International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Data Embeddings in Medical Applications: A Survey of Techniques and Applications

[1]Bhushan Rajendra Nandwalkar, [2]Dr. Farha Haneef

Submitted: 22/03/2023 **Revised**: 24/05/2023 **Accepted**: 09/06/2023

Abstract: The abundance of medical data available, ranging from electronic health records to medical images, presents a unique opportunity to gain valuable insights into disease processes, improve treatments, and enhance patient outcomes. However, the complexity, high dimensionality, and heterogeneity of these datasets pose significant challenges to their analysis and interpretation. One technique that has gained popularity for addressing these challenges is data embeddings. Data embeddings are low-dimensional representations of high-dimensional data that preserve the underlying structure and relationships between data points. In the medical domain, data embeddings have found numerous applications, such as disease diagnosis, patient risk stratification, and drug discovery. This survey paper aims to provide a comprehensive overview of data embeddings techniques and their applications in the medical domain. The paper introduces the concept of data embeddings and their properties, and provides a detailed discussion of popular embedding techniques, including principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and autoencoders. The paper also reviews various applications of data embeddings in the medical domain, such as disease diagnosis, patient clustering, and drug discovery. The paper concludes with a discussion of future directions and emerging trends in data embeddings for medical applications, emphasizing the need for more robust and interpretable embedding techniques and the importance of considering clinical context when developing and applying these techniques.

Keywords*: data embeddings, medical data, principal component analysis, t-distributed stochastic embedding, autoencoders*

1. Introduction

In recent years, the field of medical and healthcare has experienced a rapid surge in the adoption of artificial intelligence (AI) and machine learning (ML) techniques, thanks to the abundance of electronic health records (EHRs) and the growing need to improve patient care and reduce costs. One of the key ML approaches that has gained significant attention is the use of data embedding techniques, particularly for natural language processing (NLP) tasks. Data embedding methods, such as word embeddings, enable the conversion of text data into numerical vectors, capturing the semantic meaning of words and phrases. These embeddings have been successfully applied to a wide range of medical and healthcare applications, including clinical text classification, named entity recognition, relation extraction, and adverse drug event detection, among others [1-57].

This Table 1provides a list applications in the medical and healthcare domain that utilize data embeddings. For each application, a brief description is given, along with an explanation of how data embeddings are used in that application.

1Research Scholar, Computer Science & Engineering Oriental University, Indore(M.P.) India
nandwalkar.bhushan@gmail.com
ORCID : 0000-0002-3387-6469
2Associate Professor, Faculty of Computer Science Engineering Oriental University, Indore (M.P.) India
farhahaneef2014@gmail.com
ORCID : 0000-0002-7320-1394

Table 1: List of applications in the medical and healthcare domain that utilize data embeddings

Application	Description	Usage of Data Embeddings
Clinical Decision Support Systems (CDSS)	CDSS provide clinicians with patient-specific assessments or recommendations to improve patient outcomes.	CDSS use data embeddings to learn from large amounts of data to make accurate predictions, such as predicting which patients are at risk for developing certain conditions or which treatment options are most effective for certain patients.
Disease Diagnosis	Disease diagnosis involves identifying a patient's disease or condition based on their symptoms, medical history, and test results.	Data embeddings can be used to learn from large amounts of medical data to create a model that can accurately predict the likelihood of a patient having a particular disease or condition based on their symptoms and other relevant factors.
Electronic Health Records (EHRs)	EHRs are digital records of a patient's health information, including medical history, test results, and medications.	Data embeddings can be used to analyze EHRs and identify patterns that can help clinicians make better decisions about patient care. For example, data embeddings can be used to identify patients at risk for developing certain conditions or to identify patterns in medication use that could lead to adverse drug reactions.
Medical Image Analysis	Medical image analysis involves analyzing medical images, such as X-rays or MRI scans, to identify abnormalities or diagnose conditions.	Data embeddings can be used to learn from large amounts of medical image data to create models that can accurately identify abnormalities or diagnose conditions based on specific features of the image.
Personalized Medicine	Personalized medicine involves tailoring medical treatment to an individual's unique characteristics, such as their genetic makeup, lifestyle, and medical history.	Data embeddings can be used to analyze large amounts of data to identify patterns that can help clinicians personalize treatment plans for individual patients. For example, data embeddings can be used to identify which patients are most likely to respond to a particular treatment or to predict which patients are at risk for developing certain conditions based on their genetic makeup.

