Automating Clinical Documentation: A Comprehensive Systematic Review of NLP Techniques and Challenges in Digital Scribing

Md. Faiyed Bin Karim[1]. Abu Ubaida Akash[1]. Sakib Zaman[1]. Shehan Irteza Pranto[1]. Marzia Zaman[2]. Farhana Sarkar[2]. Mohammad Nurul Huda[3]. Nabeel Mohammed[4]. Shariful Islam[5]. Iman Abdollah Dehzangi[6]. Khondaker Abdullah Al Mamun[1]

Abstract

In the domain of healthcare systems, the Electronic Health Record (EHR) has enabled physicians to store patient data, track follow-ups, and collect information for future research on diseases. However, the process of documenting conversations between doctors and patients for EHR is often cumbersome and time-consuming, leading to decreased physician efficiency. Additionally, hiring a scribe to perform this task can be complicated and expensive. Recent advances in machine learning and natural language models have paved the way for automating medical documentation, specifically doctor-patient conversations. These approaches integrate Automatic Speech Recognition (ASR), Text Summarization, and Named Entity Recognition (NER) to streamline the process. However, implementing medical scribing poses several challenges and barriers that can affect the quality of generated prescriptions. Numerous techniques have been proposed to address the limitations associated with implementing medical transcription in healthcare systems. This study follows a systematic literature review approach and the PRISMA statement to analyze peer-reviewed articles from seven databases—IEEE Explore, ACM Digital Library, Arxiv, PubMed, SpringerLink, ACL Anthology, and Google Scholar. Sixty-eight articles were included in the review and were analyzed through a descriptive and thematic analysis. Of these, 26 focused on ASR technologies, 17 on text summarization methods, and 27 mentioned NER techniques. Our analysis revealed that most studies were conducted on Chinese datasets, and the Google speech-to-text API was commonly used in ASR tasks. Text summarization and NER tasks frequently employed transformer-based models, which yielded promising results. However, while many articles addressed individual tasks of medical scribing, few included all the needed tasks for comprehensive medical documentation. Additionally, models performed well in individual tasks but lacked practical implementation due to limited training data.

Keywords Digital Scribing, Medical Documentation, Systematic Review, Automatic Speech Recognition, Named Entity Recognition, Text Summarization, Machine Learning.

Khondaker Abdullah Al Mamun
mamun@cse.uiu.ac.bd

Extended author information is available on the last page of the article

1. Introduction

Recently adopted Electronic Health Records in healthcare sectors have enabled physicians to store important patient data for improved patient care, study cases with complete patient histories, and maintain proper schedules through faster data analysis [1, 2]. Through medical scribing [3], doctor-patient conversations can yield the needed information. The physician documents the entire doctor-patient conversation, which is time-consuming and tedious and requires a lot of expertise from the scribe. A 2020 survey [4] of 7510 US clinicians revealed that 38.2% had at least one burnout symptom. Medical scribes [5] are now used in clinical documentation, reducing physician burnout significantly. However, this initiative has added complexity and training and hiring costs of scribes, inefficient service, and the possibility of data violations of patients' confidential information. Thus, incorporating several automated tools for medical scribing can reduce the challenges faced through manual scribing and increase the efficiency of the work.

Recent advances in NLP and machine learning have enabled its use in many fields and automated many manual tasks [6]. ASR, NER, and text summarization make up a complete digital scribing model. These tasks also have subtasks like patient data de-identification, noise suppression, speaker diarization, entity extraction, and more. In healthcare, Nuance [7], Robin Healthcare [8], DeepScribe [9], and Amazon [10] are developing digital scribing models to replace manual doctor-patient conversations. Several research methods have been developed for digital scribing, but few have been used for clinical documentation. The efficiency of the model determines the clinical implementation of medical documents. Concerns about patient data security and privacy necessitate anonymization.

Several review studies have examined digital scribing tasks and their applicability, showing how NLP and AI tools were used to approach different medical scribing tasks, evaluating their performance using performance matrices, and identifying their limitations and drawbacks. As automated tools have not been widely used in medical documentation, these review papers found poor model performance. Limitations of existing review papers include not comparing research in different languages and clinical settings. In practice, there are environmental barriers to processing the doctor-patient conversation that the review papers did not consider. The trained models' performance is affected by accents even in the same language.

Our proposed review paper examined digital scribing models on datasets varying in language, accent, clinical settings, size, and conversation type. We also looked for ways to optimize task-specific NLP models by processing noise, punctuation, non-lexical terms, and audio frequency. The major objectives that our review gained are:

1. Observation, evaluation and comparison of NLP models performed in several individual tasks of medical scribing (ASR, NER, text summarization, de-identification).
2. Observation of the type of dataset the task-specific models were trained on. We did not solely compare the trained models, rather we evaluated them based on their dataset.
3. Assessing the limitations and challenges the proposed models faced.
4. Analysis of the preprocessing and postprocessing tasks that optimized the performance of the trained models.

The upcoming sections of this paper will offer an in-depth description of our systematic review. Section 2 will explain the methods and structure used in our review, while Section 3 will present the key outcomes of our research. Additionally, Sections 4 and 5 will provide a comprehensive discussion of our findings and highlight the limitations of the existing models.

Finally, in Section 6, we will offer concluding remarks summarizing the focus of our study.

2. Methodology

For our systematic review of our elected field, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocol (PRISMA-P) [11]. The PRISMA method encompasses a 27-item checklist that guarantees the completeness and quality of the review process while minimizing bias and maximizing the reproducibility of the findings. The PRISMA method offers an effective structure for identifying, selecting, and synthesizing relevant studies, as well as for evaluating the overall quality and risk of bias of the included studies. Therefore, our study utilized the PRISMA model for the selection of adequate articles. The research work selection and categorization for our systematic review following the PRISMA method has been displayed as a flowchart (see Fig. 1) where the number of identified, screened, included and excluded articles in each stage has been represented briefly.

2.1. Study Selection Strategy and Literature

Reviews based on studies related to the current status and recent developments on digital scribing have been performed in recent years where progress on various components of digital scribing has been observed. Table 1 shows a list of review works that have been done in recent years.

Table 1 Previous review done on the related fields

Reference	Year	Review Type	Objective
[12]	2019	systematic	Literature on the use of Speech recognition (SR) technology in healthcare, covering a wide range of medical domains, and clinical documentation, highlights the need for further investigation of the use and effectiveness of SR beyond radiology concerning the pros and cons and improvement.
[13]	2021	Scoping	Providing an overview of the current state of development, validation, and implementation of digital scribes and NLP tools for healthcare, and recommending future research to focus on clinical validity and usability.
[14]	2021	Scoping	Measurement of the potential benefits of digital scribe technology in reducing physician burnout and improving patient experience concerning the barriers to implementation such as linguistic variations, and upfront costs and suggests potential solutions to overcome these barriers.
[15]	2023	Systematic	To assess the potential of intelligent solutions for ASR with automatic documentation during a medical interview and identify future research area

To identify recent studies, the objective, scope and research questions were initially visualized and the inclusion and exclusion criteria were subsequently determined. Depending on the inclusion and exclusion criteria, the relevant studies were elected for further review. The inclusion and exclusion criteria represent the recent advancements in the relevant field. The inclusion criteria that were chosen narrowed the study to some specific topics and scenarios keeping the scope of the study more precise and the study domain more elaborate. The research objectives along with research questions are stated (see Table 2) and the selection criteria for the research articles are represented (see Table 3).

Table 2 Research Statement, Objective and Research Questions of our Study

Research Statement	To scrutinize diverse facets of digital scribing in clinical documentation, while assessing the progressions that have been achieved in correlation with varied settings and configurations.
Objective	To make a comprehensive evaluation of the efficacy of various methods, an assessment of the innovativeness and constraints of various methods applied for different tasks in digital scribing in the context of clinical documentation, explore their potential contributions across diverse healthcare settings and populations and identification of possible impediments and enablers to their adoption.
Research Questions	RQ1: What type of clinical data were collected and how were the models applied to them?
	RQ2: What explicit subtasks had been incorporated in medical documentation and what techniques were implemented for executing particular subtasks?
	RQ3: Which metrics were employed to evaluate the performance of the implemented techniques and what was the extent of their efficacy in achieving the desired outcomes?
	RQ4: What was the degree of novelty or originality demonstrated by the implemented techniques in comparison to existing approaches? Studies that do not report outcomes related to the use or impact of digital scribe aches or methods?
	RQ5: What ethical implications are associated with the adoption of digital scribing in Electronic Health Records (EHRs), including issues pertaining to patient confidentiality and safeguarding of data?
	RQ6: What is the practicality of implementing tasks of digital scribing in diverse healthcare contexts, considering the influence of demographic factors, language, clinical setup, environment and population, and identifying potential obstacles and catalysts to adoption?
	RQ7: In what manner can techniques related to digital scribing surmount the hindrances presented by manual scribing?

Table 3 Inclusion and Exclusion Criteria for our proposed study

Inclusion Criteria	Exclusion Criteria
Articles that elaborated on different tasks on automated scribing	Articles where different subtasks of digital scribing methods implemented beyond the field of electronic health record systems
Articles that are written in English	Articles which are written in any other language rather than English
Studies that are exclusively focused on the subject of medical documentation, are methodologically rigorous and relevant to the research question.	Studies that were not empirical research, such as books, surveys, review papers, conference reviews, editorials and opinion papers.
Articles that have established models and displayed the performance evaluation of their model	Studies that did not aim to achieve any subtask related to digital scribing or did not attempt to automate medical documentation in any way.
Articles that have described different practical scenarios regarding digital scribing or a specific task on digital scribing	Studies that do not report outcomes related to the use or impact of digital scribes

3. Results

3.1 Research Work Selection from the Database

To accomplish the stated research objective and address the research questions, an extensive search was conducted for scholarly articles across multiple databases, utilizing a carefully curated set of search strings. This methodological approach allowed for a comprehensive examination of the literature, enabling us to obtain a diverse range of relevant and high-quality sources to inform our study.

Upon selecting suitable databases, various combinations of search strings were formulated and tested via an advanced search of the databases. The search strings that yielded the most favourable outcomes were chosen for further screening. The selected search string included terms: "digital scribe", "clinical documentation", transcription, "natural language processing", NLP, "named entity recognition", medical*, clinical*, digital, limitation*, barrier*, environment*, population, language*, Chinese, English, “speaker diarization”, “speech recognition”, ASR, “text summarization”, “speech to text”, automate*, scribe*, NER, conversation*, EHR, healthcare, “text to snippet”. The advanced search feature of databases provides operators with the ability to create variations in search terms to obtain the maximum search results. By combining search terms with Boolean operators "AND", "OR" and “NOT”, more specific and extensive search results can be generated. Furthermore, the utilization of quotation marks " permits the search for exact phrases, whereas the asterisk * allows for the search for variations of a specific keyword or term. For example, the * is used as a wildcard to capture variations of the term "scribe," such as "scribes" or "scribing." The " is used to search for exact phrases, such as "medical scribe."

In addition, a manual search was executed on Google Scholar to retrieve pertinent

articles. Furthermore, specific papers were obtained from the references of relevant study works.

Table 4 Search string sample of the ACM Digital Library

[[Abstract: scribe*] OR [Abstract: "medical documentation"] OR [Abstract: "clinical documentation"] OR [Abstract: "text summarization"] OR [Abstract: "speech to text"] OR [Abstract: "speaker diarization"] OR [Abstract: "text to snippet"]] AND [[Abstract: medical] OR [Abstract: clinic*] OR [Abstract: healthcare] OR [Abstract: ehr]] AND [[Abstract: nlp] OR [Abstract: "natural language processing"] OR [Abstract: asr] OR [Abstract: "automatic speech recognition"] OR [Abstract: "named entity recognition"] OR [Abstract: ner] OR [Abstract: automate*] OR [Abstract: digital] OR [Abstract: limitation*] OR [Abstract: barrier*] OR [Abstract: environment*] OR [Abstract: population*] OR [Abstract: language] OR [Abstract: english] OR [Abstract: german] OR [Abstract: chinese]] AND [E-Publication Date: (01/01/2013 TO 12/31/2023)]

The search results were restricted to the time frame between 2013 and 2023 and written in English. In accordance with the PRISMA flowchart, our study entailed the initial accumulation and retention of search results in a CSV file format. Utilizing a systematic and transparent approach for conducting systematic reviews is crucial in ensuring the accuracy and reliability of the review results. In order to obtain the desired research papers, it is necessary to employ a structured method backed by a diverse database and substantial research in the relevant sector along with clear and descriptive protocols, as well as thorough and rigorous investigation. Consequently, the indicated articles underwent inspection. In the first screening, the articles were selected by evaluating their titles, abstracts, and keywords. The articles were excluded based on the exclusion criteria of our study. In the second screening, a full-text survey of the remaining articles was done, and based on the inclusion and exclusion criteria, they were evaluated and selected.

3.2 Query Output

After the application of the search strings to corresponding databases and the collection of related articles from manual searches and reference retrieval, in total 1873 articles were found. These search results included both journal and double-blinded conference papers. Table 5 represents the number of search result outputs of individual databases.

