, a mixed method of fine-tuning is used. We developed prompt-based fine-tuning for each LLM to receive optimum domain-specific responses based on the requirements of emergency medicine. The free-text format queries from medical staff consisting of health conditions, and symptoms of patients are fine-tuned to obtain optimal information for emergency medicine from LLMs on most probable diagnoses with their corresponding probabilities, rationales, urgencies, and treatments. The recommendations by the solution for intervention and treatment of emergency patients were evaluated as significant outcomes by a group of subject matter experts. These recommendations were aligned with standard guidelines and scientific evidence from publications in international medical journals.

In the next step, the results will be used for full fine-tuning of the model on emergency-specific data using a low learning rate. Full fine-tuning is particularly beneficial when enough computational resources are available. Comparing feature extraction methods, full fine-tuning allows the whole model to learn from the task-specific data and leads to a more profound adaptation of the model to emergency medicine, which provides superior performance.

For this purpose, the LLMs will be trained on a domain-specific labeled dataset, where each input data point is associated with a real outcome of the case. Supervised fine-tuning can significantly improve the model's performance on the task, making it an efficient method for customizing LLMs. The results of queries are fine-tuned by prompt engineering from each LLM, and the real outcomes of each case will be used for this phase of fine-tuning. Healthcare professionals in emergency medicine will compare the responses of engines with the correct diagnosis and treatments for each case. A multi-task approach is used since the model is fine-tuned on the diagnosis dataset and treatment dataset separately. Also, Dynamic learning will be used to extend the training of the model after initial fine-tuning. By continuously learning from new data and feedback from healthcare professionals, dynamic fine-tuning can improve the overall accuracy and performance of the model.

5. Working Prototype

5.1. Design of the Prototype

As proof of concept, our team developed a prototype to use LLM engines of ChatGPT, Claude, and Gemini for diagnosis problems of emergency patients and suggesting primary treatments. Clinical markers, state of consciousness and alertness, skin color, ocular reflex, nail refilling, pain, tiredness, drowsiness, nausea, vomiting, nausea, presence of abnormal spots on the skin, presence of fever or hypothermia, the patient's clinical history, blood pressure, oximeter saturation, heart rate, respiratory rate obtained in the area, at home and/or from the general practitioner; the electrocardiogram ECG trace, the blood gas, the markers derived from the blood count obtained in the ambulance and emergency room; and microbiological culture obtained in the emergency room and medical images are employed as system inputs using free format text, images and speech.

The prototype is developed using Python programming language. The architectural design consists of three layers as shown in Figure 1:

- **Business Layer:** The business layer contains rules, logic, and services for providing medical emergency decision support using the capabilities of LLMs.
- **Presentation Layer:** The presentation layer houses the user interface including speech-to-text and translation capabilities for interaction with medical staff.
- **Data Access Layer:** The data layer manages the interaction with the database to maintain data related to each case and corresponding responses from LLMs.

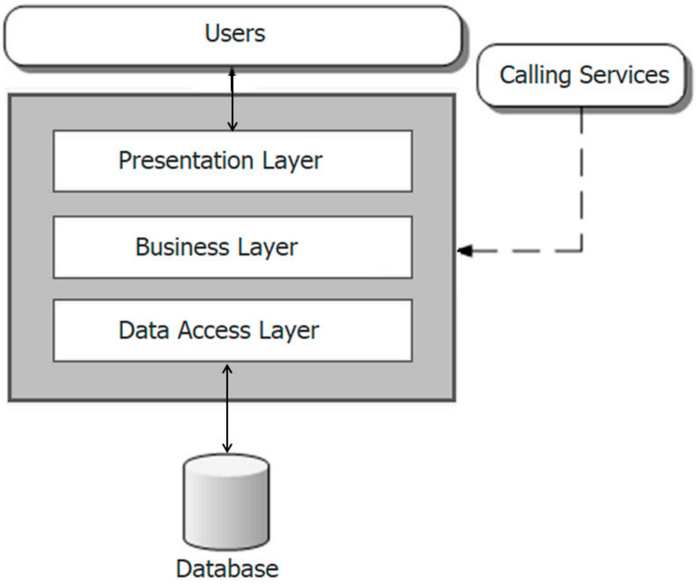


Figure 1. Architectural design of prototype.