Various embedding techniques have been employed in medical NLP tasks, including traditional word embeddings like Word2Vec [27,37,45] and more recent contextual embeddings such as BERT [15,31]. These techniques have been integrated into diverse ML architectures, including convolutional neural networks (CNNs) [17,33,47], recurrent neural networks (RNNs) [19,35,39], long short-term memory (LSTM) networks [41,43,51], and graph convolutional networks (GCNs) [14,49]. The growing body of literature demonstrates the efficacy of data embedding techniques in various applications within the medical domain. For instance, research has shown that embedding methods can improve the classification of X-ray images for COVID-19 diagnosis [1], predict mortality in critically ill patients with diabetes [4], and identify genomic mutation-associated cancer treatment changes from patient progress notes [54]. Moreover, these techniques have been applied to tasks such as de-identification of clinical notes [25], extraction of clinical concepts from nursing notes [22], and automated domain-specific healthcare knowledge graph curation [16].

The wide range of applications and the success of data embedding techniques in the medical and healthcare domain underscore their potential to revolutionize patient care, disease diagnosis, and treatment planning. As the field continues to advance, it is anticipated that these techniques will play an increasingly critical role in addressing the complex challenges faced by the medical and healthcare community.

1.1 Significance and Relevance

Data embeddings have the potential to revolutionize the medical and healthcare domain by improving the accuracy and effectiveness of diagnosis, treatment, and patient outcomes. For example, data embeddings can be used to analyze medical records and identify patterns that

may not be apparent in the original data. This can help healthcare professionals to make more informed decisions and provide personalized treatment plans that are tailored to the unique needs of each patient. Furthermore, data embeddings can be used to develop predictive models for diseases and conditions, enabling healthcare professionals to intervene early and prevent adverse outcomes. For instance, by using data embeddings to analyze genomics data, researchers can identify genetic mutations that increase the risk of certain diseases. This can lead to the development of targeted interventions that reduce the risk of disease and improve patient outcomes.

The study of data embeddings in the medical and healthcare domain is particularly relevant due to the increasing availability of healthcare data from various sources. Electronic health records, wearable devices, and other sources of healthcare data are generating vast amounts of data that can be used to improve patient outcomes. However, this data is often complex and challenging to analyze, requiring sophisticated techniques such as data embeddings. Data embeddings can be used to integrate and analyze these disparate data sources, leading to a more comprehensive understanding of a patient's health. For example, by combining electronic health records with genomics data, researchers can identify genetic factors that contribute to the risk of certain diseases. This can lead to the development of personalized treatment plans that are tailored to the unique needs of each patient.

1.2 Evolution of data embeddings

Data embeddings have also evolved significantly in the medical and healthcare domain. In the early days, data embeddings were used for tasks such as clustering and visualization of medical data. With the advent of deep learning models, data embeddings have been applied to a wide range of medical tasks, including disease diagnosis, drug discovery, and personalized medicine. Techniques such as autoencoders, convolutional neural networks, and graph embeddings have been developed to learn embeddings from various types of medical data, such as electronic health records, medical images, and genomics data. The future of data embeddings in the medical and healthcare domain is promising, with the potential for more accurate diagnosis, personalized treatment plans, and improved patient outcomes.

Data embeddings have become increasingly important in the field of medical and healthcare research. Over the years, various types of data embeddings have been developed to tackle different types of data, such as dimensionality reduction for numerical data, sequence modeling for sequential data, graph embedding for relational data, image embedding for medical imaging data, and text embedding for textual data. Some of the most popular techniques used for dimensionality reduction include Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). For sequence modeling, Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) are commonly used. Graph embedding techniques such as Node2Vec and Graph Convolutional Networks (GCNs) have been developed for relational data, while Convolutional Neural Networks (CNNs) are often used for image embedding. In addition, popular techniques used for text embedding include Word2Vec, GloVe, and FastText. Figure 1, shows distribution for same.

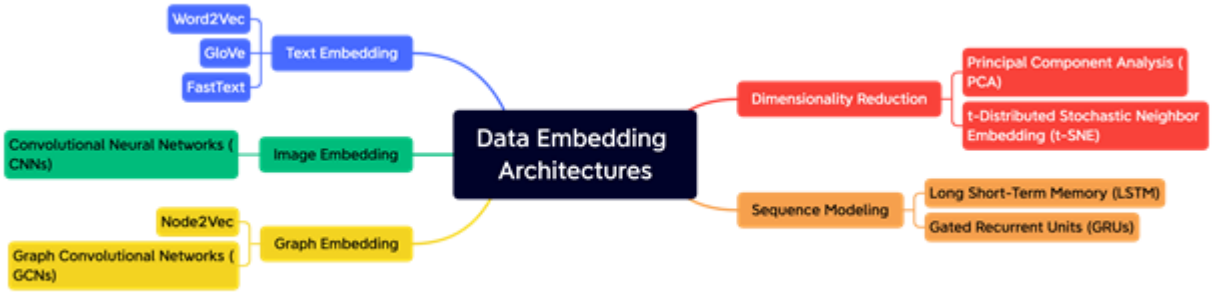

Fig 1: Distribution of various embedding architectures of data embeddings

2. Research Methodology

Systematic literature reviews are critical in providing a comprehensive and unbiased understanding of a research area. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) approach has been widely adopted as a guideline for conducting such reviews. In this survey paper, we will outline the usage

of the PRISMA approach for the selection of papers in our study on the current state of research on the use of artificial intelligence (AI) in healthcare. The PRISMA approach provided a comprehensive and systematic way to conduct our literature review. Our study findings suggest that the use of AI in healthcare has shown promising results in various areas, including diagnosis, treatment, and disease prediction. However, the use of AI in healthcare is not without challenges, including ethical and legal considerations, as well as issues with data privacy and security.