Table 5 Search Results from Selected Databases

Database	Search Result
IEEE Xplore	264
ACM Digital Library	69
Arxiv	129
MedLine	608
SpringerLink	621
ACL Anthology	167
Google Scholar	15

Total	1873

Initial screening was performed on the 1873 identified articles. Upon removing the duplicates, 1796 articles were selected from the identified 1873 articles and were gone for further evaluation. Even though SpringerLink had the highest number of search results with 621 results, only 45 articles were selected after 1st screening. 78 articles from 264 research articles (IEEE Xplore) were included which was the highest screening to identified articles ratio. The second phase of screening was done on the selected 1796 articles by studying only the titles and abstracts of the articles. Upon analysis, 253 articles were kept for full-text study. After the full-test study on 253 articles, 116 articles were included for eligibility. Finally, after further evaluation, 68 articles were eventually selected that met our inclusion criteria for the systematic review. The excluded articles were attributed to various reasons, such as articles had duplicates, were not relevant to the healthcare domain, did not specify any subtask associated with digital scribing, were not written in the English language, or identified subtasks related to the medical domain digital scribing but did not undertake any efforts to automate medical documentation.

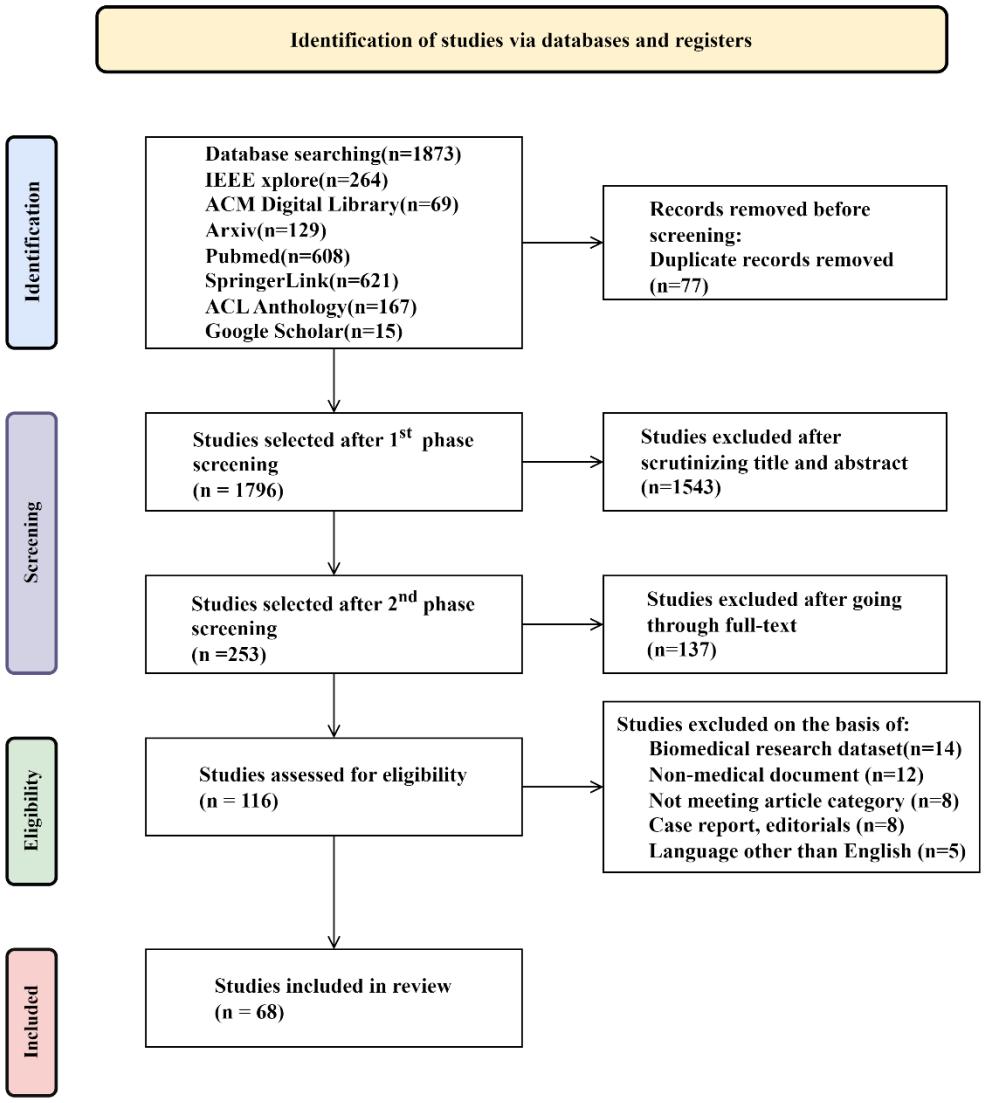

Fig. 1 PRISMA workflow diagram of article selection proposed systematic review

During the course of the studies, diverse empirical information was identified and recorded for the research. The key topics considered in article analysis are:

- Published year of the article
- Objective of the research work
- Subtask of digital scribing that was performed, subtasks included text summarization, Named Entity Recognition, Automatic Speech Recognition, speaker diarization, de-identification
- Type of data that was collected along with the language of the data and clinical setup
- Approach that was employed in achieving the objective
- Specialization that the approached model is done on the clinical document
- Evaluation of the performance of the approached models by measuring through evaluation metrics
- Limitations of approached models and the challenges they have been able to overcome

3.3 Features of Digital Scribing

In a generic digital scribing [16] tool for the Electronic Health Record (EHR) [17], several subtasks are performed to produce clinical documentation which includes Speech-to-text conversion [18], noise reduction [19], speaker diarization (SD) [20, 21], de-identification [22, 23], text summarization[24], text-to-snippet conversion which performed by Automatic Speech Recognition (ASR) [25], Named Entity Recognition (NER) and other NLP models. Based on the motive and target outcome, these subtasks are assembled to generate the intended medical documentation tool to reduce the complexity.

The ASR model is utilized to handle the physician-patient conversation and convert the audio into text of the language in which the model was trained. The ASR model first performs preprocessing for training the audio data which includes noise suppression [12, 14, 16, 26] audio normalization and segmentation [25, 26]. On the other hand, Speaker diarization (SD) involves dividing an audio recording into distinct segments based on the identity of the speaker in each segment. The segments are separated in a way that ensures that each one consists of only one speaker. After that, the outputs from SD and ASR are reconciled into a transcribed text as a dialogue script. After that, the transcribed text is processed through the NER model to extract necessary information (symptoms, patient information, disease name, medication details, follow-up suggestions, and advice). Later on, de-identification is performed by applying NER where the information of the patient is removed to ensure privacy and security. Furthermore, the key information can be summarized from the transcribed document using abstractive summarization techniques. A brief workflow diagram is displayed (see Fig. 2) for a better understanding of automated medical scribing.

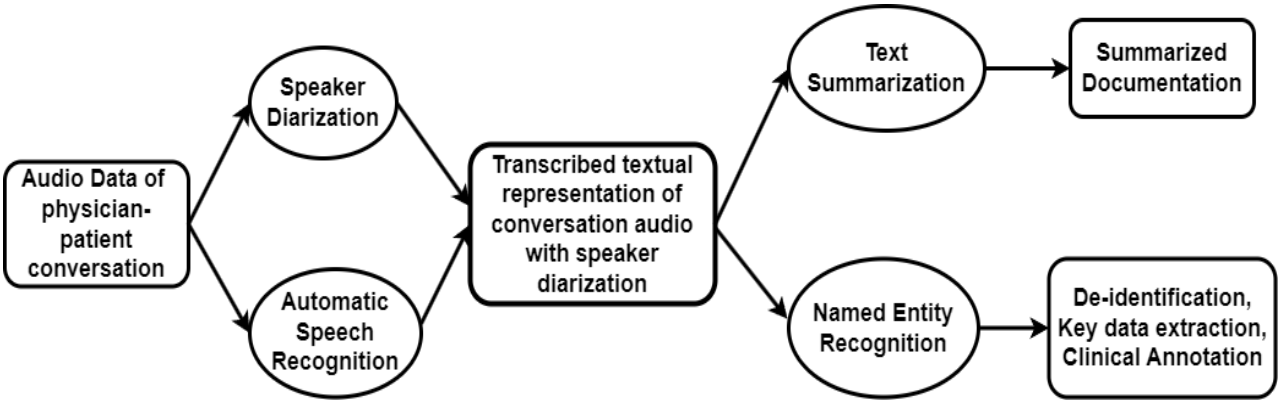

Fig. 2 Workflow Diagram of Automated Medical Scribing

3.4 Factors and Technologies in Speech-to-Text Conversion

Among the 68 selected articles that met our requirements for the study, 26 studies [18–22, 26–46] introduced Automatic Speech Recognition technologies for accomplishing different subtasks related to speech-to-text conversion. Audio data collection plays a vital role in ASR models.

3.4.1 Clinical Setup and Audio Format

To collect audio data, the recorder needs to be at an appropriate position from which the audio of the doctor and patient can be captured properly. Otherwise, there is a possibility of having different levels of intensity between the doctor's and the patient's audio. Two studies [22, 45] mentioned the positioning of the recorder between the doctor and patient where the microphone can record the audio from a distance. The quality of the audio also depends on the device used to record the audio. In three studies[29, 35, 45], the authors mentioned the tool that they used to record the audio; one study [29] used an H4n recorder and the other one used a PCM-A10 [35] recorder device for recording audio. The collected data for transcription is first required to be preprocessed properly. One of the preprocessing steps is proper formatting of the audio file as the ASR model may work on some specific formats of audio. Two studies kept the frequency of the audio at 44.1kHz [39, 41] and one study kept the frequency at 48kHz [20]. Lee et al. [35]recorded the audio and modulated it to 96kHz/24-bit. On the other hand, Chiu et al. [22] used a fast Fourier transform model where frequencies above 3800 Hz were ignored. In some of the studies, the collected audio files were converted to *.wav [20, 27, 35, 41] or *.mp3 [22] formats.

3.4.2 Overcoming Noise

In six research studies[19, 22, 26, 30, 34, 40], the ASR model was performed adequately under noisy backgrounds, where noise reduction was executed before transcribing speech audio data. In a particular investigation [19], the background noise was downsampled to 16 kHz to match the sampling rate of the clean audio. These two audio types were subsequently merged to create a dataset consisting of both noisy and clean audio segments. The resulting dataset was subsequently utilized to train the models, aiming to improve the accuracy of the ASR model. The purpose of this step is to ensure accurate transcription of the speech and obtain the correct text representation from the audio. To evaluate the performance of the models, noise can be added artificially to the audio before training the model with the data. Chiu et al. [22] added artificial noise from 20 different

backgrounds (room reverberation, background music, and café noises) where the noise ranged from 5 dB to 20 dB. Such implementation was defined as Multi-style Training (MTR). In one study[34], three types of noise were added artificially to the audio which were: cafeteria, people talking, and emergency sirens, which had three levels of intensities. The targeted models were evaluated using the combinations of the noises. A customized convolutional autoencoder was proposed for noise suppression by Menon et al. [19]. Automated Speech Recognition (ASR) technology is designed to recognize and transcribe speech into text format. Paats et al. [40] integrated open-source ASR models (Kaldi toolkit, Thrax) with different approaches of Language models on 219 radiology reports recorded in a real clinical environment that contained natural noise. The language models were trained, evaluated and modified for better performance. The modifications were performed through 8 versions and in version 8 the language model was adapted with spoken data. Throughout several stages of modifications, the final version showed improvement from a WER value of 18.4% (v1) to 5.8% (v8) and was able to be transcribed efficiently in a practical noisy environment.

3.4.3 Language variations in the datasets

The accuracy of the ASR output is strongly dependent on the language model used to train the system. The language model is created by analyzing a large corpus of text data in a particular language, which enables the system to recognize and transcribe speech in that language with a higher degree of accuracy. In different studies, the audio was transcribed to different languages such as Bahasa Indonesian [27], Javanese slang [27], English [18, 19, 28, 30, 45], Hindi [28], Gujarati [28], Serbian [26], Croatian [26], Latin phrase [26], Singapore English [29], Korean [35], Spanish[38, 41], Australian English [39], and Estonian languages[40].

3.4.4 Commonly used API models

In several cases where transcription is performed using established API models [18–20, 29, 30, 32, 34, 35, 37–40, 44–46], audio files in larger numbers in varieties of formats are acceptable where the models are available publicly. Some of the existing well-known open-source ASR technologies are: (i) Microsoft Speech API (Microsoft Azure) [18, 34]; (ii) Google Cloud API[19, 20, 29, 32, 34, 35, 37, 39, 44–46] (iii) Kaldi [30, 38, 41]; (iv) PocketSphinx [34]; (v) IBM BlueMix API [34]; (vi) Naver Clova SR (Naver Corporation) [35]; (vii) Amazon Transcribe[35, 37, 39] ; (viii) IBM Watson[39, 45] ; (ix) Thrax [40]. Table 6 shows the approached models, datasets, and performance based on the accuracy and WER of the proposed research work on speech-to-text conversion. Fig. 3 shows the WER range of existing ASR models.

In a study conducted by Xu et al. [29], the Google Cloud API was employed to transcribe audio recordings of conversations between physicians and patients diagnosed with schizophrenia. In another research work by Preum et al. [34], the Google Cloud API was employed because yielded superior results with the lowest WER in both noiseless and noisy environments. The resulting text was then utilized to extract lexical resources, which were used to evaluate the health condition of the patients. Jiang et al. [30] proposed the ASPIRE model using Kaldi, where two types of datasets were applied to the models—formatted and unformatted—both of which were in noisy environments. The formatted inputs included medical commands such as drug names and dosages

mentioned by the physician. The unformatted commands included medical inquiries and their answers. The trained model performed well on the unformatted input. The unformatted dataset was elaborated and contained additional vocabulary terms because the training set varied from the test set, and the ASR model did not perform as well on the unformatted dataset as it did on the formatted dataset.