These layers mentioned above collectively provide the following functionalities:

- Medical advice
- Image interpretation
- User interface
- Data warehouse

5.2. Medical Advice

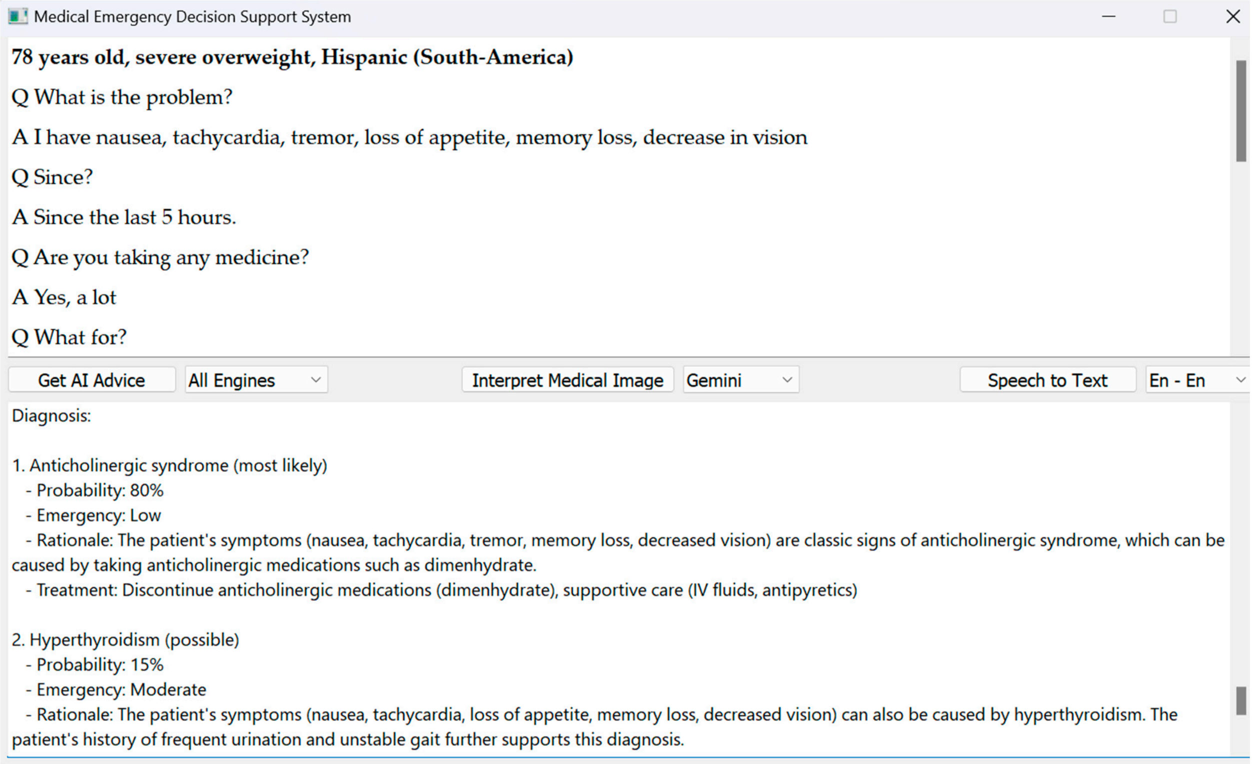
Medical advice on patient health conditions is provided using prompt-based fine-tuned LLMs. GPT-4o is used as one of the LLM engines for medical advice. GPT-4o is one of the most advanced models of OpenAI with the same capabilities as GPT 4-Turbo, but it is more efficient and faster. It accepts both text and image inputs. We also employed the Claude 3-opus engine for medical advice. Claude also handles a wide variety of tasks involving language, reasoning, analysis, and image interpretation. As of now, Opus is the most powerful model of Claude. The third LLM engine for medical advice is Google Gemini 1.5-Pro. It is a multimodal LLM with an updated context window of up to two million tokens.

