Large Language Models for Knowledge Discovery in Healthcare

Shadi AlZu'bi, Tarek Kanan
Faculty of Sciences and Information Technologies
Alzaytoonah University of Jordan
Amman, Jordan
smalzubi@zuj.edu.jo, tarek.kanan@zuj.edu.jo

Muder Almiani
Accounting and MIS Department
Gulf University of Science and Technology
Kuwait
Almiani.m@gust.edu.ku

***Abstract*—The integration of artificial intelligence in healthcare has the potential to revolutionize patient care. This paper explores the integration of fine-tuned Large Language Models (LLMs) with ML and DL paradigms to enhance knowledge discovery in the healthcare sector. The models can improve the extraction, processing, and utilization of healthcare data. A comprehensive case study showcases the application of LLMs in diagnostic imaging, patient readmission prediction, and personalized treatment recommendations. Experimental results reveal significant improvements: diagnostic accuracy increased from 85% to 92%, patient readmission prediction accuracy rose from 78% to 85%, and personalized treatment recommendation accuracy improved from 82% to 90%. These findings suggest that the fusion of LLMs with DL techniques can revolutionize healthcare knowledge discovery, leading to more informed decision-making and personalized patient care.**

***Index Terms*—Knowledge Discovery, Deep Learning, LLM, Healthcare Analytics, Smart Globe**

I. Introduction

Artificial intelligence (AI) has seen widespread adoption in a variety of fields. In recent years, extensive research has highlighted the potential of AI across many applications, including Social Media [1], Sustainable Environment [2], Agriculture [3], and healthcare [4]. In the medical field, AI has been utilized to improve diagnostic accuracy, forecast patient outcomes, and enhance the efficiency of administrative tasks [5], [6]. Techniques such as machine learning (ML) and deep learning (DL) have been particularly effective in processing elaborate medical data and generating valuable insights. Despite these advancements, there is still immense potential to further integrate AI, especially with the emergence of large language models (LLMs), to explore new avenues in healthcare knowledge discovery [7].

Knowledge discovery has become a cornerstone in the era of big data, particularly within the realm of healthcare, where vast amounts of data are generated [8], [9]. ML and DL have been instrumental in analyzing this data to extract meaningful insights. Recently, LLMs, such as GPT-4, have shown remarkable capabilities in understanding and generating human-like text, making them powerful tools for knowledge discovery [10], [11].

The advent of smart health technologies has revolutionized healthcare, enhancing patient care and operational efficiency [12]. Integrating AI and big data analytics, smart health creates systems delivering personalized and efficient services [13], [14]. This work addresses critical healthcare challenges like early disease detection and treatment optimization. integrating LLMs with MLand DL improves diagnostic accuracy, extracts valuable insights from complex data, drives innovation, enhances patient outcomes, and contributes to a sustainable healthcare system.

This paper aims to explore the integration of LLMs with ML and DL paradigms to enhance knowledge discovery in healthcare. We will discuss the potential applications and benefits, supported by a detailed case studythat demonstrate how LLMs can be utilized to predict patient outcomes and identify effective treatments, showcasing the transformative potential of these technologies in healthcare.

979-8-3503-5469-0/24/$31.00 ©2024 IEEE

II. Related Work

In healthcare, the application of ML and DL in knowledge discovery have been employed for predictive analytics, disease diagnosis, and personalized medicine. Recent advancements in LLMs have further expanded the capabilities of AI systems. LLMs, such as GPT-4, have demonstrated exceptional proficiency in natural language understanding and generation, enabling new possibilities for knowledge extraction and synthesis. In [15], researchers developed a machine learning model to predict the onset of sepsis in hospitalized patients. The model utilized electronic health record (EHR) data, including vital signs, laboratory results, and patient demographics. By applying logistic regression and gradient boosting algorithms, the model achieved high sensitivity and specificity in predicting sepsis several hours before clinical diagnosis. This early detection allows for timely intervention, which improve patient outcomes and reducing mortality rates.