Table 6 Proposed Research works on Speech to text Conversion

Approached Model	Dataset	Performance
Connectionist Temporal Classification (CTC) with Mel Frequency Cepstral Coefficient (MFCC) technique for preprocessing [27]	500 voices, 50 data for a test set of Bahasa Indonesian language, each voice of 2 seconds	Accuracy: 64% WER: 36%
Microsoft Azure [18]	ezDI dataset trained for 16 hours	21.5%
Speaker-independent ASR system with class n-gram and RNN language model with code-switching [26]	904 hours of speech data: 379 hours of Serbian (1087 speakers) and 525 hours of Croatian (1742 speakers)	Accuracy: 98.6; Perplexity Score: 59.55; WER: 1.4%
Google Cloud Speech API [29]	34 hours of audio data from 71 individuals at the Institute of Mental Health Singapore (IMH).	WER: 9%
ASPIRE model using Kaldi [30]	Formatted (30,000+ medical commands) and unformatted input (400,000+ question-answer pairs) in a noisy environment	WER (Unformatted input):4.5% SER (Formatted input): 14.4% SER (Unformatted input): 27.1%
End to end RNN-T [31]	93.3 hours of doctor-patient conversation, 34-534 utterances per conversation	Overall, CER reduction: 11.9%
Google Cloud Speech API (commercial) and ASPIRE (open source) with Sequence-to-sequence translation model [32]	Total 3807 deidentified doctor-patient conversation	WER (Google Cloud API+S2S): 34.1% BLEU (Google CLoud API+S2S): 56.4% WER (ASPIRE+S2S): 34.5% BLEU (ASPIRE+S2S): 55.8%
Naver Clova SR, Google	112 doctor-patient conversation audio	Highest Accuracy:

Speech-to-Text API and Amazon Transcribe [35]	files	75.1% (Naver Clova SR)

Table 6 (continued)

Approached Model	Dataset	Performance
Dragon Medical Edition 2 by Nuance Inc. [36]	audio recordings were sent to a server after being recorded, with no specific mention of the amount of data	WER: 22.4%
Google Speech-to-Text Clinical Conversation (Google ASR) and Amazon Transcribe Medical (Amazon ASR) [37]	Anonymized 36 primary care counters, 135647 spoken words with 3284 (2.4%) NLCS (Non-Lexical Conversational Sound). Among these NLCS, 76 (0.06% of total words, 2.3% of all NLCS) were clinically relevant terms	WER (Google ASR): 11.8% WER (Amazon ASR): 12.8% WER (Google ASR on added NLCS): 40.8% WER (Amazon ASR on added NLCS): 57.2%
Neural Networks model (NNET3 of Kaldi, based on sequential models including I-Vectors module) [38]	trained with almost 800 hours of audio of Spanish speakers from Argentina	Accuracy: 91.7% WER: 8.3%
open-source software components (Kaldi toolkit, Thrax) with 8 different adaptations of the language model [40]	219 Estonian language radiology reports containing 19928 words in a real clinical environment	Maximum improvement of WER (at version 8): 5.3% WER at version 1: 18.4%
Connectionist Temporal Classification (CTC), and end-to-end models with attention (LAS: Listen, Attend and Spell) [22]	anonymized conversations with approximately 14,000 hours of audio	WER (CTC): 20.1% WER (LAS): 18.3%
RNN-T for ASR and SD [21]	Approximately 15K hours of doctor-patient conversation, each conversation 10 minutes long	WDER (Word-level Diarization Error Rate): 2.2% WER (Word Error Rate): 19.3%
Google Cloud text-to-speech [45]	6,693 real doctor-patient conversations, each having a 9 min 28-sec duration; recorded in a clinical setting using distant microphones of varying quality.	WER: 50%

Google Cloud text-to-speech [46]	Medical Conversations dataset from AMC'17 [22] and a combined dataset SWFI [47, 48]	WER: 38.3%

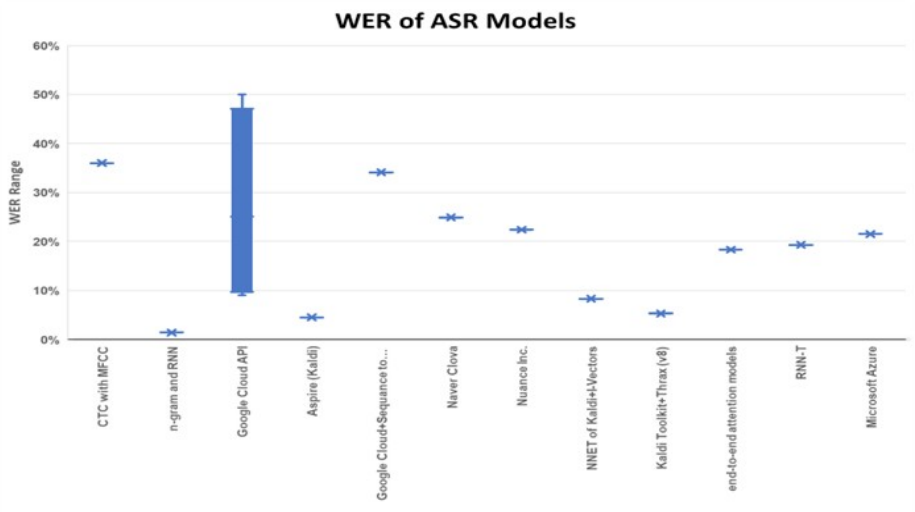

Fig. 3 WER Range of Existing ASR Models

3.4.5 Speaker Diarization

Speaker diarization [20, 21, 41] is an essential component in the conversion of speech to text, as it enables the identification of individual speakers and their respective speech segments. In an approach developed by Shafey et al. [21], speaker diarization and speech-to-text conversion were performed separately using a Recurrent Neural Network Transducer (RNN-T) which had improvement in WDER rating from 15.8% to 2.2%. Another study [41] introduced the development of a Kaldi X-Vectors-based system involving the training of new speaker diarization models. X-vectors are a specific type of deep neural network architecture that can be utilized for both speaker recognition and diarization.

3.4.6 Postprocessing

The preprocessing and postprocessing stages of ASR play critical roles in improving the accuracy and readability of the output text and optimizing these stages can significantly enhance the performance of an ASR system. Adding punctuation and truecasing of the transcribed text makes the text more understandable and enables the model to be more robust. Five studies [33, 36, 41, 43, 45] emphasized the punctuation and true casing of the transcription. Sunkara et al. [33] introduced pre-trained masked natural language processing (NLP) models, namely BERT, BioBERT, and RoBERTa, which were employed to predict punctuation and apply an accurate casing framework. Data augmentation was performed to enhance the robustness of the automatic speech recognition (ASR) model and reduce common errors. A postprocessing step was focused on an ASR model in a study by Lybarger et al. [36], where the transcribed text data were processed by adding punctuations, capitalizing letters where needed; repeated phrases were removed, gender dictations were corrected, and misspelt medical terms were updated. This postprocessing was performed by manual editing by physicians. Selvaraj et al. [45] represented that the Google text-to-speech API is not verbatim like IBM-STT and has a proper punctuation setting, which can add punctuation to the transcript.

3.4.7 Limitations

With the rapid development of technology, the use of ASR models has become versatile and more accurate. However, several drawbacks limit the functionalities of ASR models and need to be done manually. Table 7 visualizes a brief idea about some of the limitations that the studies mentioned.

Table 7 Limitations of the Proposed ASR

Model Used	Limitations and Drawbacks
Google speech-to-text API [19]	• Unable to capture some words, especially when the speaker has a heavy accent • more data and fine-tuning of parameters needed to

	improve the accuracy of noise suppression and transcription
Google Speech-to-Text model [20]	• Slower for the physical exam section concerning the history section due to typing error in the physical section • Speech recognition error and punctuation insertion errors resulting in a decrement in dictation speed
RNN-T for ASR and SD [21]	• No evaluation of performance gain from lexical cues • Punctuation and capitalization have not been introduced
Connectionist Temporal Classification (CTC), and end-to-end models with attention (LAS: Listen, Attend and Spell) [22]	• Missing out words in longer utterances • WER of patient audio was more than the WER of the doctor's speech as the audio recording device was closer to the doctor than the patient • The format (MP3) of audio was more likely to decrease the efficiency of the ASR model • The LAS model had less casual conversational terms compared to medical terms to learn, for which it was giving errors while transcribing casual terms
Google Cloud Speech API [29]	• Complication in converting Singapore English to text • Relatively small sample size
Naver Clova SR, Google Speech-to-Text, and Amazon Transcribe[35]	• Removal of words • Spelling mistake • Changing the spelling of the word resulting in different meaning
Dragon Medical Edition 2 by Nuance Inc. [36]	• Misspelled medical terms • Repetition of phrases • Capitalization errors • Wrong use of Punctuation and incorrect Acronyms

Table 7 (continued)

Model Used	Limitations and Drawbacks
Google Speech-to-Text Clinical Conversation (Google ASR) and Amazon Transcribe Medical (Amazon ASR) [37]	• ASR models could not transcribe the NLCS (Non-Lexical Conversational Sounds, specifically the clinically relevant terms) Had a very high WER value when NLCS were included
Neural Networks model (NNET3 of Kaldi, based on	• Error in recognizing numbers, which is a sensitive information

sequential models including I-Vectors module) [38]	
Amazon Web Services, Google Cloud, and IBM Watson [39]	• Australian English was available for Amazon Web Services and Google Cloud but not for IBM Watson • Lower accuracy of ASR for people with neurodegenerative disease
Google Cloud Speech to text API [45]	• High word error rate • The model can only process the English dialect

3.5 Text Summarization Techniques

3.5.1 Dataset Language and Features

Out of 68 finalized studies, a total of 16 studies [18–20, 44, 49–60] exhibited variations in the techniques employed for text summarization. Our study found a few research works that used datasets in Chinese [43], English [44], Arabic [50] and Indian [59] language for text summarization. As the conversation audio between the doctor and the patient is transcribed into a text file, it is further optimized as the text file has too lengthy and unnecessary lines and terms. The optimization is performed by executing text summarization of conversational text data using Natural Language Processing method [61]. Throughout such procedure, the conversational data is reduced to the point where important information such as patient history, disease symptoms, clinical notes, doctor's suggestions, medication suggestions, radiology reports, pathology reports, and discharge summaries are extracted and represented in a condensed version.

3.5.2 Types of Text Summarization Methods

Two primary methods can be used by NLP techniques to perform text summarization. These two basic methods are: (i) Extractive Summarization and (ii) Abstractive Summarization.

The extractive text summarization [51, 52, 56, 60, 62] techniques employ statistical or machine learning algorithms to discern significant sentences based on various factors, including word frequency, sentence length, and position within the document. One study [60] suggested that a digital scribe should aim to capture approximately 20% of the conversation to create a useful extractive summary. Analyzing patterns related to highlighted words, medical terminology, and speaker turns could assist in developing rules for information extraction. Wang et al. [20] used six extractive summarization methods, namely LexRank, TextRank, LSA, Luhn, SumBasic, and KLSum, for the summarization of clinical trial descriptions, among which the TextRank model gained maximum efficiency in summarizing the clinical trial text.

On the other hand, abstractive summarization involves the generation of novel sentences or phrases that encapsulate the essence of the source text. Abstractive summarization is considered to be a more intricate process compared to extractive summarization [63]. The latter involves selecting and merging existing sentences from the source text. On the other hand, Abstractive summarization necessitates a more profound comprehension of language and context and often utilizes complex NLP techniques. Using an encoder-decoder architecture, the Explainer Extractor model generates abstractive summaries from clinical data. It offers insights into the significance and relevance of patient information, helping medical professionals make informed decisions.

In contrast, the patient record (PR) extractor is a model that extracts essential information from patient records, such as diagnoses, medications, and laboratory results. In one study [57], text summarization was performed on patient care episode entries, where both the Explainer Extractor and PR extractor were performed, and the Explainer Extractor performed better than the PR extractor. In another study [55], the system was employed to automatically generate a discharge summary report for patients in the medical domain. The IDS has three main parts: In-Hospital Course Management, Discharge Medication, Instructions and Follow-up, and Discharge Symptoms. The model includes various techniques for preprocessing patient data obtained from electronic medical records. These techniques involve breaking down the text into coherent blocks of information, removing stop words that add noise to the data, reducing words to their root form to group similar words, and generating an array of keywords obtained from segmentation.

3.5.3 Statistical and mathematical models

One study [19] included the term frequency-inverse document frequency (TF-IDF) model for summarization tasks, where the text file was first preprocessed; through this process, the punctuation, stop words, and least used words were removed, and feature engineering was performed. TF-IDF vectors (TF-Term Frequency, IDF- Inverse Document Frequency) along with Word Count Vectors were utilized and later on classification and categorization of the features were performed using the SVM model.

The Behavioral Tree Framework (BT) [44] can be utilized as a behavioural modeling framework for cognitive assistants. The framework can be used for converting text, identifying important concepts, transforming them into vector space, selecting protocols, and suggesting protocol execution or intervention. Protocol execution and intervention suggestions are done through parallel nodes, where multiple applicable protocols are executed concurrently. Protocol nodes are sequential and follow the EMS protocols' conditions and logic sub-trees.