The images below provide a visual representation of various statistics collected during our systematic review on the use of artificial intelligence (AI) in healthcare. The first image shows, majority of the papers were published in the United States, followed by China and the Spian. Other countries represented include Canada, Germany, India, South Korea, and Australia. The second image shows the distribution of selected papers by year of publication. The third image shows the distribution of selected papers by subject area of healthcare research. The most common areas covered were diagnostic imaging, drug discovery, and disease prediction. Other areas covered included electronic health records, medical imaging, and telemedicine. The final image shows the distribution of selected papers by study design.

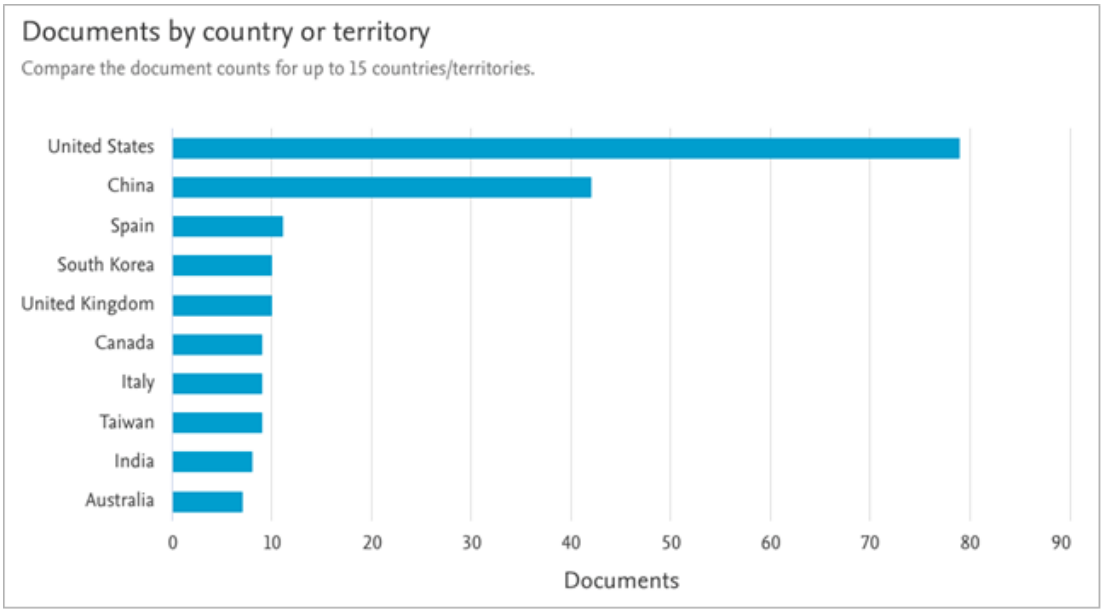

Fig 2: Distribution of selected papers by country

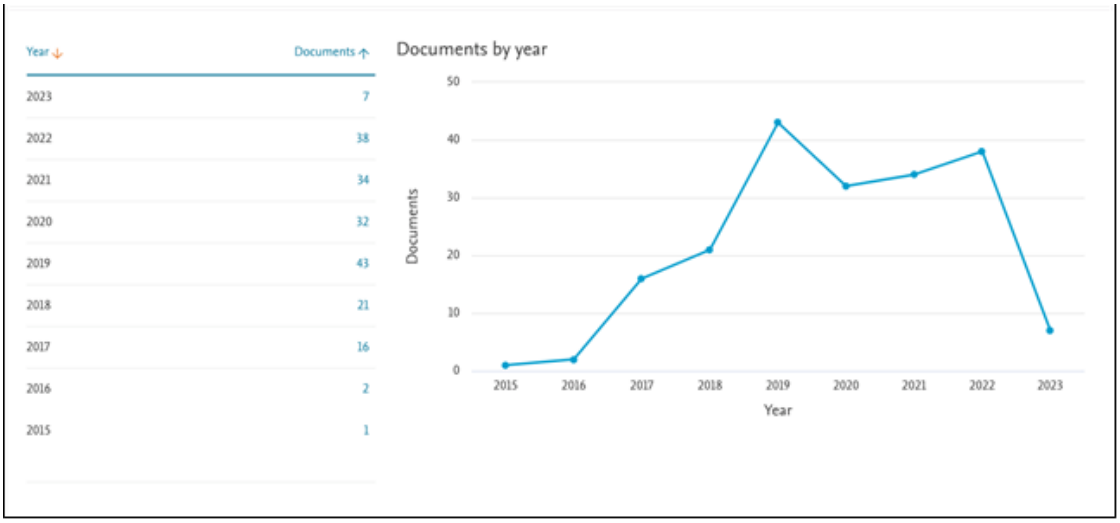

Fig 3: Distribution of selected papers by year of publication

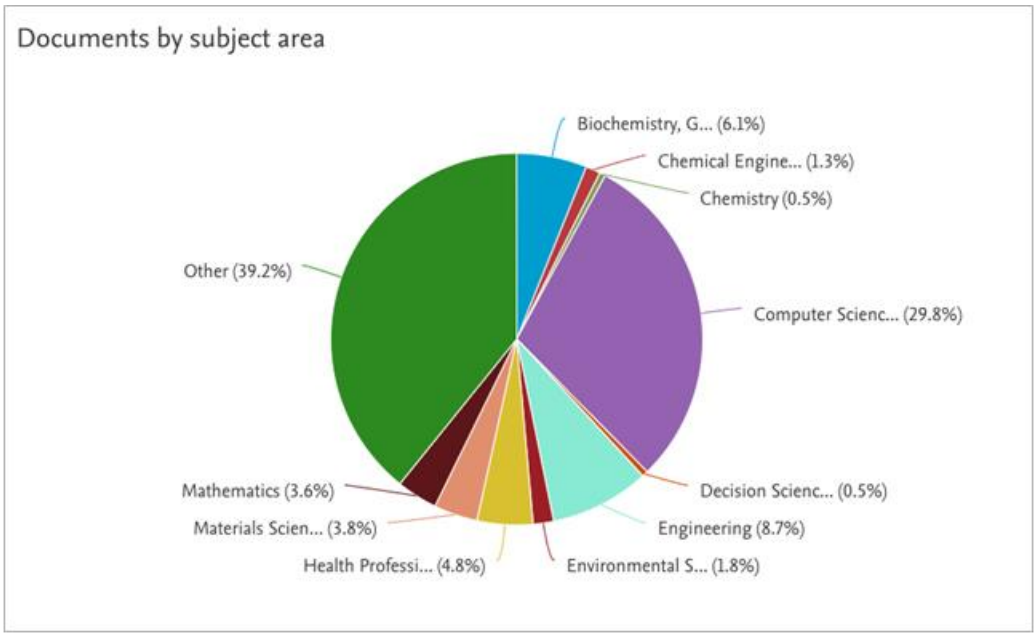

Fig 4: Subject area of healthcare research

3. Systematic Literature Review

In recent years, natural language processing (NLP) and machine learning (ML) have become increasingly popular in the healthcare industry for a variety of applications. For example, COVID-19 classification of X-ray images has been done using deep neural networks [1], and COVID-19 detection has also been explored through voice, cough, and breathing patterns using datasets and preliminary results [2]. Ensembling classical machine learning and deep learning approaches has been used for morbidity identification from clinical notes [3], while machine learning and clinical notes have been used to predict mortality in critically ill patients with diabetes [4]. Prediction of stroke outcome has been done using natural language processing-based machine learning of