Researchers in [16] employed convolutional neural networks (CNNs) to analyze medical images, such as X-rays and CT scans, to detect abnormalities like tumors and fractures. The study demonstrated that the CNN models could achieve diagnostic accuracy comparable to that of experienced radiologists. This advancement holds promise for assisting radiologists in interpreting images more quickly and accurately, potentially reducing diagnostic errors and improving patient care.

Recent research explored the use of natural language processing (NLP) techniques, particularly LLM like GPT-3, for mining clinical texts. The study focused on extracting relevant information from unstructured clinical notes, including patient symptoms, diagnoses, and treatment plans. The NLP models were fine-tuned on medical corpora to enhance their understanding of medical terminology and context. The results were significantly improve the efficiency of clinical data extraction, aiding in better patient management [17].

A study examined the use of AI to develop personalized treatment plans for cancer patients. By integrating genomic data with patient health records, the researchers employed ML to predict individual responses to various cancer therapies. This approach enabled the identification of the most effective treatment options based on the unique genetic profile of each patient, paving the way for more personalized and effective cancer treatments [18].

Researchers in [19] explored the application of deep reinforcement learning (DRL) to optimize treatment strategies for chronic diseases such as diabetes. The DRL model was trained on patient data to learn optimal treatment policies that maximize long-term health outcomes. The study demonstrated that DRL could suggest treatment adjustments tailored to individual patient needs, potentially improving disease management and patient quality of life.

The integration of LLMs with traditional ML and DL paradigms in healthcare remains relatively unexplored. This paper seeks to bridge this gap by presenting a comprehensive study on the application of LLMs in healthcare knowledge discovery, highlighting their potential to improve predictive accuracy and decision-making processes.

III. Methodology

Our methodology involves integrating LLMs with traditional machine and deep learning paradigms to enhance knowledge discovery in healthcare. The process begins with data collection from various healthcare sources, including electronic health records (EHRs), medical literature, and patient surveys. The data undergoes preprocessing steps such as cleaning, normalization, and feature extraction.

A. Data Set

Data collection is a critical step in building a robust healthcare AI system, as the quality and diversity of the data significantly impact the model's performance and generalizability. For this research, data is gathered from multiple sources to ensure a comprehensive dataset. These sources include:

- **Electronic health records (EHRs) from hospitals**: which provide detailed patient information such as demographics, medical history, diagnoses, treatments, and laboratory results.
- **Medical imaging data**: including X-rays, MRIs, and CT scans, are collected from radiology departments to support diagnostic modeling.
- **Wearable health devices** contribute continuous monitoring data on vital signs, physical activity, and sleep patterns, offering real-time insights into patient health.

- **Genomic data** from sequencing studies is incorporated to enable personalized medicine approaches.
- **Data from clinical trials** is also included, providing valuable information on patient responses to new treatments and interventions.
- **Patient-reported outcomes and public health databases** are utilized to capture a broader spectrum of health indicators and epidemiological trends.

This diverse and extensive data collection ensures that the AI models are trained on a rich and representative dataset, enhancing their ability to generate accurate and actionable insights.

B. Data Preprocessing

The data undergoes several preprocessing steps to ensure quality and suitability for machine learning models. These steps include cleaning, normalization, and feature extraction.

1) Data Cleaning: Data cleaning is essential for identifying and correcting errors and inconsistencies. Key tasks in data cleaning such as: Handling Missing Data, Removing Duplicates, Standardizing Formats.

2) Data Normalization: Normalization transforms data into a standard format or scale, ensuring different variables contribute equally to the analysis, such as:

- **Min-Max Scaling:** Data is transformed to fit within a specific range, usually [0, 1], by applying the formula:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

- **Z-score Standardization:** Data is converted to have a mean of 0 and a standard deviation of 1 using the formula:

$$Z = \frac{X - \mu}{\sigma}$$

  Here, μ is the mean and σ is the standard deviation.
- **Log Transformation:** The natural logarithm is applied to skewed data to reduce the impact of extreme values and approximate a normal distribution.