One study [49]used the NegEx tool for text summarization of clinical data. The n-gram technique involves categorizing the words in a sentence and creating a sequence of tokens from them. The value of n determines the number of words in each token. In medical scribing, negated disease filtration using NegEx helps to remove false positives, i.e. diseases or conditions that are mentioned in the text but are not present in the patient's medical history. The NegEx algorithm first identifies negation cues, such as "no", "not", "denies", and "without", which indicate that the medical concept mentioned in the text is negated. Then, it looks for medical concepts within a specific context window around the negation cue and marks them as negated.

3.5.4 Transformer-based Models

The Transformer language model [50, 53] is a deep neural network architecture commonly used in NLP for language translation and text summarization tasks. When applied to medical scribing, a Transformer language model can be trained on sizeable medical text datasets to learn patterns and relationships among medical concepts. The model can assist medical scribes in real time by suggesting the next likely word or phrase as notes are being typed out. Enarvi et al.[53]compared the effectiveness of RNNs and Transformer models for generating medical reports from patient-doctor conversations. The study showed that Transformer models are better at summarizing relevant clinical conversation excerpts and generating medical information that is factually correct and easily understandable, even when computing and time resources are limited.

One study [64]implemented the Well Behaved (WB) transformer model on Chinese clinical Named Entity Recognition on three different datasets where the model outperformed a lattice-LSTM model. In another study [65], a Multi-Granularity Transformer (MGT) was used to extract medical terms and their corresponding status from Chinese clinical dialogue. The proposed model incorporates character-level and word-level features and cross-turn interaction to capture semantic dependencies over long distances, improving state inference. The Multi-Granularity Transformer (MGT) is a type of transformer model that utilizes multiple granularities in its self-attention mechanism. This allows it to process long documents more efficiently by breaking them down into smaller chunks. MGT also uses a hierarchical architecture to optimize its performance further.

Two studies [54, 66] implemented the BERT model for text summarization. BERT (Bidirectional Encoder Representations from Transformers) can be used as the token-level encoder [54]in the text summarization where the model encodes the input text at the token level, which means that each word or sub-word in the input text is represented as a vector. The encoded tokens are then fed into the utterance-level encoder to generate representations for each utterance in the conversation. The representations are then tagged with labels indicating whether the utterance contains a problem statement or a treatment recommendation. Finally, the labeled declarations are concatenated to form the summary. Using BERT as the token-level encoder allows the model to capture the contextual information of each token, which is essential for generating accurate representations of the input text. In addition, BART (Bidirectional and Auto-Regressive Transformer) [58] can generate high-quality summaries of EHR and progress notes through extractive and abstractive summarization. The BART model is transformer-based. So, BART is based on a sequence-to-sequence Transformer architecture, while BERT is based on a bidirectional Transformer encoder.

The T5 (Text-to-Text Transfer Transformer) [67] is a large-scale language model that can be fine-tuned for various natural language processing tasks such as text classification, machine translation, and text summarization. One study included the T5 model for text summarization [47], generating a list of care plan problems using progress notes from a patient's electronic health record during hospitalization. Furthermore, domain adaptive pre-training (DAPT) was used to improve the performance of T5 on domain-specific data.

Song et al. [54] proposed a model where the Hierarchical Encoder-Tagging (HET) task summarizes medical conversations by identifying and tagging meaningful utterances containing problem statements or treatment recommendations. The HET model follows a hierarchical structure where the input utterances are first encoded at the token level using BERT and ZEN and then at the utterance level using an utterance-level encoder such as LSTM or Bi-LSTM.

3.5.5 Other Neural Networks

Sanjeev et al. [18] demonstrated the use of an NLP tool on unstructured data for tokenization and removal of stop words. Afterwards, a developed neural network containing 1 input layer with 12 nodes, 2 hidden layers with 16 and 18 nodes respectively and 1 output layer with 14 nodes acts as an AI chatbot and makes decisions on medications for diseases related to the heart.

Dialogue2Note [52] is a deep learning-based model that is designed to convert doctor-patient conversations into accurate and structured clinical notes. The model has an encoder-decoder architecture. The automated dialogue2note sentence alignment system can be used to create realistic training data for natural language generation systems. The dialogue2note snippet

summarization system involves generating a corresponding clinical note sentence given the gold standard dialogue snippet text. One study [59] included the COCOA system where relevant clinical information from Indian clinical records of an ICU in India was extracted using the model. In countries such as India, the adoption rate of structured clinical records is very low and the bulk of clinical documentation is still either paper-based or semi-structured electronic formats.

3.5.6 Metamap tool

In one study [44], the first step was UMLS Concept Extraction, which extracted biomedical concepts from the text alongside their negation condition, semantic type, and position information using MetaMap. Each concept was assigned a Clinical Unique Identifier (CUI) in the UMLS. Following this, Concept filtering was subsequently performed to filter out irrelevant concepts from the extracted set using a predefined set of specific concepts from the EMS protocol. Value Retrieval was then conducted to extract additional information related to the concepts, including their corresponding numeric values and preferred names. Among the evaluation metrics shown (see Table 8), the F1-Score can be calculated because compared with other performance metrics, namely accuracy, precision, and recall, the F1-score is the combination or mean value of recall and precision. Fig. 4 represents the F1 scores of the existing text summarization models.

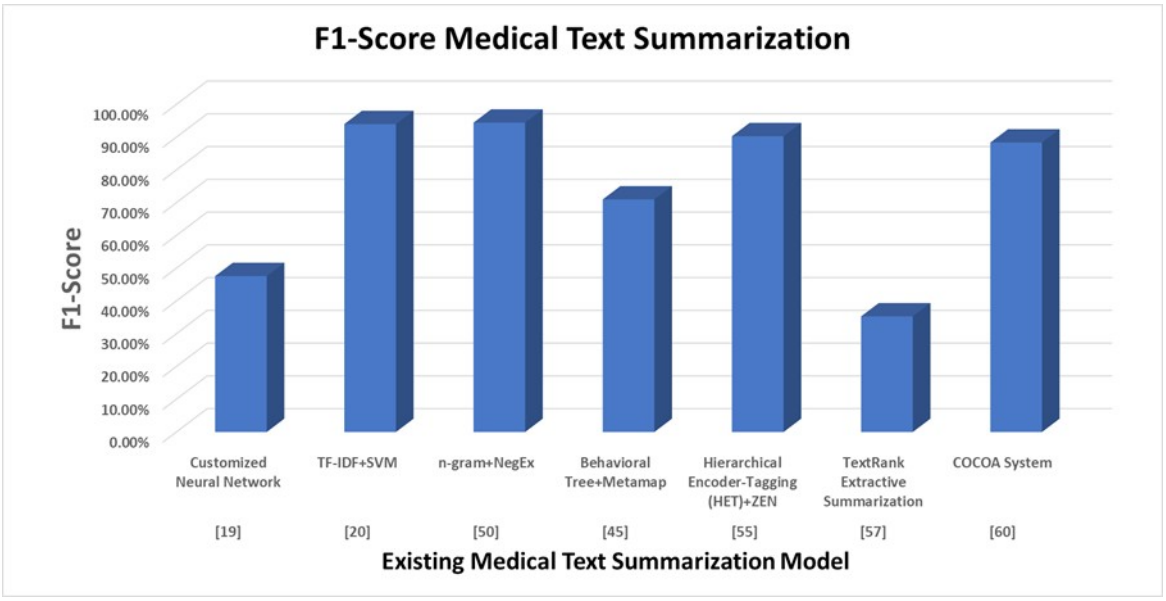

Fig. 4 F1-Scores of Existing Text Summarization Methods

Table 8 represents the dataset used, results and limitations of the research approach to clarify the existing studies

Table 8 Research works done on text summarization

Model	Dataset	Result	Limitation and drawbacks
NLP for preprocessing and customized Neural	16 hours of audio files and medical transcripts on the ezDI dataset	Prediction Accuracy: 88% Micro Precision:	• Prediction can be done on diseases only related to the heart

Networks for prescribing medication for heart diseases [18]		71.4% Micro Recall: 35.7% Micro F1-score: 47.6%	
TF-IDF and SVM for text categorization and labeling [19]		Mean Accuracy: 94.98% Precision: 94% Recall: 94% F1-score: 94%	● Important data lost in extractive summarization; physician input needed. ● The limited availability of labeled medical data.
N-gram technique and NegEx (Negated Disease Filtration) [49]	1050 notes in 20 various clinical domains including urology, nephrology, radiology, orthopaedic, and cardiovascular etc	Accuracy: 97.81% Precision: 98.50% Recall: 90.71% F1-score: 94.44%	● Not Mentioned
Behavior Tree Framework and Metamap [44]	8302 real EMS call records from an urban, high volume regional ambulance agency in the U.S.	Accuracy: 89% Recall: 66% Precision: 76% F1-score: 71%	● inaccurate concept detection ● need for more precise detection techniques ● accounting for contextual limitations during interventions
Dialogue2Note sentence alignment and snippet summarization [52]	Audio and clinical notes of 500 clinical visits	No particular result given	● content incongruence, ● clinical note-creation challenges ● inconsistent order of appearance ● a relatively small dataset
Transformer+Pointer Generator [53]	a dataset that consists of around 800k encounters from 280 doctors.	No particular result given	● repeated sentences ● unfounded clinical statements

Table 8 (continued)

Model	Dataset	Result	Limitation and drawbacks
TextRank Extractive Summarization [56]	The dataset containing 277,000 clinical trial records	Rogue-1 Recall: 0.3805 Rouge-1 Precision: 34.86%	● abstractive summarization methods to provide better performance ● The lack of context and

		Rogue-1 F1-score: 35.3% Rouge- F-score: 30.03%	complex sentence structures in some clinical trial descriptions. ● Study may not be representative of all clinical trial descriptions
Explainer Extractor [57]	1.7 M paragraphs with 48 unique headings	55% summaries adequate	● Further research is needed for its potential in other applications
T5 with DAPT model [58]	768 annotated progress notes from MIMIC-III	No particular result mentioned	● Use of a limited dataset for annotation and evaluation ● The dataset may also carry social bias features that can affect fairness and equity during model training ● progress notes may increase in length due to copy-and-paste behaviour ● The summarization task requires complex cognitive processes to arrive at an accurate diagnosis
COCOA System (Rule-based NLP system [59]	corpus of 250 annotated discharge summaries from an ICU in India	F1-scores for diseases/symptoms, procedures, and lab parameters: 0.856, 0.834, and 0.961 respectively	● No normalization of diseases, symptoms, or anatomical parts to an ontology ● the study did not use disease mentions ● The NLP engine is limited by spelling errors in the corpus, particularly for progression descriptions with odd punctuation.

3.6 Named Entity Recognition

Twenty-five studies demonstrated the application of NER [23, 46, 64, 65, 68–88]. The Named Entity Recognition (NER) technique extracts all the necessary information from unstructured text data which may include data such as symptoms, diseases, medications, procedures, dosage, frequency, route of administration, duration of treatment, anatomical location and many more. Utilizing the NER model various sub-tasks such as classification, annotation, deidentification,

relation extraction, recognition, normalization etc.

3.6.1 Language variation in the NER dataset

Diverse languages have unique grammatical structures, vocabularies, and idiomatic expressions, which pose challenges for natural language processing (NLP) algorithms to accurately capture and summarize information. The performance of automated scribing models in Named Entity Recognition (NER) can be affected by diverse languages, which can lead to variations in grammar, vocabulary, idioms, sentence structures, and homophones. Furthermore, accent, dialect, and regional language differences can further impact the accuracy of NER. Therefore, it is crucial to develop approaches that can handle the linguistic nuances of different languages and to train the NER models on diverse language data. Chinese [64, 65, 71, 73–77, 89], Spanish [79, 90], Swedish[90] , German [82, 83], and Russian [84] were seen to be used as datasets for Named Entity Recognition of medical documents. Frei et al.[83] described the creation of a German medical NLP model using multiple techniques, including translation, annotation projection, and transfer learning via model fine-tuning. The model performed well on external datasets and was trained on a small, task-specific dataset. The paper also discusses related work in the field of NLP for medical text and the difficulties of developing such models for German due to limited data and privacy concerns. Some of the challenges which were faced while building the model mentioned in the study were: limited and poor-quality datasets, the need for gold-standard annotated labels, undisclosed data in internal datasets, legal disputes over privacy concerns, and the absence of publicly accessible German medical datasets. Furthermore, limited research has focused on German medical NLP models due to the dominance of the English language.

3.6.2 Annotation and Labeling

However, 19 studies [23, 46, 57, 59, 60, 65, 67, 68, 70, 73, 76, 79–81, 83–85, 87, 88, 90] mentioned manual dataset annotation or evaluation for evaluation of their approached models. Annotation in medical scribes refers to the process of identifying and labeling specific entities, such as diseases, symptoms, treatments, and medications, in text data to train and test NER models. The resulting annotated dataset can be used to train and evaluate machine learning models, which can then be applied to new, unlabeled data to automatically extract relevant information. In one study [57], the experts were instructed to read the keyword summary, read all the original nursing entries from the care episodes, and finally return to the keyword summary to score it. The manual evaluation and statistical analysis were blinded to which method was applied. In another study by Weegar et al[90]. , a smaller set of text data was organized that had been manually annotated for clinical entities to train and evaluate their neural network for Named Entity Recognition. Specifically, the annotated corpus consisted of entities relevant to the Disease and Drug categories in Spanish, as well as those
relevant to the Body part, Disorder, and Finding categories in Swedish. In an approach by Singh et al. [23], two selected datasets were annotated and entities were grouped into seven classes: Name, Profession, Location, Age, Date, Contact, and ID. An Italian dataset was annotated by 3 Italian native speakers for de-identification in one study [87]. In another study [88], the dataset was first manually deidentified and annotated manually

3.6.3 Open-source NLP models

Shah-Mohammadi et al. [70] implemented and compared the performances of 3 models namely Spark-NLP, CLAMP and ACM. The clinical pipeline of CLAMP consists of a tokenizer, POS tagger, section identifier, deep learning-based named entity recognizer (LSTM-based deep learning model), assertion classifier, attribute recognizer, concept mapper, temporal recognizer, and temporal relation. Entities of various categories can be identified, including temporal, test, lab value, treatment, drug, strength, negation, problem, subject, body location, severity, family history, history, dosage, form, route, frequency, condition, duration, course, and generic.