5.3. Image Interpretation

Interpretation of medical images is also provided using prompt-based fine-tuned LLMs capable of processing image inputs. GPT-4o is used as one of the engines for image interpretation. Up until now, GPT-4o has shown the best image processing performance across OpenAI models. We also employed the Claude 3-opus engine for image interpretation. Opus is the state-of-the-art vision model of Claude that can generate human-like text outputs based on both image and text inputs. The third engine used for image interpretation is Google Gemini 1.5-flash. It is a lightweight and fast AI engine with multimodal reasoning and a context window of up to one million tokens.

5.4. User Interface

NLP capabilities of selected LLMs have been employed for interaction with medical staff using free-format text. In addition, the speech-to-text capability is used to facilitate interaction with medical staff in emergency situations. Google Speech to Text is used for converting speech input to text and Google translation is used for translation between different languages to facilitate the communication of medical personnel with the system in different countries (Figure 2).



78 years old, severe overweight, Hispanic (South-America)

Q What is the problem?

A I have nausea, tachycardia, tremor, loss of appetite, memory loss, decrease in vision

Q Since?

A Since the last 5 hours.

Q Are you taking any medicine?

A Yes, a lot

Q What for?

Diagnosis:

1. Anticholinergic syndrome (most likely)
   - Probability: 80%
   - Emergency: Low
   - Rationale: The patient's symptoms (nausea, tachycardia, tremor, memory loss, decreased vision) are classic signs of anticholinergic syndrome, which can be caused by taking anticholinergic medications such as dimenhydrate.
   - Treatment: Discontinue anticholinergic medications (dimenhydrate), supportive care (IV fluids, antipyretics)

2. Hyperthyroidism (possible)
   - Probability: 15%
   - Emergency: Moderate
   - Rationale: The patient's symptoms (nausea, tachycardia, loss of appetite, memory loss, decreased vision) can also be caused by hyperthyroidism. The patient's history of frequent urination and unstable gait further supports this diagnosis.

Figure 2. User interface in the prototype.

5.5. Data-Warehousing

The decision-support system is capable of handling structured and unstructured data of each case and corresponding predictions from LLMs. A data warehouse maintains anonymized data of patients and interaction with AI engines. This data can be used in dynamic fine-tuning of the model. Also, the information can be used for publication and future phases of the project. In this prototype, Microsoft SQL Server is used for data warehousing.

6. Assessment of the Prototype

6.1. Medical Assessment

In the pre-hospital phase, AI can support family, community, and ambulance doctors in making decisions right before ED admission [36]. By analyzing patient data, such as health status, medical history, current symptoms with vital and physical signs, and the patient's medical history, AI can make a strong contribution to the management before even arriving at the ED to choose the most appropriate course of action [36]. In this project, we studied and developed an AI model capable of helping the doctor quickly and effectively, leading toward the most valuable diagnostic paths and making decisions on therapeutic priorities [36,37].

Infectious diagnostics as well as cardiovascular, imaging, and neurological diagnostics are areas in which AI has shown great potential. AI is capable of analyzing complex data patterns, such as electrocardiogram recordings, medical imaging (MRI and CT), and data to identify potentially lethal conditions such as a heart attack or sepsis [36,38]. Our project considered the use of the large dataset of ChatGPT, Claude, Gemini, and Google for medical diagnostics. Through continuous and constant evaluation procedures, our AI increasingly seeks answers and compatible solutions to questions related to states of emergency at home, territorial medical centers, and ambulances.

The currently available AI algorithms can analyze complex medical data with unprecedented precision and speed. In summary, chatbots such as ChatGPT, Claude, Gemini, and Google have shown a lot of potential in medicine [36,38]. Nevertheless, our study showed that chatbots could be considered valuable sources of information, and the responses generated are credible and comprehensive and could be a strong tool in helping doctors formulate more precise diagnoses and proceed with earlier therapies. We proposed that a crucial point for effectively exploiting this technology is close collaboration between medical professionals and artificial intelligence experts [39]. The responses from different integrated AI engines for different diseases and different disease categories may vary. Also, the accuracy of each integrated AI engine may depend on the specifics of the input prompts. This challenge will be addressed by the proprietary intelligent components of our system, which are currently under development and will be described in detail in our future publications.

6.2. Comparison with Other Solutions

Comparative analysis of our AI-based system with the existing AI engines can be performed using different metrics.