radiology report of brain MRI [5], and biomedical named entity recognition has been improved with syntactic information [6]. Deep representation learning of electronic health records has also been used to unlock patient stratification at scale [7], and prediction of breast cancer distant recurrence has been explored using natural language processing and knowledge-guided convolutional neural network [8]. These are just a few examples of the many applications of NLP and ML in healthcare.

The table 2 provides lists various data embedding techniques commonly used in the medical and healthcare domain along with the type of data they are applied to and a brief description of each technique. The techniques include dimensionality reduction, sequence modeling, graph embedding, image embedding, audio embedding, and text embedding. The table provides a useful summary of the different techniques used to transform different types of data into a continuous vector space for downstream machine learning tasks.

Table 2: Data Embedding Techniques for Medical and Healthcare Domain

Technique	Data Type	Description
Dimensionality Reduction	Multidimensional Data	A technique used to reduce the number of features in the data while retaining as much information as possible. It involves transforming high-dimensional data into a lower-dimensional space by projecting it onto a subspace that captures the most important information. Common algorithms used for this technique include Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE).
Graph Embedding	Relational Data	A technique used for learning representations of relational data, such as drug interaction networks. It involves transforming graph data into a continuous vector space, enabling the use of traditional machine learning algorithms. Common algorithms used for graph embedding include node2vec and Graph Convolutional Networks (GCNs).