3.6.4 Transformer based models

Two studies utilized [64, 65] transformer models for NER tasks. In one study, Zhang et al. [64]developed a transformer model called WB-Transformer for Chinese medical NER using encoding. This model encoded characters and words from Chinese electronic medical records (EMRs) separately, reducing word segmentation errors and providing word boundary information. WB-Transformer was more efficient than Lattice-LSTM, even at batch size 1. Li et al. [65] proposed an MGT model for Chinese clinical conversation extraction using character, word, and sentence-level interaction information in model generation. This method helped the model capture semantically-dependent long-distance data.

In eight studies [72–74, 80, 82, 84, 85, 88] the BERT model was used for NER. There can be several versions of BERT models that spend time on the structure and the dataset in which they have been trained. BioBERT is trained on a large biomedical text corpus and can be fine-tuned for various biomedical natural language processing tasks, such as named entity recognition and biomedical relation extraction. Meripo et al. [80] introduced appointment span extraction from the medical conversation and represented this as a sequence tagging problem in which the hybrid bioBERT model was used. The BioBERT hybrid model utilized in this research employed a combination of two weak supervision techniques, namely inaccurate and incomplete supervision. The model was initially trained with inaccurate supervision on a sizable dataset of conversations that contained labels with varying degrees of noise. Subsequently, the model was fine-tuned with incomplete supervision on a limited subset of conversations that feature manually annotated labels. This method enabled the model to acquire more precise and reliable predictions while reducing the need for extensive manual annotation. The approached model outperformed the ELMo and other BERT variants. Another BERT model that is pre-trained on biomedical text is SciBERT. Memarzadeh et al. [81] demonstrated the use of the SciBERT model for clinical keyphrase extraction. The SciBERT is almost similar to BioBERT, but it has applicability towards a wider range of tasks. One study [83] demonstrated the utilization of GottBERT and German BERT for clinical annotation in Germany, where translation was also performed. The primary distinction between GottBERT and German BERT is the pretraining dataset's size, where GottBERT was trained on a larger dataset (145GB) compared to German BERT's 12GB. The German BERT model exhibited inferior performance compared to GottBERT due to differences in pretraining dataset sizes. The model is an updated version of GERNERMED demonstrated by the same author in a previous study [82]. Yalunin et al. [84] introduced two models, RuBioRoBERTa and RuBioBERT, for performing Named Entity Recognition (NER) in the medical domain of Russian language text. The RuBioRoBERTa model and an improved version of the RuBioBERT model were constructed with additional pretraining techniques. Nesterov et al. [85] used the EhrRuBERT model, a specialized BERT model specifically trained for Electronic Health Record data for clinical Named Entity Recognition. The model could generalize entities and perform human-level extraction efficiently from the most frequent data. The proposed model was trained with binary

cross-entropy loss and the Adam optimizer.

3.6.5 Mathematical and graphical models

The TF-IDF model has applications in NER tasks. Memarzadeh et al. [81] utilized the TF-IDF model along with SciBERT for the extraction of key concepts from clinical documents. The research evaluated various unsupervised techniques for extracting keyphrases from clinical texts and concludes that statistical approaches, as well as the TF-IDF method, outperform machine learning methods such as KeyBERT.

Conditional Random Fields (CRF) [91] is a type of probabilistic model that can capture contextual information to accurately predict named entities in text. By considering the dependencies between labels and input features, CRF can effectively model these relationships and make accurate predictions. Furthermore, CRF models can incorporate different features such as word embeddings, morphological features, and contextual features to improve performance. Nine studies [23, 69, 72, 74, 75, 79, 87, 89, 90]evaluated CRF for performing medical Named Entity Recognition tasks. Liu et al. [89] implemented CRF on Chinese electronic medical records. Features namely bag-of-characters, part of speech, dictionary features, and word clustering features were observed and their features were examined. In another study [90], CRF was used as a sequential tagger to predict the label of a word based on the word's context and the labels that had been assigned to the previous words in the sequence. Singh et al.[23] demonstrated an ensemble learning model which included the CRF model for the deidentification of clinical data from diverse locations.

3.6.6 Neural Network

Long Short-Term Memory (LSTM) is a subtype of recurrent neural networks (RNN) which have been developed to overcome the issue of vanishing gradients observed in conventional RNNs. The LSTM networks are proficient in learning long-term dependencies in sequential data by selectively retaining or discarding information over time. Bi-LSTM is a variation of LSTM that captures both the past and future context of the input sequence. It is often employed in natural language processing tasks, like sentiment analysis and named entity recognition, to enhance the model's performance. Seven studies [71, 72, 75, 76, 79, 87, 90] introduced Bi-LSTM structures for Named Entity Recognition tasks. Weegar et al. [90] proposed a Bi-LSTM model with word embedding for Spanish and Swedish NER to extract information from clinical texts which improved the performance of the NER method compared to the previous method. Table 9 shows the model involved in the NER process along with the applicated dataset and their accuracy. includes some limitations of these NER approaches.

Table 9 Models Established on Named Entity Recognition

Model	Dataset	Accuracy

Med7 NER model [68]	2018 n2c2 challenge (train), MIMIC-III (train) and OxCRIS dataset (test)	F1-score: 95.7% (94.4%) Precision: 0.982 (94.1%) Recall: 0.933 (94.7%) on n2c2 gold test set (OxCRIS)
Well-Behaved Transformer Model+Self Attention [64]	CEMR (China Electronic Medical Record) dataset, CCKS (China Conference on Knowledge Graph and Semantic Computing) 2019 benchmark dataset, and ALCD (Alibaba Cloud Labeled Chinese Dataset for diabetes) dataset.	Recall: 82.47% F1-score: 83.29%
CLAMP model (LSTM-based deep learning model) [70]	Track2 i2b2 2014 NLP challenge dataset. The dataset consists of 521 medical texts in XML format	Precision: 97% F1-score: 90%
ZEN2-base+BiLSTM+CRF+Radical features [71]	About 200,000 Chinese medical QA from ChiMed [92]	F1-score: 84.43%
MLNER+BERT [73]	1000 marked data from Chinese medicine's manual from TianChi Entity Recognition	F1-score: 70.87% Recall:76.39% Precision: 66.09%
Bi-LSTM with word embedding [90]	Spanish corpus (disease, drug) of 342 MB, Swedish corpus (disorder, finding, body part) of 1.2 GB data	Average F-measure: 71.90% (Spanish) 75.96% (Swedish)
TsERL model (Bi-LSTM+CRF) [75]	Two public Chinese medical NER datasets: CCKS2021 and CmeEE	F1-Score (CCKS2021): 84.95% Recall (CCKS2021): 86.56% Precision (CCKS2021): 83.39%
Multilayer BiLSTM [76]	3588 sentences of real Chinese medical dataset containing 5 types of entities: "symptom", "check", "check result", "disease", and "treatment"	F1-score: 78.69% Recall: 77.8% Precision: 79.6%
BiLSTM+CRFlayer [79]	Spanish lung cancer dataset	F1-score: 90%

Table 9 (Continued)

Model	Dataset	Accuracy
BioBERT hybrid model [80]	24k hand-written transcripts of de-identified medical conversations	Precision: 77.23%, Recall: 81.53%, and F1 score: 79.32%
Multi-Granularity Transformer (MGT) model [65]	Chunyu dataset: 1,120 dialogues from a Chinese medical website; CMDD dataset: 2,067 conversations from a Chinese online pediatric health community.	F1-score: 0.877 (Chunyu dataset); 0.771 CMDD dataset)
GERNERMED (combination of rule-based methods and machine learning models) with default SpaCy parameters [82]	n2c2 German dataset (8599 sentences with 172695 tokens)	Average F1-score: 81.54%
GERNERMED++ (GottBERT) [83]	German translated 2018 n2c2 shared task on Adverse Drug Event (ADE) and medication extraction	Precision: 92.4% Recall: 95% F1-score: 94.6%

Fig. 5 shows a comparison of F1-Scores among existing medical NER methods for better understanding.

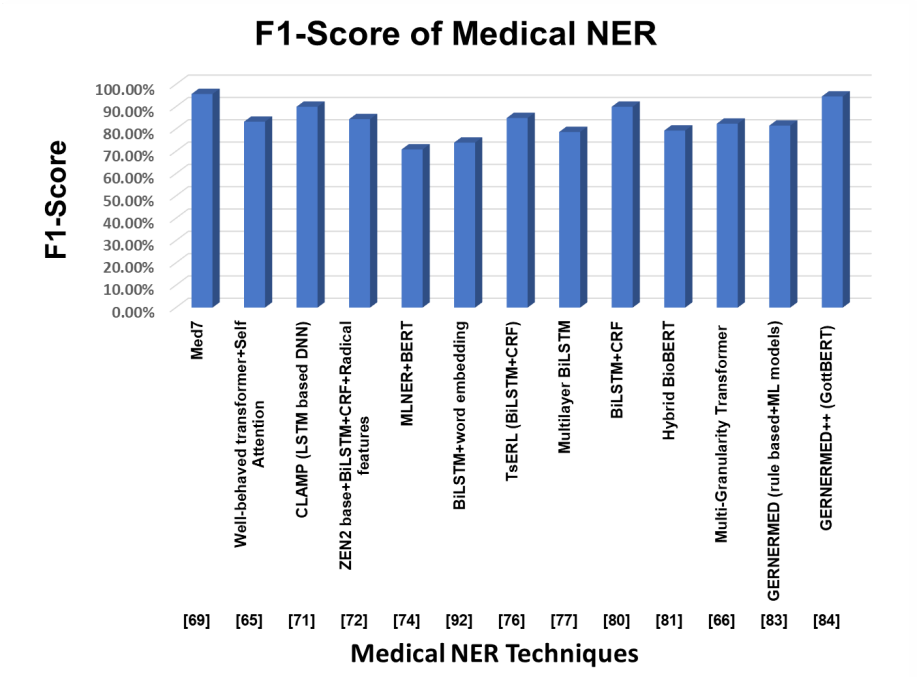

Fig. *5* F1-Score of Existing NER Techniques in Medical Domain

Table 10 Limitations of some of the NER approaches

Paper	Limitations and drawbacks
[80]	1. Limited set of annotations 2. Difficulty in spotting confirmation clues 3. Ambiguity in appointment reason and time mentions 4. Transcription error rate in automatic transcripts 5. Lower F1 score for time span extraction
[81]	1. Tokenization issues 2. UMLS assignment is error-prone without filtering. 3. MIMIC-III mortality rate is 23.2%, higher for older patients. 4. Repetitive and low-value concepts
[83]	1. limited understanding of how well the models generalize to unseen data 2. Scarcity of independent annotated datasets 3. The creation of gold-standard annotations is an expensive task 4. character-level annotation may not preserve the same level of granularity as in other types of annotation.

3.7 Prevention of Patients Privacy and Ensuring Security

In accordance with the regulations set forth by the Health Insurance Portability and Accountability Act (HIPAA) [93]n the United States, the anonymity of patients must be ensured when generating automated medical documents. Such measures serve to enable healthcare organizations to uphold patient privacy and confidentiality. Such measurement prevents several incidents such as discrimination, stigmatization, and identity theft. Medical de-identification refers to a series of techniques and procedures aimed at the anonymization or obscuring of personal identifying information that may be present within medical records or other forms of health-related data, thus effectively mitigating the risk of inadvertent disclosure of sensitive patient information which may include patient's name, address, profession, IDs, date, social security number, insurance details, and contact information.

3.7.1 NLP Models for De-identification

Nine studies [23, 46, 72, 86–88, 94–96] have incorporated deidentification as part of their methodology. The de-identification approaches were established based on different regional datasets which include from the USA [94], New Zealand [23], Italy [87], France [88], and Australia [72, 95]. Singh et al[23]. proposed an approach where the deidentification of protected health information (PHI) in electronic health records (EHR) was performed using an ensemble framework which combined the rule-based CRF model and incremental Naive Bayes algorithm. Two datasets were used in this model: the publicly available i2b2 dataset and the local dataset collected in New Zealand. The diversified datasets enabled the model to perform in various clinical settings. In another study [86], Clinical reports across different stages of prostate cancer management, such as consultations, on-treatment visits, phone encounters, treatment records, and follow-ups, were collected and subjected to de-identification. The corpus was tokenized by the NLTK tokenizer [97] and later on the punctuations and numbers were removed. The processed data were subsequently embedded using the Word2vec model, and binary classification was

performed to classify the name and non-name data. Catelli et al. [87] combined the Bi-LSTM and CRF models as a sequence architecture to perform de-identification on the Italian dataset from the Society of Radiology (SIRM). The model with stacked embedding achieved an F1-score of 98.6%, outperforming a state-of-the-art BERT BASE model in all evaluation metrics except for the binary token level.