Our prototype employs fine-tuned large language models that integrate diverse healthcare data, leading to highly accurate prediction outcomes. In contrast, traditional scoring systems like the Sequential Organ Failure Assessment and the Systemic Inflammatory Response Syndrome criteria have been widely used but often suffer from lower accuracy due to their reliance on a limited set of clinical parameters [40]. Machine learning models such as the Modified Early Warning Score also provide good predictive performance but may not capture the full complexity of sepsis [41].

Another aspect that characterizes our prototype is the speed: by leveraging advanced LLMs and real-time data processing, our system can quickly analyze large volumes of data and provide rapid risk assessments, whereas many current AI engines may require significant time-consuming preprocessing [42].

Scalability is another metric that makes our approach innovative, compared with existing technologies. The design of our data collection framework and the use of cloud-based infrastructure ensure that our system can scale to handle large datasets and multiple healthcare settings. This makes it suitable for widespread deployment, whereas many current models are designed for specific datasets or settings and may struggle to scale effectively without significant modification [43].

One widely studied AI tool is the Epic Sepsis Model, integrated with Epic's electronic medical record software. This model generates sepsis risk estimates every 20 min for hospitalized patients, but it struggles to differentiate between high and low-risk patients before they receive treatment. According to recent findings, the Epic Sepsis Model correctly identified high-risk patients 87% of the time when using all available data but only 62% before patients met clinical criteria for sepsis, and just 53% before a blood culture was ordered. This suggests that the model heavily relies on clinical-initiated diagnostic tests, reducing its effectiveness as an early warning system.

Our system also addresses a critical gap highlighted in the literature: the need for immediate actionable insights. Unlike the Epic Sepsis Model, which often cues in diagnostic tests and treatments already initiated by clinicians, our system continuously updates risk assessments and suggests proactive treatment protocols based on a patient's evolving condition. This dynamic interaction is vital for improving clinical outcomes and aligns with findings from studies that emphasize the importance of real-time data processing and intervention. In response to benchmarks becoming old very quickly, some of the newer databases are constantly being updated with new and relevant data points. This is why AI models have technically not yet managed to match an optimal performance in some areas such as sepsis diagnosis, even though they are on the way.

Using data from our archive, Google Scholar, and PubMed, we visualized how quickly our AI prototype has started to beat database benchmarks such as ChatGpt, Gemini, and Claude, as well as whether it yet reached human levels of skill. Each AI model was devised around specific skills, like diagnosis accuracy, treatment advice emergency detection, voice recognition, language understanding, or imaging readings. Each score contrasts with the following benchmarks: 0 maximally performing baseline to 4, equal to the highest AI performance on the current dataset. By creating a scale between these points, the progress of AI models on each dataset could be tracked. Each point on a line means the best result and as the line trends upwards, our AI model gets closer and closer to matching the best performance which is represented now by ChatGPT (Table 2 above).

7. Conclusions and Future Directions

In this paper, we introduced a software infrastructure to prototype the integration of three main LLM engines to support the accurate prediction of medical diagnostics and treatment. Looking at Table 1, the ChatGPT model outperformed those of Gemini and Claude for most features. Specifically, ChatGPT was more accurate, providing better treatment advice and explanations of that advice than the other two engines. The Claude LLM model was better at detecting urgencies than ChatGPT and Gemini, but all three models show similar capabilities when providing alternative diagnoses. From a user interaction point of view, both ChatGPT and Claude performed similarly, with both outperforming Gemini in user-friendliness. In all, ChatGPT appears to provide the best base model to support the successful prediction of medical diagnostics and treatment. These results can be expanded by supporting other engines and relying on extensive fine-tuning to accomplish better accuracy.

Combining the results of LLMs fine-tuned for emergency medicine is a challenging task that can be addressed by a few advanced AI techniques. Multi-agent AI (MAAI) involves systems where multiple autonomous AI agents interact within an environment. These agents can be cooperative or competitive and they work to achieve collective goals. Interaction protocols define how agents communicate and collaborate with each other.

One of the applications of MAAI in medicine is in the diagnosis of diseases. Some studies [44] show that multi-agent systems can significantly enhance diagnostic accuracy by enabling collaborative decision-making among agents, each specializing in different diagnostic tasks.