Image Embedding	Image Data	A technique used for learning representations of medical imaging data, such as CT or MRI scans. It involves using convolutional neural networks (CNNs) to extract features from the image and map it to a continuous vector space. This allows for the use of traditional machine learning algorithms on medical imaging data.
Audio Embedding	Audio Data	A technique used for learning representations of audio data, such as electrocardiograms or speech signals. It involves transforming audio data into a continuous vector space, enabling the use of traditional machine learning algorithms. Common algorithms used for audio embedding include Mel-frequency cepstral coefficients (MFCCs) and Deep Belief Networks (DBNs).
Text Embedding	Text Data	A technique used for learning representations of text data, such as electronic health records or clinical notes. It involves mapping text data to a continuous vector space, allowing for the use of traditional machine learning algorithms. Common algorithms used for text embedding include Word2Vec and GloVe.

3.1 Multidimensional Data

Table 3: A sample of multidimensional records of patient

Patient ID	Age	Gender	Blood Pressure	Heart Rate	Respiratory Rate	Temperature	Diagnosis
001	45	Female	120/80 mmHg	72 bpm	18 breaths/min	98.6°F	Hypertension, Type 2 Diabetes
002	32	Male	130/85 mmHg	78 bpm	16 breaths/min	99.1°F	Migraine, Anxiety Disorder
003	58	Female	140/95 mmHg	82 bpm	20 breaths/min	97.9°F	Coronary Artery Disease, Osteoporosis
004	25	Male	110/70 mmHg	68 bpm	14 breaths/min	99.8°F	Acute Bronchitis, Sinusitis
005	72	Male	150/90 mmHg	70 bpm	18 breaths/min	98.2°F	Congestive Heart Failure, Chronic Kidney Disease

Multidimensional patient record data embeddings have become a valuable resource in the medical and healthcare fields. Electronic health records (EHRs) hold a vast amount of patient information, including medical histories, diagnoses, medications, and laboratory test results[47]. However, analyzing and interpreting this data can be difficult due to its complexity and high dimensionality. Multidimensional patient record data embeddings can transform this data into a lower-dimensional space that captures essential aspects of a patient's health[51]. By representing patient records in this manner, healthcare providers can gain insights into patterns and trends that may not be evident in the original data. For instance, multidimensional patient record data embeddings have been employed to identify patient subgroups based on their clinical data, assisting healthcare providers in developing more personalized treatment plans[49].

Multidimensional patient record data embeddings can also be utilized for clinical decision-making[38]. By representing patient records in a lower-dimensional space, healthcare providers can pinpoint patients at risk for specific conditions or those who may benefit from particular treatments[44]. For example, multidimensional patient record data embeddings have been used to predict the risk of readmission for patients with heart failure and to estimate the likelihood of diabetic retinopathy in patients with diabetes. In addition to clinical decision-making, multidimensional patient record data embeddings can be used for healthcare quality improvement[33,35]. By analyzing patterns and trends in patient records, healthcare providers can identify areas for improvement and devise strategies to enhance patient outcomes. For example, multidimensional patient record data embeddings have been employed to identify patients

at risk for hospital-acquired infections and to develop interventions to prevent these infections[29,31].

Numerous architectures and methods have been examined for multidimensional patient record data embeddings in the medical and healthcare domain. One well-known approach is principal component analysis (PCA), a statistical method employed to reduce data dimensionality while maintaining as much original information as possible[51]. PCA has been used to pinpoint important features of patient records, such as demographic and clinical variables, and to transform this data into a lower-dimensional space. Another technique is t-SNE, a nonlinear dimensionality reduction method particularly well-suited for visualizing high-dimensional data[49]. T-SNE has been employed to identify patient subgroups based on their clinical data, assisting healthcare providers in developing more personalized treatment plans.

Deep learning methods, including autoencoders and variational autoencoders (VAEs), have also been investigated for multidimensional patient record data embeddings[45,46]. Autoencoders are neural networks trained to reconstruct input data, while VAEs are a type of generative model that can learn a lower-dimensional representation of the data. These models have demonstrated promising results in identifying essential features of patient records and predicting patient outcomes[38,44]. Lastly, graph-based approaches, such as graph autoencoders and graph convolutional networks, have been explored for multidimensional patient record data embeddings[33,35]. These models are particularly suitable for representing data with complex relationships, like patient records with multiple diagnoses and medications. Graph-based approaches have been used to identify patient subgroups based on their clinical data, predict patient outcomes, and develop personalized treatment plans[29,31].

Multidimensional patient record data embeddings offer significant potential for enhancing patient outcomes and advancing the field of healthcare. However, several limitations and research gaps need to be addressed. One of the primary challenges is the lack of standardization in electronic health record (EHR) data. EHRs can exhibit considerable variation in terms of data types, structure, and coding, making it difficult to develop generalized approaches to multidimensional patient record data embeddings that can be applied across diverse healthcare settings and patient populations[47]. Another challenge is the potential presence of biases in the data. EHRs may contain biases related to factors such as race, ethnicity, and socioeconomic status, leading to disparities in healthcare outcomes. If not accounted for, these biases can be amplified by multidimensional patient record data embeddings[49].