Cohn et al. [46] proposed a de-identification task of public health information (PHI) on both medical conversation text data and audio. With respect to the text data, NER models such as rule-based models and machine learning models were applied. For de-identifying audio data Automatic Speech Recognition (ASR), Named Entity Recognition (NER) on the transcript, and aligning the text to the audio were established. The objective of this de-ID task pipeline was to generate a modified audio stream that removes Personal Health Information (PHI) from the original input. Tchouka et al. [88] proposed a transformer-based BERT model for the French language named CamemBERT on a large set of unlabeled medical notes obtained from a French public hospital. Hyperparameter tuning was later on done using a Tree-structured Parzen Estimator algorithm while training the model. Another de-identification approach was made in one study [72], where de-identification was done on Personally Identifiable Information (PII) from Australian hospital discharge summaries. The study applied an ensemble model which stacked SVM models with 3 (BiLSTM + CRF, CNN + CRF, and BERT). Although several advances have been made in terms of deidentification techniques, due to the limited datasets available, the models could not perform well in different environments. El-Hayek et al.[95]utilized four existing de-identification tools (HMS Scrubber, MIT De-id, Philter, MIST) on Australian patient progress notes and evaluated their performance. Although the Philter model gave the best output, it was suggested that the existing models are not suitable enough and further modifications are required.

Table 11 Limitations of the proposed deidentification methods

Research Work	Limitations and drawbacks
[23]	● utilized model made use of basic lexical and orthographic features, failing to fully leverage the corpus's comprehensive attributes. ● Validation set is similar to text set which may not always be the case ● Only F1- score was measured ● More data was required for risk-based evaluation method
[88]	● The substitution strategy for date and age information is limited ● The proposed approach may not perform as well on informal medical documents ● The effectiveness of the proposed approach may depend on the specific dataset
[72]	● sensitive information such as genetic data, social security numbers, and email addresses were not considered ● Proposed model needs to be applied to other healthcare centers ● the study did not address the ethical and legal issues related to de-identification ● No assessment of de-ID impact on clinical research or quality.

3.7.2 GPT-4 on De-identification:

Recent success of AI and machine learning models have spread towards diversified fields which led to the establishment of ChatGPT and GPT-4. Liu et al.[96] introduced DeID-GPT, a de-identification tool based on ChatGPT and GPT-4 which showed remarkable results in anonymizing patients' sensitive information. ChatGPT and GPT-4 are Large Language Models which are trained on very large data. In this study, the model was evaluated using the 2014 i2b2/UTHealth de-identification challenge dataset [98] where the DeID-GPT outperformed all the other LLM models showing accurate performance in diversified data and overcame the limitations of previous LLM models. The approached model gained an accuracy of over 99% showing optimal results in a zero-shot scenario.

4. Discussion

4.1. Study Components

A systematic review involves gathering information from published studies related to medical scribes, analyzing the data, and presenting a concise and accurate summary of the findings. The comparative analysis revealed that a generic medical scribing model contains 3 major components for creating a complete prescription from a doctor-patient conversation. These components include ASR, text summarization and NER. While there has been limited progress toward the development of a fully automated medical scribe, numerous studies have investigated various components of the system in isolation by applying different techniques.

4.2. Key Factors

For each major subtask, several factors need to be considered. Rather than solely assessing the performance of established models, our study has placed greater emphasis on the adaptability and versatility of the proposed techniques across various linguistic, regional, and clinical contexts. Through utilizing these techniques across different regions and settings, their robustness and effectiveness in a wide range of scenarios are observed. This is particularly significant, as the use of electronic healthcare systems continues to expand, and there is a growing need for effective techniques that can be applied across a variety of contexts. We conducted an empirical analysis to investigate and contrast the effectiveness of various models pertaining to diverse subtasks, datasets, regions, and environments. Additionally, we explored how the performance of these models could be enhanced through the integration of complementary models. Our research methodology was designed to address the research questions that were previously articulated and was structured to provide meaningful responses to those inquiries through the utilization of a variety of research approaches selected from our search queries. Our research questions addressed the influence of demographic factors, language, clinical setup, environment and population on the tasks related to digital scribing.

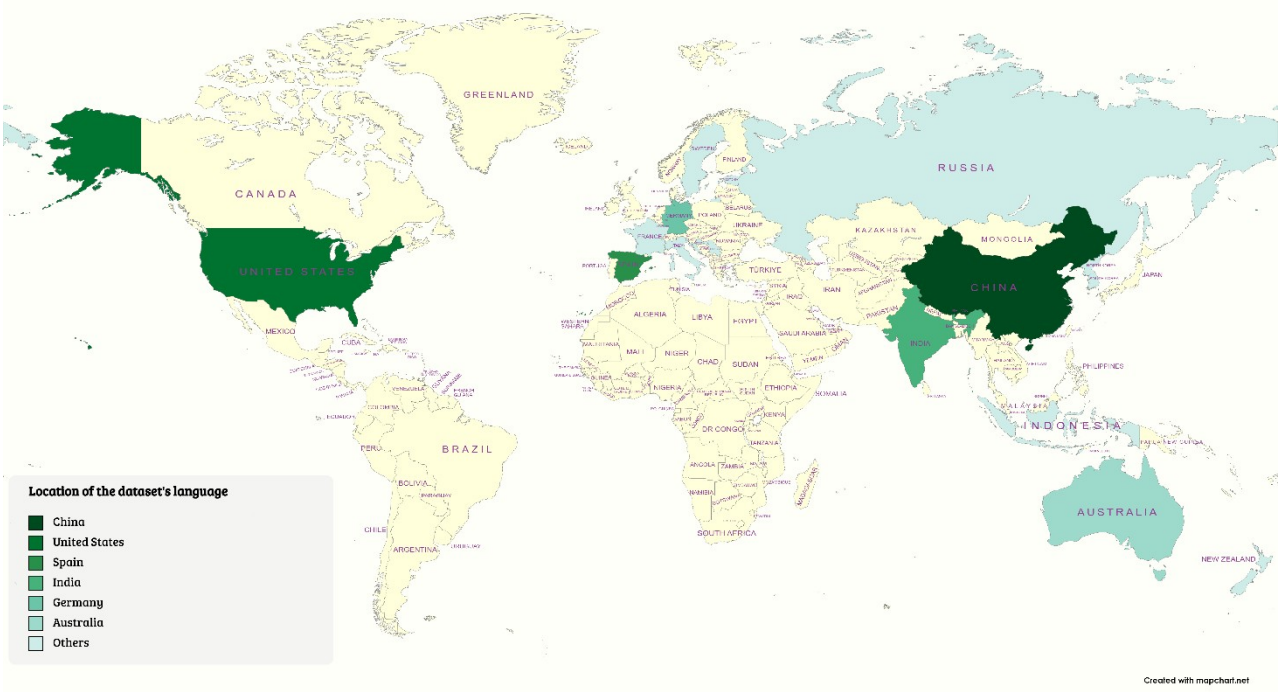

Fig. 6 Geographical Representation of Regions from which datasets of the selected articles have been collected

4.3. Key Findings of the Models

With the rapid development of natural language processing and automatic speech recognition, an increasing number of technologies are becoming accessible for medical documentation purposes. The observed approaches were evaluated through several evaluation metrics. The ASR models were evaluated through the score of WER (Word Error Rate), WDER (Word-level Diarization Error Rate), and BLEU (Bilingual Evaluation Understudy). Among these metrics, the WER is considered to be the most ideal evaluation metric. The research we have elected for ASR has the WER ranged from 1.8% up to 38.3% even though one study [37] had the WER of 40.8% (Google text to speech API) and 57.2% (Amazon ASR) due to the presence of non-lexical conversational sounds (NLCS). After assessing the Word Error Rate (WER) of 20 different models chosen from the identified approaches, it has been determined that the mean WER of all the evaluated models is 22.9935% which is quite high compared to an available adequate state-of-the-art ASR model which has the WER of approximately 5% [99]. Most of the ASR models utilized Google speech-to-text API for the transcription. Among all the studies related to ASR models, 42.31% of studies used the Google speech-to-text API model. The selected models are trained by utilizing datasets that vary in size from minimal to medium of different languages. The prime reason for the variation is the versatility in languages and accents; and the presence of lexical resources. Besides that, the clinical setup, surrounding environment, and background noise have an impact on the performance of the ASR model. The lowest WER recorded is 1.8% where 904 hours of Serbian and Croatian audio datasets were used on an RNN speech model. On the other hand, high WER (38.3%) was found due to the bulk size of the English dataset [46]. Thus, the language and size of the dataset play a vital role in the performance of the ASR models.

Approaches that made significant progress [18, 21, 26, 29, 30, 37, 40, 100] in transcription utilized Google text-to-speech API, Kaldi toolkit or integrated RNN-based methods with preprocessing and postprocessing steps. Noise reduction, punctuation addition & truecasing have

been performed as preprocessing and postprocessing steps in some new research studies. One study also mentioned downsampling the audio to 16 kHz to mitigate noise and get a clear audio file. The ASR models are capable of performing real-time transcription of audio conversation which can reduce the time consumption significantly compared to manual scribing.

From the study on the existing text summarization models, it has been found that the extractive text summarization models were utilized in more approaches compared to the abstractive text summarization methods even though the abstractive text summarization method shows superior outcomes. 4 major evaluation metrics were measured in most of the approaches which are accuracy, recall, precision, and F1-score. There were a few variations in languages for the text summarization task. We have considered the F1-score for evaluating the performance of the model. The F1-Scores were between 35.31% to 96.1%. The F1-score dropped by a big margin in the Textrank [56] extractive summarization technique. Overall, the average F1-Score for all text summarization models was 78.78%. Transformer-based models are seen to be applied in most approaches such as the BERT model, and T5 model. However, superior results were seen in the rule-based COCOA model.

In our findings, entity extraction of doctor-patient conversation included categories namely drug names, route of administration, frequency, dosage, strength, form, duration, disease name, symptoms, abnormal inspection result, test, treatment, degree of illness, body part, and family history. Furthermore, some rare categories such as diabetics, coronary artery disease (CAD), hypertension, hyperlipidemia, obesity, smoking status, image inspection and other non-medical entities. There were quite variations in dataset languages for performing the NER task. In most studies, Chinese datasets were seen to be used. A total of nine studies used Chinese datasets for training the NER models. Among the models used for clinical NER tasks, various types of BERT models have been observed to be used and the BERT models are utilized in most studies. In addition, the CRF models are also seen to be used in an equal number of studies. For evaluating the performances of the NER models, we have selected the F1-score to be the benchmark. The F1-score is combined with both precision and recall which gives a more balanced measurement of the confusion matrix. The score gives a more precise evaluation while one class is more dominating over the other one and the precision and recall results are not compatible. It was found that the F1 scores of the NER models ranged from 70.87% to 95.7%. On the Med7 NER model [68], the dataset was trained on MIMIC-III and had an average macro F1-score of 95.7%. The model was furthermore tested on the OxCRIS dataset and achieved an F1-score of 94.4%. The models performed quite well in entity extraction tasks with satisfactory results. The NER task in the German language has shown improvement where the GERNERMED++, and upgraded German NER model of GARNERED showed a 16.0167% improvement in the F1-score. explains the performance of the approached models for the NER tasks.

In both text summarization and NER tasks, MIMIC-III models were seen to be used in several approaches [58, 68, 81, 83]. Pretraining the model with this dataset made the model fine-tuned and increased its performance. The transformer-based models could also perform well in de-identification tasks but could not give a reliable result and were bound to limited datasets. But the failings were overcome by the latest implementation of DeID-GPT by GPT-4 which gave extraordinary results which proves the model to be reliable on almost any sort of dataset. The approached model outperformed results of all the existing models.