Cooperative Multi-agent AI can leverage the capabilities of various fine-tuned LLMs specializing in processing specific types of medical data such as lab results, medical images, and patient history to provide a comprehensive view of patients' health conditions. Also, various fine-tuned LLMs can focus on various aspects of medical decisions such as diagnosis, treatment, required tests, etc. Each instance of LLMs can be fine-tuned and specialized in one specific task in emergency medicine and collaborate with other agents using Multi-Agent Coordination techniques, which ensure that agents can work together and make joint decisions effectively.

One other technique to combine the results of fine-tuned LLMs is stacking individual models to create a super-learner. Stacking involves training multiple base models and a meta-model that teaches them to combine their outputs. Combining multiple models to leverage their collective strengths often leads to improved accuracy and better performance compared to individual models.

Some studies show the ability of stacking methods to improve the accuracy of prediction and diagnosis of diseases. The authors of [45] used stacking of different neural network models to predict sepsis and the model achieved robust and interpretable predictions.

Results of LLMs fine-tuned in various aspects of emergency medicine can be integrated using a super-learner. By combining multiple fine-tuned LLMs using a meta-learner, the resulting model is less likely to be affected by the weaknesses of any single LLM; however, it takes advantage of the strengths of fine-tuned LLMs specialized in different tasks.

The limitations of our approach may include an insufficient amount of actual emergency cases for additional training of the integrated LLMs and differences in the medical practices and regulations in different countries. These limitations can be addressed by conducting comprehensive case collection from the participating emergency services and special fine-tuning for different countries.

Further research is needed to develop frameworks to realize multi-agent systems and super-learners from fine-tuned LLMs. The integration issues and interoperability are challenges that need to be addressed by these systems. Integrating these systems with emerging technologies like the Internet of Medical Things (IoMT) could further enhance the capabilities of medical decision-support systems in emergency medicine. Our future plans include additional proprietary AI-based components to manage the variability and accuracy of responses from different integrated AI engines.

Author Contributions: Conceptualization, S.K.A.; Methodology, S.K.A., A.M., R.H. and C.G.I.; Software, A.M. and R.H.; Validation, S.K.A., A.M., R.H., N.I. and M.M. (Manuel Mersini); Formal analysis, A.M., R.H. and C.G.I.; Investigation, A.M., R.H., K.C.D.N., N.I., A.C., M.M. (Manuel Mersini) and C.G.I.; Resources, F.I., M.B., R.L., L.C., M.M. (Manuel Mersini), J.M.L., M.M. (Marilena Minieri), L.S. and C.G.I.; Data curation, R.H., N.I. and C.G.I.; Writing—original draft, C.G.I.; Writing—review and editing, S.K.A. and C.G.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: The authors would like to thank graduate students Jiang Han, Vikram Sawant, Shuhan Ji, and Jing Cao for their contribution and help in the project.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lewis, T.L.; Wyatt, J.C. mHealth and mobile medical apps: A framework to assess risk and promote safer use. *J. Med. Internet Res.* **2014**, *16*, e210. [CrossRef] [PubMed]
2. Mathews, S.C.; McShea, M.J.; Hanley, C.L.; Ravitz, A.; Labrique, A.B.; Cohen, A.B. Digital health: A path to validation. *npj Digit. Med.* **2019**, *2*, 38. [CrossRef] [PubMed]