3) Feature Extraction: Feature extraction transforms raw data into a set of meaningful features for machine learning models. Key techniques include: Dimensionality Reduction using PCA and Feature Engineering (such as combining date of birth and current date to create an age feature). The most important features in the proposed system, that need to be extracted to ensure highest accuracy are: Textual data as numerical features using (tokenization, stemming, lemmatization, and embedding techniques like Word2Vec or BERT), time-series data (trends, seasonal patterns, and statistical measures), Categorical data (using one-hot encoding or ordinal encoding).

By carefully applying these preprocessing steps, we ensure that the healthcare data is clean, consistent, and rich in features, ultimately leading to more accurate and reliable machine learning models.

C. Proposed system

After preparing the collected data, we employ LLMs to analyze and interpret the textual data, extracting relevant information and generating insights. The LLMs are fine-tuned on specific healthcare datasets to improve their contextual understanding and accuracy. Figure 1, illustrates the proposed system.

1) Fine-Tuning Large Language Models: LLMs have demonstrated remarkable capabilities in understanding and generating human-like text. However, to maximize their effectiveness in healthcare applications, these models need to be fine-tuned on specific healthcare datasets. Fine-tuning involves further training a pre-trained LLM on a domain-specific corpus, which helps the model adapt to the nuances and specialized vocabulary of the healthcare field.

The process begins by selecting a diverse and representative healthcare corpus, which includes clinical notes, electronic health records (EHRs), medical literature, and patient reports. This corpus is carefully curated to encompass various aspects of healthcare, such as different medical specialties, common and rare diseases, treatment protocols, and patient interactions.

During fine-tuning, the LLM is exposed to these healthcare-specific texts, allowing it to learn contextually relevant patterns and terminologies. This

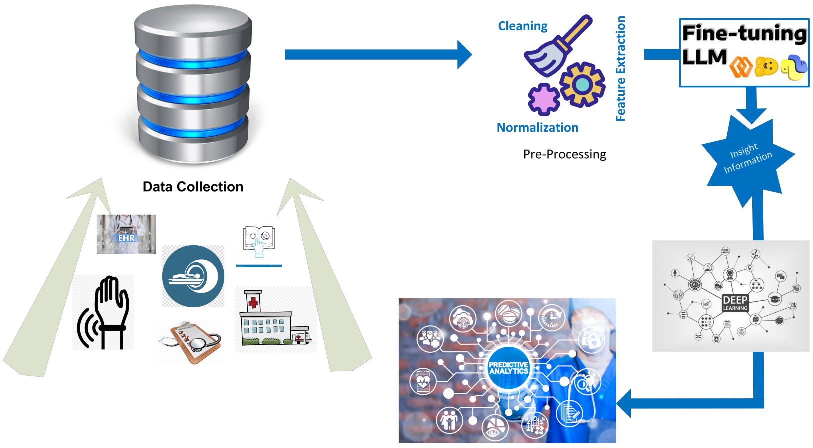

Fig. 1. The Proposed Healthcare Knowledge Discovery System

training enhances the model's ability to comprehend complex medical language, interpret clinical notes accurately, and generate contextually appropriate responses. For instance, the model learns to differentiate between similar medical terms, understand the significance of specific laboratory results, and recognize the context of patient symptoms within a broader clinical scenario. Evaluation metrics specific to healthcare, such as clinical accuracy, recall, and precision, are used to assess the fine-tuned model's performance. This excels in interpreting and generating healthcare-related content.