Table 12 Different NER Tasks done by the approached models

Model Name	Reference	Task Performed
Transformer Model	[64]	Well Behave Transformer model ● employed an encoding method that effectively encodes Chinese EMR characters separately. ● used a layer that models both word and character features, followed by decoding that considers character relationships. ● obtained information on word boundaries while minimizing the influence of errors in word segmentation
	[65]	Multi-Granularity Transformer ● merged word-level and character-level strings ● aggregated the representation of the strings ● measured the distance between tokens
TF-IDF	[81]	● performed keyphrase extraction ● evaluated the significance of each term in a document by measuring its frequency as well as its rarity across all documents ● No use of tokenization, which ensured that the meaning of phrases composed of multiple words was maintained
BERT	[81]	Sci-BERT ● Represented extracted keywords ● Gave optimal outcome in text classification ● Sequence labeling ● dependency parsing
	[80]	BioBERT ● Enhanced appointment span extraction ● Pre-trained on large PubMed abstracts ● Outperformed ELMo and BERT variants
	[83]	GottBERT and GermanBERT ● Performed German medical NER
	[84]	RuBioRoBERTa ● Performed Russian medical NER ● Classifying words from drug review ● Predicted diseases from patient symptoms ● Suggested relevant symptoms ● Provided binary response on a medical given question ● Made relation between a statement and hypothesis

Table 12 (continued)

Model Name	Reference	Task Performed
	[85]	EhrRuBERT ● Trained medical tokenizer learns specialized token embeddings from medical data ● Applied dynamic class weightings for faster training ● Utilized the entire text representation to extract all entities simultaneously.
CRF	[69]	● Followed normalization and expand system to extract biomedical and clinical terms
	[89]	● Performed Chinese medical NER such as bag-of-characters, part of speech, dictionary feature, and word clustering feature
	[74]	● Ignores structural errors in the output of the preceding feature extraction
	[90]	● Performed as a sequential tagger to predict entity labels for current words based on contextual representation and previously predicted labels. ● utilized dense characterizations, well-suited for clinical text containing extensive vocabularies and numerous less commonly used words.
	[75]	● Found the interdependence between consecutive labels in the NER task
Bi-LSTM	[71]	● Integrated information by merging forward and backward LSTM ● encoded the semantic features of the current word.
	[90]	● Acquire contextualized representations of the input words. ● Obtained word and character embedding ● Merged both character and word vectors ● Created representation for out-of-vocabulary words
	[76]	● Captured contextual information at various depths from given text. ● Improved the semantic understanding through different sizes of hidden layers ● Extracted local contextual features ● Found contextual semantic features within the text through the Semantic Feature Extraction Layer ● Capture word-level features

4.4. Recommendation

Even though several models have shown promising results, there has been seen the scarcity of datasets which has affected the performance most. The datasets were of a specific language and accent for which created a barrier in the diversity of the models. Thus, for future improvement, Furthermore, the models seemed to perform poorly when non-lexical terms were included in the test set. For tokenization, there can be out-of-vocabulary words which don't seem to exist in tokenizer vocabulary due to the small size of the dataset. The primary suggestions we can provide for future endeavor on medical scribing include:

- The models have to be trained in which large corpus and variations in accent have to be made. Furthermore, more non-lexical terms are needed to be added in the dataset on which the dataset will be trained.
- The frequency of the audio should be properly scaled for the trained model
- Audio needs to be recorded on an efficient recorder which can minimize the noise level of the recording and capture clean audio. Furthermore, the clinical setting and environment around the recording area should be kept as noise-free as possible.
- If noise is also recorded, it needs to be suppressed using the adequate noise-reducing tool as a preprocessing step of the ASR model. Also, the ASR models should be trained with both clean and noisy audio.
- Proper annotations are needed on datasets. Thus, the annotations should be gold standard annotations which will include topics like part-of-speech tags, named entities, sentiment labels, and semantic relations
- As the transformer models are performing more stably, they should be used and merged with other models for better results.
- Latest innovations like GPT-4 have been seen to be implemented on a large scale of data with various languages for de-identification. Such approaches should be made on other tasks such as NER and text summarization.
- For out-of-vocabulary words or unknown tokens, generation of similar words is required. Generated models such as Hidden Markov models, and Gaussian Mixture models should be incorporated.

5. Conclusion

The recent development on medical scribing is accelerating at a very slow pace with up-to-date NLP models. Utilizing three tasks successively (ASR, text summarization and NER) can produce a complete medical document. From our review, we have been able to understand that the ASR, NER and text summarization models have not been able to show satisfactory results for which these models can be employed practically but the progress is visible and such practical implementation can be hoped to be observed in near future. The models performed depending on the datasets they were trained on. Among the ASR models, Google text-to-speech API was used the most and the model performed moderately. But the neural network-based models showed highest efficiency having the lowest value of WER. On the other hand, the CRF, BERT, Bi-LSTM and other transformer-based models were most commonly used NER models which were used for

extracting keywords of specified categories, tokenization, creating entity relations, labeling and many more. For text summarization various models were used of which the BERT model had quite a few utilizations. Besides, transformer language models were also used. Datasets containing the personal information of patients in different languages were subjected to de-identification techniques to ensure the preservation of their privacy. Our review also showcases factors which affect the training model. To mitigate the losses and optimize the performance of the training model, preprocessing and postprocessing were performed on the factors. From the limitations observed from the research work, our review work concludes that more work needs to be done on healthcare sectors based on AI and NLP tools to establish easier and more convenient ways of medical documentation. As technology is advancing faster and AI-based tools are being used more and more, it can be hoped that the technologies are soon to be incorporated in healthcare sectors to improve the clinical documentation process.

Declarations

Acknowledgements The authors would like to thank Institute of Advanced Research (IAR) for funding and supporting this research. Project code: UIU-IAR-01-2022-SE-24.

Author Contribution The first and second author spearheaded the study design, material preparation, data analysis and collection and the first author also wrote the first draft of the manuscript. The third author contributed to the final preparation of the manuscript. All other authors involved in the conception of the study, read, advised and approved the final version of the manuscript.

Competing Interests: The authors declare no competing interests.

6. Reference

1. Weiner JP (2012) Doctor-patient communication in the e-health era. Isr J Health Policy Res 1:33. https://doi.org/10.1186/2045-4015-1-33
2. Cowie MR, Blomster JI, Curtis LH, et al (2017) Electronic health records to facilitate clinical research. Clinical Research in Cardiology 106:1–9. https://doi.org/10.1007/s00392-016-1025-6
3. Gellert GA, Ramirez R, Webster SL (2015) The Rise of the Medical Scribe Industry. JAMA 313:1315. https://doi.org/10.1001/jama.2014.17128
4. Shanafelt TD, West CP, Sinsky C, et al (2022) Changes in Burnout and Satisfaction With Work-Life Integration in Physicians and the General US Working Population Between 2011 and 2020. Mayo Clin Proc 97:491–506. https://doi.org/10.1016/j.mayocp.2021.11.021
5. Pozdnyakova A, Laiteerapong N, Volerman A, et al (2018) Impact of Medical Scribes on Physician and Patient Satisfaction in Primary Care. J Gen Intern Med 33:1109–1115. https://doi.org/10.1007/s11606-018-4434-6
6. Finley G, Edwards E, Robinson A, et al (2018) An automated medical scribe for documenting clinical encounters. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 11–15
7. "Ambient clinical intelligence: the exam of the future has arrived. ," Nuance Communications, 2019. https://www.nuance.com/healthcare.html (accessed May 12, 2023).
8. “Robin Healthcare | automated clinic notes, coding and more.,” Robin Healthcare, 2019. https://www.robinhealthcare.com/ (accessed May 12, 2023).
9. “DeepScribe - AI-Powered Medical Scribe,” Deepscribe, 2017. https://www.deepscribe.ai/ (accessed May

12, 2023).
10. “Amazon comprehend medical,” Amazon Web Services, Inc, 2018. https://aws.amazon.com/comprehend/medical/ (accessed May 12, 2023).
11. Page MJ, McKenzie JE, Bossuyt PM, et al (2021) The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. International Journal of Surgery 88:105906. https://doi.org/10.1016/j.ijsu.2021.105906
12. Blackley S V, Huynh J, Wang L, et al (2019) Speech recognition for clinical documentation from 1990 to 2018: a systematic review. Journal of the American Medical Informatics Association 26:324–338. https://doi.org/10.1093/jamia/ocy179
13. van Buchem MM, Boosman H, Bauer MP, et al (2021) The digital scribe in clinical practice: a scoping review and research agenda. NPJ Digit Med 4:57. https://doi.org/10.1038/s41746-021-00432-5
14. Ghatnekar S, Faletsky A, Nambudiri VE (2021) Digital scribe utility and barriers to implementation in clinical practice: a scoping review. Health Technol (Berl) 11:803–809. https://doi.org/10.1007/s12553-021-00568-0
15. Falcetta FS, de Almeida FK, Lemos JCS, et al (2023) Automatic documentation of professional health interactions: A systematic review. Artif Intell Med 137:102487. https://doi.org/10.1016/j.artmed.2023.102487
16. Quiroz JC, Laranjo L, Kocaballi AB, et al (2019) Challenges of developing a digital scribe to reduce clinical documentation burden. NPJ Digit Med 2:114. https://doi.org/10.1038/s41746-019-0190-1
17. Sahney R, Sharma M (2018) Electronic health records: A general overview. Curr Med Res Pract 8:67–70. https://doi.org/10.1016/j.cmrp.2018.03.004
18. Sanjeev S, Sai Ponnekanti G, Pradeep Reddy G (2021) Advanced Healthcare System using Artificial Intelligence. In: 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, pp 76–81
19. Menon NG, Shrivastava A, Bhavana ND, Simon J (2021) Deep Learning based Transcribing and Summarizing Clinical Conversations. In: 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). IEEE, pp 358–365
20. Wang J, Lavender M, Hoque E, et al (2021) A patient-centered digital scribe for automatic medical documentation. JAMIA Open 4:. https://doi.org/10.1093/jamiaopen/ooab003
21. Shafey L El, Soltau H, Shafran I (2019) Joint Speech Recognition and Speaker Diarization via Sequence Transduction. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. ISCA, ISCA, pp 396–400
22. Chiu C-C, Tripathi A, Chou K, et al (2018) Speech Recognition for Medical Conversations. In: Interspeech 2018. ISCA, ISCA, pp 2972–2976
23. Singh B, Sun Q, Koh YS, et al (2020) Detecting Protected Health Information with an Incremental Learning Ensemble: A Case Study on New Zealand Clinical Text. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp 719–728
24. Kuber A, Kulthe S, Kulkarni Y, et al (2022) Extensive Study of Automatic Text Summarization on Biomedical Texts. In: 2022 6th International Conference On Computing, Communication, Control And Automation (ICCUBEA. IEEE, pp 1–7
25. O’Shaughnessy D (2008) Invited paper: Automatic speech recognition: History, methods and challenges. Pattern Recognit 41:2965–2979. https://doi.org/10.1016/j.patcog.2008.05.008
26. Pakoci E, Pekar D, Popovic B, et al (2022) Overcoming Data Sparsity in Automatic Transcription of Dictated Medical Findings. In: 2022 30th European Signal Processing Conference (EUSIPCO). IEEE, pp 454–458
27. Laksono TP, Hidayatullah AF, Ratnasari CI (2018) Speech to Text of Patient Complaints for Bahasa Indonesia. In: 2018 International Conference on Asian Language Processing (IALP). IEEE, pp 79–84
28. Badlani S, Aditya T, Dave M, Chaudhari S (2021) Multilingual Healthcare Chatbot Using Machine Learning. In: 2021 2nd International Conference for Emerging Technology (INCET). IEEE, pp 1–6
29. Xu S, Yang Z, Chakraborty D, et al (2019) Automated Lexical Analysis of Interviews with Schizophrenic Patients. 9th International Workshop on Spoken Dialogue System Technology 185--197
30. Jiang Y, Poellabauer C (2021) A Sequence-to-sequence Based Error Correction Model for Medical

Automatic Speech Recognition. In: Proceedings - 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021. Institute of Electrical and Electronics Engineers Inc., pp 3029–3035
31. Signal JOA-P, Summit IPAssociationA, Tokyo C, et al 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) : proceedings : 14-17 December 2021, Tokyo, Japan
32. Gyllensten AC, Sahlgren M (2018) Measuring Issue Ownership using Word Embeddings. In: WASSA 2018 - 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Proceedings of the Workshop. Association for Computational Linguistics (ACL), pp 149–155
33. Sunkara M, Ronanki S, Dixit K, et al (2020) Robust Prediction of Punctuation and Truecasing for Medical ASR. In: Proceedings of the First Workshop on Natural Language Processing for Medical Conversations. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 53–62
34. Preum S, Shu S, Hotaki M, et al (2019) CognitiveEMS. ACM SIGBED Review 16:51–60. https://doi.org/10.1145/3357495.3357502
35. Lee SH, Park J, Yang K, et al (2022) Accuracy of Cloud-Based Speech Recognition Open Application Programming Interface for Medical Terms of Korean. J Korean Med Sci 37:. https://doi.org/10.3346/jkms.2022.37.e144
36. Lybarger KJ, Ostendorf M, Riskin E, et al (2018) Asynchronous Speech Recognition Affects Physician Editing of Notes. Appl Clin Inform 9:782–790. https://doi.org/10.1055/s-0038-1673417
37. Tran BD, Latif K, Reynolds TL, et al (2023) "Mm-hm," "Uh-uh": are non-lexical conversational sounds deal breakers for the ambient clinical documentation technology? Journal of the American Medical Informatics Association 30:703–711. https://doi.org/10.1093/jamia/ocad001
38. Renato A, Berinsky H, Daus M, et al (2019) Design and evaluation of an automatic speech recognition model for clinical notes in Spanish in a mobile online environment. In: Studies in Health Technology and Informatics. IOS Press, pp 1761–1762
39. Schultz BG, Tarigoppula VSA, Noffs G, et al (2021) Automatic speech recognition in neurodegenerative disease. Int J Speech Technol 24:771–779. https://doi.org/10.1007/s10772-021-09836-w
40. Paats A, Alumäe T, Meister E, Fridolin I (2018) Retrospective Analysis of Clinical Performance of an Estonian Speech Recognition System for Radiology: Effects of Different Acoustic and Language Models. J Digit Imaging 31:615–621. https://doi.org/10.1007/s10278-018-0085-8
41. Perez N, Álvarez A, Pozo A Del, et al (2022) ESAN: Automating medical scribing in Spanish ESAN: Automatización de la toma de notas clínicas
42. Roychowdhury P, Castillo-Bustamante M, Gandhi D, et al (2023) Evaluating the accuracy of speech to text applications for cochlear implant candidates during COVID-19. Cochlear Implants Int 24:1–5. https://doi.org/10.1080/14670100.2022.2120450
43. Salloum W, Finley G, Edwards E, et al (2017) Deep Learning for Punctuation Restoration in Medical Reports. In: BioNLP 2017. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 159–164
44. Shu S, Preum S, Pitchford HM, et al (2019) A Behavior Tree Cognitive Assistant System for Emergency Medical Services. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp 6188–6195
45. Selvaraj SP, Konam S (2021) Medication Regimen Extraction from Medical Conversations. In: Explainable AI in Healthcare and Medicine. Springer Link, pp 195–209
46. Cohn I, Laish I, Beryozkin G, et al (2019) Audio De-identification - a New Entity Recognition Task. In: Proceedings of the 2019 Conference of the North. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 197–204
47. Godfrey JJ, Holliman EC, McDaniel J (1992) SWITCHBOARD: telephone speech corpus for research and development. In: [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, pp 517–520 vol.1
48. Cieri C, Miller D, Walker K (2004) The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)
49. Hussain M, Choi D-J, Lee S (2020) Semantic based Clinical Notes Mining for Factual Information Extraction. In: 2020 International Conference on Information Networking (ICOIN). IEEE, pp 46–48