3. Ayers, J.W.; Poliak, A.; Dredze, M.; Leas, E.C.; Zhu, Z.; Kelley, J.B.; Faix, D.J.; Goodman, A.M.; Longhurst, C.A.; Hogarth, M.; et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **2023**, *183*, 589–596. [CrossRef] [PubMed]
4. Newman-Toker, D.E.; Peterson, S.M.; Badihian, S.; Hassoon, A.; Nassery, N.; Parizadeh, D.; Wilson, L.M.; Jia, Y.; Omron, R.; Tharmarajah, S.; et al. *Diagnostic Errors in the Emergency Department: A Systematic Review [Internet]*; Report No.: 22(23)-EHC043; Agency for Healthcare Research and Quality (US): Rockville, MD, USA, 2022. [PubMed]
5. Persson, I.; Östling, A.; Arlbrandt, M.; Söderberg, J.; Becedas, D. A Machine Learning Sepsis Prediction Algorithm for Intended Intensive Care Unit Use (NAVOY Sepsis): Proof-of-Concept Study. *JMIR Form Res.* **2021**, *5*, e28000. [CrossRef]
6. Kroenke, K.; Mangelsdorff, A. Common symptoms in ambulatory care: Incidence, evaluation, therapy, and outcome. *Am. J. Med.* **1989**, *86*, 262–266. [CrossRef]
7. Chen, L.M.; Zheng, W.M.; Dong, X.M.; Zheng, Y.M.; Shi, W.M.; Zhang, W. Analysis of misdiagnosed or delayed-diagnosed Leprosy bacillus infection from 1990 to 2020 with a prophet time series prediction in Hubei Province, China. *Medicine* **2023**, *102*, e34714. [CrossRef]
8. Abdel-Razig, S.; Stoller, J.K. Global "systemness" in medical education: A rationale and framework to assess performance. *Med. Teach.* **2023**, *45*, 1431–1435. [CrossRef] [PubMed]
9. Sqalli, M.T.; Al-Thani, D. AI-supported health coaching model for patients with chronic diseases. In Proceedings of the 16th International Symposium on Wireless Communication Systems 2019, Oulu, Finland, 27–30 August 2019; pp. 452–456.
10. Zhou, L. A rapid, accurate and machine-agnostic segmentation and quantification method for CT-Based COVID-19 diagnosis. *IEEE Trans. Med. Imaging* **2020**, *39*, 2638–2652. [CrossRef]
11. Delen, D.; Walker, G.; Kadam, A. Predicting breast cancer survivability: A comparison of three data mining methods. *Artif. Intell. Med.* **2005**, *34*, 113–127. [CrossRef]
12. Dahm, M.R.; Cattanach, W.; Williams, M.; Basseal, J.M.; Gleason, K.; Crock, C. Communication of Diagnostic Uncertainty in Primary Care and Its Impact on Patient Experience: An Integrative Systematic Review. *J. Gen. Intern. Med.* **2023**, *38*, 738–754. [CrossRef]
13. Rhodes, K.V.; Pollock, D.A. The future of emergency medicine public health research. *Emerg. Med. Clin. N. Am.* **2006**, *24*, 1053–1073. [CrossRef] [PubMed] [PubMed Central]
14. Mostafa, R.; El-Atawi, K. Strategies to Measure and Improve Emergency Department Performance: A Review. *Cureus* **2024**, *16*, e52879. [CrossRef] [PubMed] [PubMed Central]
15. Sørup, C.M.; Jacobsen, P.; Forberg, J.L. Evaluation of emergency department performance—A systematic review on recommended performance and quality-in-care measures. *Scand. J. Trauma Resusc. Emerg. Med.* **2013**, *21*, 62. [CrossRef] [PubMed]
16. Horeczko, T.; Green, J.P.; Panacek, E. Epidemiology of the Systemic Inflammatory Response Syndrome (SIRS) in the emergency department. *West. J. Emerg. Med.* **2014**, *15*, 329–336. [CrossRef]
17. Coburn, B.; Morris, A.M.; Tomlinson, G.; Detsky, A.S. Does this adult patient with suspected bacteremia require blood cultures? *JAMA* **2012**, *308*, 502–511. [CrossRef]
18. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2017**, *2*, 230–243. [CrossRef]
19. Zhang, D.; Liu, X.; Shao, M.; Sun, Y.; Lian, Q.; Zhang, H. The value of artificial intelligence and imaging diagnosis in the fight against COVID-19. *Pers. Ubiquitous Comput.* **2023**, *27*, 783–792. [CrossRef]
20. Ronicke, S.; Hirsch, M.C.; Türk, E.; Larionov, K.; Tientcheu, D.; Wagner, A.D. Can a decision support system accelerate rare disease diagnosis? Evaluating the potential impact of Ada DX in a retrospective study. *Orphanet J. Rare Dis.* **2019**, *14*, 69. [CrossRef]
21. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. Towards expert-level medical question answering with large language models. *arXiv* **2023**, arXiv:2305.09617.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
23. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef] [PubMed]
24. Yadav, M.; Sahu, N.K.; Chaturvedi, M.; Gupta, S.; Lone, H.R. Fine-tuning Large Language Models for Automated Diagnostic Screening Summaries. *arXiv* **2024**, arXiv:2403.20145.
25. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165.
26. Laumer, S.; Maier, C.; Gubler, F.T. Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis. In Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden, 8–14 June 2019.
27. Xie, Q.; Liu, Z.; Yuille, A.; Tu, Z. Self-training with noisy student improves ImageNet classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10687–10698.
28. Rittinghouse, J.W.; Ransome, J.F. *Cloud Computing: Implementation, Management, and Security*; CRC Press: Boca Raton, FL, USA, 2017.
29. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling laws for neural language models. *arXiv* **2020**, arXiv:2001.08361.

30. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog* **2019**, *1*, 9.
31. Peters, M.E.; Ruder, S.; Smith, N.A. To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. *arXiv* **2019**, arXiv:1903.05987.
32. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
33. Rebuffi, S.A.; Bilen, H.; Vedaldi, A. Learning multiple visual domains with residual adapters. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 506–516.
34. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
35. Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 1–6 August 2021; pp. 4582–4597.
36. Aleksandra, S.; Robert, K.; Klaudia, K.; Dawid, L.; Mariusz, S. Artificial Intelligence in Optimizing the Functioning of Emergency Departments; a Systematic Review of Current Solutions. *Arch. Acad. Emerg. Med.* **2024**, *12*, e22.
37. E Annis, I.; Jordan, R.; Thomas, K.C. Quickly identifying people at risk of opioid use disorder in emergency departments: Trade-offs between a machine learning approach and a simple EHR flag strategy. *BMJ Open* **2022**, *12*, e059414. [CrossRef]
38. Casano, N.; Santini, S.J.; Vittorini, P.; Sinatti, G.; Carducci, P.; Mastroianni, C.M.; Ciardi, M.R.; Pasculli, P.; Petrucci, E.; Marinangeli, F.; et al. Application of machine learning approach in emergency department to support clinical decision making for SARS-CoV-2 infected patients. *J. Integr. Bioinform.* **2023**, *20*, 20220047. [CrossRef] [PubMed]
39. Mueller, B.; Street, W.N.; Carnahan, R.M.; Lee, S. Evaluating the performance of machine learning methods for risk estimation of delirium in patients hospitalized from the emergency department. *Acta Psychiatr. Scand.* **2023**, *147*, 493–505. [CrossRef] [PubMed]
40. Singer, M.; Deutschman, C.S.; Seymour, C.W.; Shankar-Hari, M.; Annane, D.; Bauer, M.; Bellomo, R.; Bernard, G.R.; Chiche, J.-D.; Coopersmith, C.M.; et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* **2016**, *315*, 801–810. [CrossRef]
41. Churpek, M.M.; Adhikari, R.; Edelson, D.P. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation* **2016**, *102*, 1–5. [CrossRef]
42. Desautels, T.; Calvert, J.; Hoffman, J.; Jay, M.; Kerem, Y.; Shieh, L.; Shimabukuro, D.; Chettipally, U.; Feldman, M.D.; Barton, C.; et al. Prediction of Sepsis in the Intensive Care Unit with Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Public Health Surveill.* **2016**, *4*, e28. [CrossRef]
43. A Goldstein, B.; Navar, A.M.; Pencina, M.J.; A Ioannidis, J.P. Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *J. Am. Med. Inform. Assoc.* **2017**, *24*, 198–208. [CrossRef]
44. Iqbal, S.; Altaf, W.; Aslam, M.; Mahmood, W.; Khan, M.U.G. Application of intelligent agents in health-care. *Artif. Intell. Rev.* **2016**, *46*, 83–112. [CrossRef]
45. El-Rashidy, N.; Abuhmed, T.; Alarabi, L.; El-Bakry, H.M.; Abdelrazek, S.; Ali, F.; El-Sappagh, S. Sepsis prediction in intensive care unit based on genetic feature optimization and stacked deep ensemble learning. *Neural Comput. Appl.* **2022**, *34*, 3603–3632. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.