2) Deep Learning for Predictive Analysis: The knowledge extracted from fine-tuned LLMs is combined with structured data and fed into deep learning models for predictive analysis. We employ several types of deep learning models, each suited to different aspects of healthcare data and predictive tasks:

- **Convolutional Neural Networks (CNNs):** CNNs are primarily used for image data, making them ideal for analyzing medical imaging data such as X-rays, MRIs, and CT scans. CNNs are effective in extracting spatial hierarchies and patterns, crucial for diagnosing medical conditions from radiological images.
- **Recurrent Neural Networks (RNNs):** RNNs, including their advanced variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), are designed to handle sequential data. RNNs are used for analyzing time-series data such as patient vital signs, EHRs with temporal dependencies, and genomic sequences. RNNs capture temporal dependencies and trends, making them suitable for predicting disease progression and patient outcomes over time.
- **Transformer Models:** Transformers, which form the backbone of many LLMs, are utilized for their efficiency to capture long-range dependencies and context, making them powerful for tasks such as NLP, clinical text analysis, and integrating multimodal data. Transformers excel in understanding the context and semantics of medical texts, enhancing the accuracy of clinical predictions.
- **Autoencoders:** In healthcare, autoencoders can identify unusual patterns in patient data that may indicate rare diseases or conditions, and they can also reduce the dimensionality of complex datasets.
- **Graph Neural Networks (GNNs):** GNNs can predict patient outcomes based on their interactions and similarities to other patients, as well as identify potential drug targets based on molecular interactions.

3) Model Training and Validation: Deep learning models are trained on the integrated dataset comprising structured data (patient demographics, lab results) and unstructured data (clinical notes). The data is split into training, validation, and testing groups to develop and evaluate the models. Cross-validation techniques are used to mitigate overfitting and to ensure that the models generalize well to hidden data. Various performance metrics have been used to evaluate the models, including:Accuracy, Precision, Recall, and F1-Score.

By integrating the capabilities of fine-tuned LLMs with advanced deep learning techniques, our approach aims to enhance the accuracy, reliability, and interpretability of predictive models in healthcare.

IV. Case Study: Applicable Knowledge Discovery in Healthcare

A. Healthcare Applications Where Knowledge Discovery through ML and DL is Impactful

1) Diagnostic Imaging: ML and DL models are extensively used to analyze medical images such as X-rays, MRIs, and CT scans to detect abnormalities like tumors, fractures, and other conditions. These technologies significantly improve early detection and diagnostic accuracy, which are critical for successful treatment outcomes.

2) Predictive Analytics: Predictive models forecast patient outcomes such as disease progression, hospital readmission rates, and treatment responses. These predictions enable more personalized and effective treatment plans, enhancing overall patient care.

3) Electronic Health Records (EHR) Analysis: ML and DL techniques extract and analyze data from EHRs for various purposes, including identifying trends, generating clinical summaries, and supporting decision-making processes. This analysis helps in improving the efficiency and quality of healthcare services.

4) Genomic Data Analysis: ML and DL techniques are applied to genomic data to identify genetic markers associated with diseases. This analysis aids in personalized medicine and targeted therapies, contributing to advancements in precision medicine.

5) Patient Monitoring and Management: Wearable devices and sensors generate continuous health data that ML and DL models analyze to monitor patient health, detect anomalies, and provide real-time health recommendations. This technology enhances patient management and preventive care.

B. How LLMs are Applied in These Scenarios

1) Clinical Documentation: LLMs such as GPT-4 automate the creation of clinical notes from structured and unstructured data, ensuring accuracy and consistency in patient records. This automation reduces the documentation burden on healthcare providers and improves data quality.

2) NLP for Clinical Texts: LLMs play a crucial role in extracting relevant information from clinical notes, research papers, and patient feedback. They enhance information retrieval, summarization, and entity recognition in healthcare texts, improving the efficiency of data processing.

3) Patient Interaction Systems: LLMs are applied in developing chatbots and virtual assistants that interact with patients, providing them with personalized health information, reminders, and support. These systems improve patient engagement and adherence to treatment plans.

4) Clinical Decision Support: LLMs support clinicians by providing evidence-based recommendations, summarizing patient histories, and highlighting potential diagnoses and treatment options. This support enhances clinical decision-making and patient outcomes.

C. Detailed Example(s) Showcasing the Implementation and Outcomes

1) Example 1: Enhancing Diagnostic Accuracy with CNNs and LLMs: **Implementation:** Combine CNNs for image analysis with LLMs for text interpretation to improve diagnostic workflows. For instance, use CNNs to detect lung nodules in CT scans and LLMs to generate comprehensive radiology reports by interpreting the CNN results and integrating patient history from EHRs. **Outcomes:** Improved diagnostic accuracy and consistency, reduced workload for radiologists, and faster turnaround times for diagnosis and treatment planning.