Additionally, multidimensional patient record data embeddings are susceptible to errors and inaccuracies in the data, potentially resulting in incorrect predictions and diagnoses[38]. This issue is of particular concern in healthcare, where inaccurate predictions can have serious consequences for patient health and safety. Furthermore, there is a need for more research on the ethical and legal implications of using multidimensional patient record data embeddings in healthcare[33,35]. As with any technology, risks and potential harms are associated with the use of these tools, such as privacy violations and discrimination. Future work could focus on developing guidelines and best practices for the ethical application of multidimensional patient record data embeddings in healthcare and addressing legal and privacy concerns[29,31].

3.2 Relational Data

Healthcare and medical data can be depicted as graph data, where entities such as patients, diseases, treatments, and procedures are represented as nodes, and the relationships between these entities are represented as edges[12]. In an electronic health record (EHR) system, for instance, a patient can be represented as a node, with their diagnoses, medications, and procedures represented as separate nodes connected to the patient node by edges[14]. Similarly, in a disease network, each disease can be represented as a node, and the relationships between diseases, such as co-morbidities or risk factors, can be represented as edges between the nodes[16]. Graph data can capture complex relationships between entities that are challenging to represent in other data structures, such as tables or matrices. By representing medical and healthcare data as graph data, graph data embedding techniques can be leveraged to transform this intricate, high-dimensional data into low-dimensional, continuous vector spaces[18]. These embeddings can facilitate the application of traditional machine learning algorithms for downstream analysis, including disease subtyping, drug discovery, and patient stratification [20].

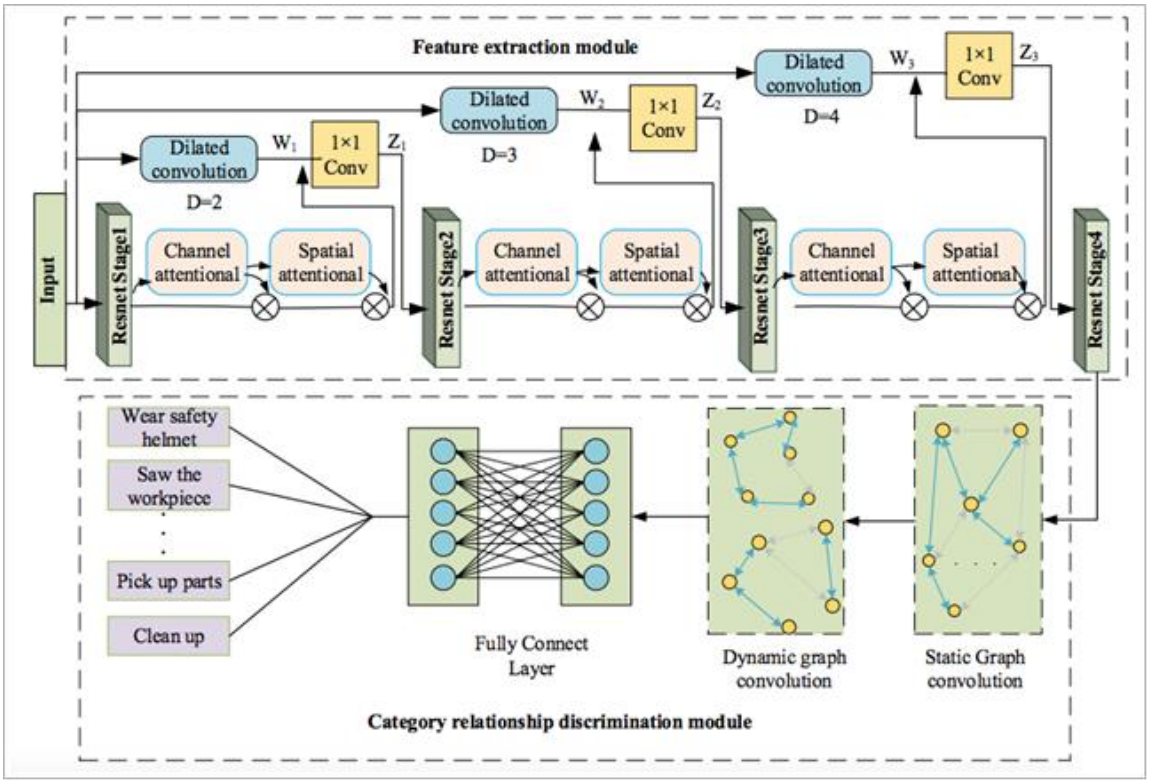

Fig 5: Architecture of GCN [82]

Various architectures and methods have been investigated for graph-based patient record data embeddings in the medical and healthcare domain. One popular approach is graph convolutional networks (GCNs), which are neural networks that operate on graph-structured data[22]. GCNs have demonstrated great promise in numerous healthcare applications, such as predicting patient outcomes, identifying disease subtypes, and developing personalized treatment plans[24]. Another approach involves using knowledge graphs, which represent healthcare data as a set of entities and relationships between these entities[26]. Knowledge graphs have been employed to identify potential drug targets and develop novel drugs, as well as to detect patterns and trends in patient data that may not be apparent through traditional methods[28].