50. Zakraoui J, AlJa'am JM, Salah I (2022) Domain-Specific Text Generation for Arabic Text Summarization. In: 2022 International Conference on Computer and Applications (ICCA). IEEE, pp 1–4
51. Ansary MdS (2021) A Hybrid Approach for Extractive Summarization of Medical Documents. In: 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON). IEEE, pp 1–4
52. Yim W, Yetisgen M (2021) Towards Automating Medical Scribing : Clinic Visit Dialogue2Note Sentence Alignment and Snippet Summarization. In: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 10–20
53. Enarvi S, Amoia M, Teba MD-A, et al (2020) Generating Medical Reports from Patient-Doctor Conversations Using Sequence-to-Sequence Models. In: Proceedings of the First Workshop on Natural Language Processing for Medical Conversations. Association for Computational Linguistics, pp 22–30
54. Song Y, Tian Y, Wang N, Xia F (2020) Summarizing Medical Conversations via Identifying Important Utterances. In: Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Stroudsburg, PA, USA, pp 717–729
55. Abulkhair M, ALHarbi N, Fahad A, et al (2013) Intelligent integration of discharge summary: A formative model. In: Proceedings - International Conference on Intelligent Systems, Modelling and Simulation, ISMS. pp 99–104
56. Gulden C, Kirchner M, Schüttler C, et al (2019) Extractive summarization of clinical trial descriptions. Int J Med Inform 129:114–121. https://doi.org/10.1016/j.ijmedinf.2019.05.019
57. Reunamo A, Peltonen LM, Mustonen R, et al (2022) Text Classification Model Explainability for Keyword Extraction-Towards Keyword-Based Summarization of Nursing Care Episodes. In: Studies in Health Technology and Informatics. IOS Press BV, pp 632–636
58. Gao Y, Dligach D, Miller T, et al (2022) Summarizing Patients Problems from Hospital Progress Notes Using Pre-trained Sequence-to-Sequence Models. Proceedings of the 29th International Conference on Computational Linguistics 2979–2991
59. Ramanan S V, Radhakrishna K, Waghmare A, et al (2016) Dense Annotation of Free-Text Critical Care Discharge Summaries from an Indian Hospital and Associated Performance of a Clinical NLP Annotator. J Med Syst 40:. https://doi.org/10.1007/s10916-016-0541-2
60. Quiroz JC, Laranjo L, Kocaballi AB, et al (2020) Identifying relevant information in medical conversations to summarize a clinician-patient encounter. Health Informatics J 26:2906–2914. https://doi.org/10.1177/1460458220951719
61. Pivovarov R, Elhadad N (2015) Automated methods for the summarization of electronic health records. Journal of the American Medical Informatics Association 22:938–947. https://doi.org/10.1093/jamia/ocv032
62. Lee EK, Uppal K (2020) CERC: an interactive content extraction, recognition, and construction tool for clinical and biomedical text. BMC Med Inform Decis Mak 20:. https://doi.org/10.1186/s12911-020-01330-8
63. Dalal V, Malik L (2013) A Survey of Extractive and Abstractive Text Summarization Techniques. In: 2013 6th International Conference on Emerging Trends in Engineering and Technology. IEEE, pp 109–110
64. Zhang Z, Qin X, Qiu Y, Liu D (2021) Well-Behaved Transformer for Chinese Medical NER. In: Proceedings - 2021 3rd International Conference on Natural Language Processing, ICNLP 2021. Institute of Electrical and Electronics Engineers Inc., pp 162–167
65. Li M, Xiang L, Kang X, et al (2021) Medical Term and Status Generation from Chinese Clinical Dialogue with Multi-Granularity Transformer. IEEE/ACM Trans Audio Speech Lang Process 29:3362–3374. https://doi.org/10.1109/TASLP.2021.3122301
66. Kshatriya BSA, Sagheb E, Wi C-I, et al (2021) Identification of asthma control factor in clinical notes using a hybrid deep learning model. BMC Med Inform Decis Mak 21:272. https://doi.org/10.1186/s12911-021-01633-4
67. Raffel C, Shazeer N, Roberts A, et al (2019) Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 21:1–67
68. Kormilitzin A, Vaci N, Liu Q, Nevado-Holgado A (2021) Med7: A transferable clinical natural language processing model for electronic health records. Artif Intell Med 118:.

https://doi.org/10.1016/j.artmed.2021.102086
69. Ghiasvand O, Kate RJ (2015) Biomedical Named Entity Recognition with less Supervision. In: 2015 International Conference on Healthcare Informatics. IEEE, pp 495–495
70. Shah-Mohammadi F, Finkelstein J (2022) A Comparison of Automated Named Entity Recognition Tools Applied to Clinical Text. In: ICIIBMS 2022 - 7th International Conference on Intelligent Informatics and Biomedical Sciences. Institute of Electrical and Electronics Engineers Inc., pp 227–231
71. Tan H, Yang Z, Ning J, et al (2021) Chinese Medical Named Entity Recognition Based on Chinese Character Radical Features and Pre-trained Language Models. In: 2021 International Conference on Asian Language Processing, IALP 2021. Institute of Electrical and Electronics Engineers Inc., pp 121–124
72. Liu L, Perez-Concha O, Nguyen A, et al (2022) De-identifying Australian hospital discharge summaries: An end-to-end framework using ensemble of deep learning models. J Biomed Inform 135:. https://doi.org/10.1016/j.jbi.2022.104215
73. Xiao Y, Zhao Q, Li J, et al (2021) MLNer: Exploiting multi-source lexicon information fusion for named entity recognition in Chinese medical text. In: Proceedings - 2021 IEEE 45th Annual Computers, Software, and Applications Conference, COMPSAC 2021. Institute of Electrical and Electronics Engineers Inc., pp 1079–1084
74. Tang G (2022) Named Entity Recognition in Chinese Electronic Medical Records Based on ALBERT-IDCNN-CRF. In: 2022 IEEE 8th International Conference on Computer and Communications (ICCC). IEEE, pp 1753–1757
75. Yang D, Yang H, Wu B (2022) TsERL: Two-stage Enhancement of Radical and Lexicon for Chinese Medical Named Entity Recognition. In: Proceedings - 2022 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2022. Institute of Electrical and Electronics Engineers Inc., pp 2719–2726
76. Li D, Li J, Zhu Z, Mahmood T (2022) Chinese Medical Named Entity Recognition Based on Multi-word Segmentation and Multi-layer BILSTM. In: Proceedings - 2022 IEEE 46th Annual Computers, Software, and Applications Conference, COMPSAC 2022. Institute of Electrical and Electronics Engineers Inc., pp 1414–1419
77. Zong J, Han J (2022) Entity Recognition of Chinese Electronic Medical Record Based on Gated Graph Neural Network. In: 2022 14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA). IEEE, pp 1208–1213
78. Zafari H, Zulkernine F (2021) Chatsum: An Intelligent Medical Chat Summarization Tool. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, pp 1–2
79. Solarte-Pabon O, Blazquez-Herranz A, Torrente M, et al (2021) Extracting Cancer Treatments from Clinical Text written in Spanish: A Deep Learning Approach. In: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, pp 1–6
80. Meripo NV, Konam S (2021) Extracting Appointment Spans from Medical Conversations. In: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 41–46
81. Memarzadeh H, Ghadiri N, Samwald M, Lotfi Shahreza M (2023) Applying unsupervised keyphrase methods on concepts extracted from discharge sheets. Pattern Analysis and Applications 26:1715–1727. https://doi.org/10.1007/s10044-023-01198-0
82. Frei J, Kramer F (2022) GERNERMED: An open German medical NER model[Formula presented]. Software Impacts 11:. https://doi.org/10.1016/j.simpa.2021.100212
83. Frei J, Frei-Stuber L, Kramer F (2022) GERNERMED++: Transfer Learning in German Medical NLP. https://doi.org/10.48550/arXiv.2206.14504
84. Yalunin A, Nesterov A, Umerenkov D (2022) RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. https://doi.org/10.48550/arXiv.2204.03951
85. Nesterov A, Umerenkov D (2022) Distantly supervised end-to-end medical entity extraction from electronic health records with human-level quality. https://doi.org/10.48550/arXiv.2201.10463
86. Zhou H, Ruan D (2020) An Embedding-based Medical Note De-identification Approach with Minimal Annotation. In: 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE). IEEE, pp 263–268

87. Catelli R, Gargiulo F, Casola V, et al (2021) A Novel COVID-19 Data Set and an Effective Deep Learning Approach for the De-Identification of Italian Medical Records. IEEE Access 9:19097–19110. https://doi.org/10.1109/ACCESS.2021.3054479
88. Tchouka Y, Couchot J-F, Laiymani D (2023) An Easy-to-Use and Robust Approach for the Differentially Private De-Identification of Clinical Textual Documents. In: Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies. SCITEPRESS - Science and Technology Publications, pp 94–104
89. Liu K, Hu Q, Liu J, Xing C (2017) Named Entity Recognition in Chinese Electronic Medical Records Based on CRF. In: 2017 14th Web Information Systems and Applications Conference (WISA). IEEE, Liuzhou, China, pp 105–110
90. Weegar R, Perez A, Casillas A, Oronoz M (2018) Deep Medical Entity Recognition for Swedish and Spanish. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, Madrid, Spain, pp 1595–1601
91. Lafferty J, McCallum A, Pereira F (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the International Conference on Machine Learning (ICML'01)
92. Tian Y, Ma W, Xia F, Song Y (2019) ChiMed: A Chinese Medical Corpus for Question Answering. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 250–260
93. PUBLIC LAW 104-191 AUG. 21, 1996 HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT OF 1996 Public Law 104-191 104th Congress
94. Catelli R, Gargiulo F, Damiano E, et al (2021) Clinical de-identification using sub-document analysis and ELECTRA. In: 2021 IEEE International Conference on Digital Health (ICDH). IEEE, pp 266–275
95. El-Hayek C, Barzegar S, Faux N, et al (2023) An evaluation of existing text de-identification tools for use with patient progress notes from Australian general practice. Int J Med Inform 173:105021. https://doi.org/10.1016/j.ijmedinf.2023.105021
96. Liu Z, Huang Y, Yu X, et al (2023) DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4. https://doi.org/10.48550/arXiv.2303.11032
97. Loper E (2004) NLTK: Building a Pedagogical Toolkit in Python. PyCon DC 2004
98. Stubbs A, Uzuner Ö (2015) Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform 58:S20–S29. https://doi.org/10.1016/j.jbi.2015.07.020
99. Hu K, Pang R, Sainath TN, Strohman T (2021) Transformer Based Deliberation for Two-Pass Speech Recognition. In: 2021 IEEE Spoken Language Technology Workshop (SLT). IEEE, pp 68–74
100. Kojima A (2021) Large-Context Automatic Speech Recognition Based on RNN Transducer. In: Asia-Pacific Signal and Information Processing Association. Annual Summit and Conference Tokyo, Japan; Online, Asia-Pacific Signal and Information Processing Association, IEEE Signal Processing Society. Tokyo, Japan

Affiliations

Md. Faiyed Bin Karim[1]. Abu Ubaida Akash[1]. Sakib Zaman[1]. Shehan Irteza Pranto[1]. Marzia Zaman[2]. Farhana Sarkar[2]. Mohammad Nurul Huda[3]. Nabeel Mohammed[4]. Shariful Islam[5]. Iman Abdollah Dehzangi[6]. Khondaker Abdullah Al Mamun[1,3]

Md. Faiyed Bin Karim
faiyed-karim@iut-dhaka.edu

Abu Ubaida Akash
abu.ubaida.akash@usherbrooke.ca

Sakib Zaman
Sakibzaman169@gmail.com

Shehan Irteza Pranto
shehanirteza@gmail.com

Marzia Zaman
marzia.zama321@gmail.com

Farhana Sarker
farhana.sarker@ulab.edu.Bd

Mohammad Nurul Huda
mnh@cse.uiu.ac.bd

Nabeel Mohammed
Nabeel.mohammed@northsouth.edu

Shariful Islam
shariful.islam@deakin.edu.au

Iman Abdollah Dehzangiss
i.dehzangi@rutgers.edu

1 AIMS Lab, IRIIC, United International University, Dhaka, Bangladesh
2 CMED Health Ltd, Dhaka, Bangladesh
3 United International University, Dhaka Bangladesh
4 North South University, Dhaka Bangladesh
5 Deakin University, Geelong, Australia
6 Rutgers, The State University of New Jersey, United States