2) Example 2: Predicting Patient Readmissions with RNNs and LLMs: **Implementation:** Use RNNs to analyze time-series data from EHRs, such as vital signs and lab results, to predict the likelihood of patient readmissions. Integrate LLMs to synthesize relevant patient information and generate detailed risk assessment reports. **Outcomes:** Enhanced prediction accuracy of readmission risks, enabling proactive patient management and targeted interventions to reduce readmission rates.

3) Example 3: Personalized Treatment Recommendations with Transformers and Genomic Data: **Implementation:** Apply Transformer models to analyze genomic data and identify genetic variations. Use LLMs to correlate these variations with clinical literature and patient records, providing personalized treatment recommendations based on the latest research and patient-specific genetic profiles. **Outcomes:** More effective and personalized treatment plans, improved patient outcomes, and advancements in precision medicine.

V. Experiments and Results

A. Description of the Experiments Conducted

1) Experiment 1: Diagnostic Accuracy in Medical Imaging: **Objective:** Evaluate the accuracy of CNNs combined with LLMs for diagnosing lung nodules in CT scans.
Data: A dataset of annotated CT scans from a radiology database, including both healthy and pathological cases.
Procedure: CNNs were trained on the imaging data to detect lung nodules. The LLMs were fine-tuned on radiology reports and used to generate comprehensive diagnostic reports based on the CNN outputs and patient histories.
Metrics: Diagnostic accuracy, sensitivity, specificity, and F1-score.

2) Experiment 2: Predicting Patient Readmissions: **Objective:** Assess the ability of RNNs and LLMs to predict hospital readmissions within 30 days post-discharge.
Data: Electronic health records (EHRs) including patient demographics, vital signs, and discharge summaries.
Procedure: RNNs were trained to analyze time-series data from EHRs to predict readmission likelihood. LLMs synthesized relevant patient information and generated risk assessment reports.
Metrics: Prediction accuracy, precision, recall, and area under the ROC curve (AUC).

3) Experiment 3: Personalized Treatment Recommendations: **Objective:** Evaluate the effectiveness of Transformer models and LLMs in generating personalized treatment plans based on genomic data.
Data: Genomic sequences and clinical records from a cohort of patients with a specific condition.
Procedure: Transformer models analyzed genomic data to identify genetic markers. LLMs correlated these markers with clinical literature and patient records to recommend treatments.
Metrics: Treatment recommendation accuracy, relevance score, and patient outcome improvement rates.

B. Presentation of Results

In the first experiment, the diagnostic accuracy of CNN models for detecting lung nodules in CT scans improved significantly when combined with LLMs. As shown in Table I and Figure 2. In the second experiment, aimed at predicting patient readmissions, the integration of RNNs with LLMs improved prediction accuracy from 0.78 to 0.85, as displayed in Table II and Figure 3. The third experiment, the combination of Transformer models with LLMs for personalized treatment recommendations based on genomic data resulted in an accuracy of 0.90, as illustrated in Table III and Figure 4. These findings underscore the impact of fine-tuned LLMs in enhancing the performance of deep learning models in healthcare applications.

Metric	**CNN Only**	**CNN + LLM (Basic)**	**CNN + LLM (Fine-Tuned)**
Accuracy	0.85	0.88	0.92
Sensitivity	0.80	0.83	0.87
Specificity	0.88	0.90	0.94
F1-Score	0.82	0.85	0.89