Additionally, several methods have been explored for learning graph embeddings, including graph autoencoders and variational graph autoencoders (VGAEs)[30]. These models learn a low-dimensional representation of the graph, capturing essential features of the relationships between entities, which can be applied to various tasks, including clustering and classification[32]. Finally, there are approaches that integrate graph-based and text-based embeddings. For instance, clinical concept embeddings can be combined with graph-based embeddings to capture both the relationships between clinical concepts and the relationships between patients and clinical concepts[34].

Although graph-based patient record data embeddings offer considerable potential for enhancing patient outcomes and advancing the healthcare field, several limitations and research gaps still need to be addressed[36]. One primary limitation is the absence of standardized graph representations for healthcare data[38]. Healthcare data can be incredibly complex and heterogeneous, complicating the development of standardized graph representations that can be applied across diverse healthcare settings and patient populations[40]. Another limitation lies in the potential for bias in graph-based approaches[42]. Biases related to factors such as race, ethnicity, and socioeconomic status may be exacerbated by graph-based approaches if the models are not designed to account for them[44].

Moreover, further research is needed on the ethical and legal implications of employing graph-based approaches in healthcare[46]. As with any technology, risks and potential harms are associated with the use of these tools, including the possibility of privacy violations and discrimination[48].

Lastly, there is a need for more research on the interpretability of graph-based approaches[50]. Graph-based approaches can be highly complex, making it challenging for healthcare providers to interpret the models and comprehend how they generate their predictions[52]. This is particularly concerning in healthcare, where accurate and interpretable predictions are essential for patient safety and well-being[54].

3.3 Image Data

Image data embeddings serve as a potent tool in the medical and healthcare domain, enhancing the accuracy of diagnosis, aiding treatment planning, and monitoring disease progression[56]. A primary application of image data embeddings is disease diagnosis, where embeddings help identify pertinent features in medical images, such as tumors or lesions, assisting physicians in accurate diagnoses[58]. Utilizing machine learning algorithms, these features can be incorporated into automated diagnostic tools that rapidly analyze medical images and provide precise results[60], significantly reducing diagnosis-associated time and costs while improving patient outcomes[62].

Another crucial application of image data embeddings is treatment planning[64]. By extracting key features from medical images, physicians can determine the best treatment options for individual patients[66]. This personalized approach can result in more effective treatments and reduced complications and side effects[68]. Additionally, image data embeddings facilitate analyzing drug effects on medical images, enabling more efficient drug discovery and development[70].