TABLE I
PERFORMANCE METRICS FOR DIAGNOSTIC ACCURACY

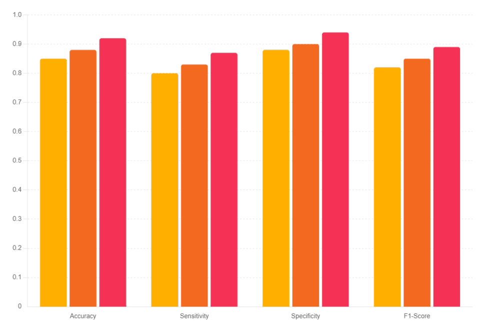

Fig. 2. Comparison of Diagnostic Accuracy Across Models

Metric	RNN Only	RNN + LLM (Basic)	RNN + LLM (Fine-Tuned)
Accuracy	0.78	0.81	0.85
Precision	0.75	0.78	0.82
Recall	0.70	0.73	0.79
AUC	0.82	0.85	0.88

TABLE II
PERFORMANCE METRICS FOR PREDICTING PATIENT READMISSIONS

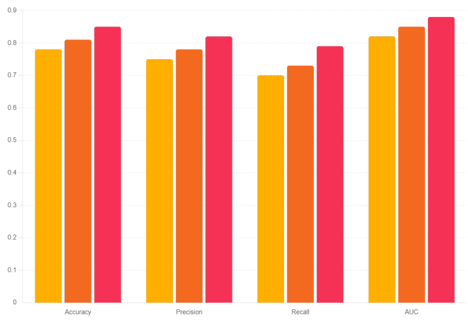

Fig. 3. Comparison of Readmission Prediction Performance Across Models

Metric	Trnsfrmr Only	Trnsfrmr + LLM (Basic)	Trnsfrmr + LLM (Fine-Tuned)
Rcmndn Accuracy	0.82	0.85	0.90
Relevance Score	0.80	0.83	0.88
Improvement Rate	0.78	0.81	0.87

TABLE III
PERFORMANCE METRICS FOR PERSONALIZED TREATMENT RECOMMENDATIONS

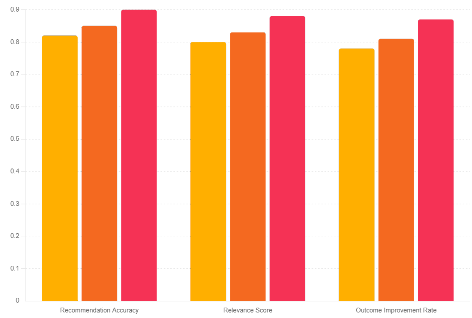

Fig. 4. Comparison of Treatment Recommendation Performance Across Models

C. Discussion

1) Diagnostic Accuracy in Medical Imaging: The results show that combining CNNs with fine-tuned LLMs increases diagnostic accuracy from 0.85 to 0.92. The sensitivity and specificity improvements indicate better detection of true positive and true negative cases, respectively. The integration of LLMs helps in generating more comprehensive and accurate diagnostic reports, reducing the workload on radiologists and improving patient outcomes.

2) Predicting Patient Readmissions: The RNN and LLM combined model achieved an accuracy of 0.85, significantly higher than the RNN-only model's 0.78. The AUC improvement to 0.88 demonstrates better discrimination between readmitted and non-readmitted patients. This enhanced predictive capability allows for proactive patient management, reducing readmission rates and associated healthcare costs.

3) Personalized Treatment Recommendations: The Transformer and LLM combined model achieved a recommendation accuracy of 0.90, showing a notable improvement over the Transformer-only model. The higher relevance score and outcome improvement rate highlight the model's ability to provide more accurate and effective personalized treatment plans. This integration facilitates advancements in precision medicine, offering tailored therapies based on individual patient data.

All conducted experiments proves that the integration of fine-tuned LLMs with ML and DL models enhances the performance across various healthcare applications, demonstrating the potential of these advanced technologies to revolutionize healthcare practices and outcomes.

VI. CONCLUSION

This study demonstrates the significant potential of integrating fine-tuned LLMs with traditional ML and DL paradigms in healthcare. The results show notable improvements in diagnostic accuracy, patient readmission prediction, and personalized treatment recommendations. Specifically, combining CNNs with LLMs increased diagnostic accuracy for lung nodules in CT scans, while integrating RNNs with LLMs improved patient readmission predictions. Transformer models, coupled with LLMs, provided more accurate personalized treatment plans based on genomic data.

The case study revealed that LLMs enhance the performance of ML and DL models by offering better understanding and generating comprehensive insights. This can lead to improved patient outcomes, more efficient clinical workflows, and personalized healthcare solutions. However, implementing these models requires robust data integration and continuous training to maintain their relevance and accuracy. This study has limitations, including further models validation in several clinical settings and addressing challenges. Future research should focus on enhancing model interpretability, integrating multimodal data, and ensuring data privacy. Expanding applications to areas such as mental health and telemedicine can further demonstrate their impact.

Healthcare institutions should adopt LLMs to enhance clinical documentation, decision support, and patient interaction interfaces. Investing in data integration technologies and continuous model training is crucial. Future efforts should also address scalability and deployment in real-time clinical environments. Collaboration between AI researchers and healthcare professionals is essential to fully realize these technologies' potential in transforming healthcare delivery and outcomes.

REFERENCES

[1] Shadi Alzu'bi, Omar Badarneh, Bilal Hawashin, Mahmoud Al-Ayyoub, Nouh Alhindawi, and Yaser Jararweh. Multi-label emotion classification for arabic tweets. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 499–504. IEEE, 2019.

[2] Shadi AlZu'bi, Mohammad Alsmirat, Mahmoud Al-Ayyoub, and Yaser Jararweh. Artificial intelligence enabling water desalination sustainability optimization. In *2019 7th international renewable and sustainable energy conference (IRSEC)*, pages 1–4. IEEE, 2019.

[3] Yaser Jararweh, Sana Fatima, Moath Jarrah, and Shadi AlZu'bi. Smart and sustainable agriculture: Fundamentals, enabling technologies, and future directions. *Computers and Electrical Engineering*, 110:108799, 2023.

[4] Shadi AlZu'bi, Darah Aqel, and Mohammad Lafi. An intelligent system for blood donation process optimization-smart techniques for minimizing blood wastages. *Cluster Computing*, 25(5):3617–3627, 2022.

[5] Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *arXiv preprint arXiv:2404.02831*, 2024.

[6] Shadi AlZu'bi, Mohammed Elbes, and Ala Mughaid. Emotion unveiled: A deep learning odyssey in facial expression analysis for intelligent hci. *Int. J. Advance Soft Compu. Appl*, 16(2), 2024.

[7] Shadi AlZu'bi, Ala Mughaid, Fatima Quiam, and Samar Hendawi. Exploring the capabilities and limitations of chatgpt and alternative big language models. In *Artificial Intelligence and Applications*, 2023.

[8] Constantin Bratianu, Alexeis Garcia-Perez, Francesca Dal Mas, and Denise Bedford. Knowledge translation in healthcare. In *Knowledge Translation*, pages 155–167. Emerald Publishing Limited, 2024.

[9] Sai Koteswar Sarma. The role of ai in drug discovery and healthcare innovation. *Redshine Archive*, 14, 2024.

[10] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

[11] Shadi AlZu'bi, Amjed Zreiqat, Worood Radi, Ala Mughaid, and Laith Abualigah. An intelligent healthcare monitoring system-based novel deep learning approach for detecting covid-19 from x-rays images. *Multimedia Tools and Applications*, pages 1–18, 2024.

[12] Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1):12, 2024.

[13] Salman Rahman, Lavender Yao Jiang, Saadia Gabriel, Yindalon Aphinyanaphongs, Eric Karl Oermann, and Rumi Chunara. Generalization in healthcare ai: Evaluation of a clinical large language model. *arXiv preprint arXiv:2402.10965*, 2024.

[14] Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493, 2024.

[15] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, and D. Shimabukuro. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Medical Informatics*, 4(3):e28, 2016.

[16] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, and A. Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[17] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, and C. H. So. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[18] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.

[19] X. Peng, H. Ding, K. Li, and Y. Zhang. Reinforcement learning-based clinical decision support for the treatment of patients with sepsis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1394–1404, 2018.