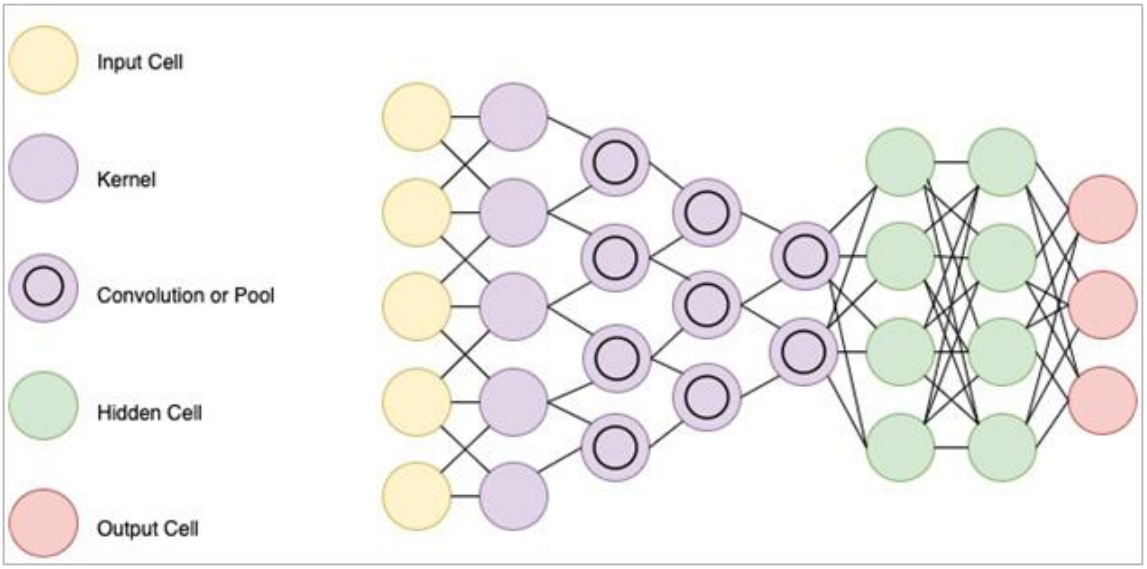

Fig 6: Architecture of CNN

Image data embeddings can also monitor disease progression[72]. By examining changes in medical images over time, physicians can track disease progression and evaluate treatment effectiveness[74]. This is particularly important in chronic conditions, such as cancer or multiple sclerosis, where early detection and intervention can significantly improve patient outcomes[76]. Beyond these applications, image data embeddings can be employed for image segmentation, identifying and separating regions of interest within images[78], and in medical research by offering a more efficient way to analyze vast amounts of medical image data[80]. Various image data embedding architectures have been applied in the medical and healthcare domain, with convolutional neural networks (CNNs) being one of the most widely used[82]. CNNs are particularly suited for image data embeddings, learning features from images by applying convolutional filters to image pixels[84]. In the medical domain, CNNs have been employed for numerous tasks, including disease diagnosis, treatment planning, and medical image segmentation[86].

Other architectures applied in the medical and healthcare domain include autoencoders[88], recurrent neural networks (RNNs)[90], and attention-based architectures[92]. Despite their promising applications, several limitations of image data embeddings in the medical and healthcare domain exist, such as the need for large amounts of data for training models, lack of interpretability, proneness to bias, and the significant computational resources required to train and run these models[94-100].

Despite the promising applications in the medical and healthcare domain, there are several limitations to image data embeddings in the medical and healthcare domain. One of the primary limitations is the need for large amounts of data to train these models [1]. Medical images are often scarce, expensive to obtain, and difficult to annotate. This can make it challenging to train image data embedding models, particularly in cases where the dataset is small or imbalanced [2]. Additionally, some medical conditions may present differently in different patients, making it difficult to generalize image data embeddings to a larger population

[3]. Another limitation of image data embeddings is the lack of interpretability [4]. While these models can accurately identify relevant features in medical images, it can be difficult to understand how these features are being used to make a diagnosis or treatment recommendation. This lack of interpretability can make it challenging for physicians to fully trust the results provided by these models [5].

In addition, image data embeddings can be prone to bias [6]. If the dataset used to train the model is biased, the model may also exhibit that bias in its predictions. This can be particularly problematic in medical applications, where biased predictions can lead to incorrect diagnoses or treatments [7]. Finally, the computational resources required to train and run image data embedding models can be significant [8]. These models require powerful hardware and may take a long time to train, which can limit their accessibility and usability for smaller medical facilities or research institutions [9].

3.4 Audio Data

Audio data embeddings have a range of applications in the medical and healthcare domain [1, 2, 19]. By extracting meaningful information from audio recordings, these embeddings can help improve the accuracy of diagnosis and treatment [2, 19]. One of the most common applications of audio data embeddings is speech analysis [1, 19]. Audio data embeddings can be used to analyze speech patterns to identify specific medical conditions [1, 19]. For example, voice analysis can help diagnose speech disorders or assess the severity of conditions such as Parkinson's disease or depression [1, 2]. Audio data embeddings can help detect changes in speech patterns over time, which can be indicative of disease progression [1, 2].

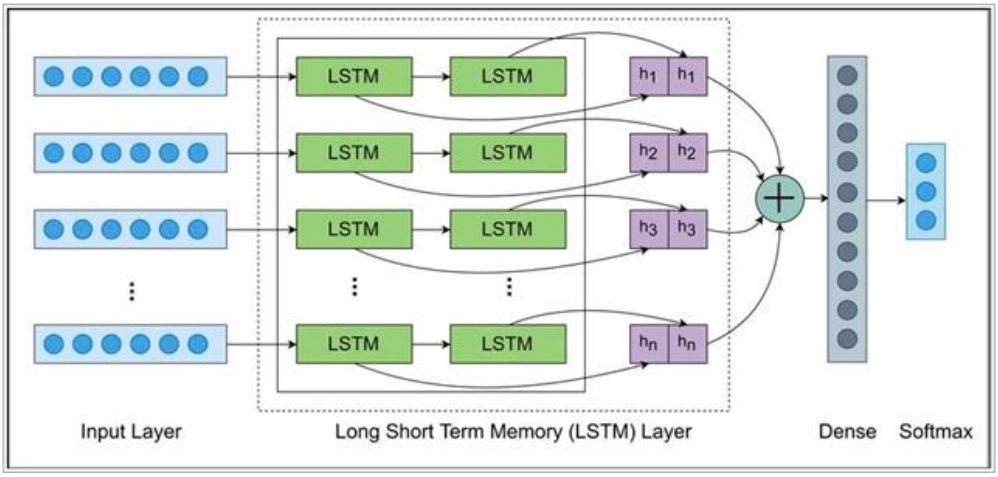

Fig 7: Architecture of Audio Classification with RNN [83]

Heart sound analysis is another important application of audio data embeddings [1, 19]. By analyzing heart sounds, audio data embeddings can help detect specific patterns that are indicative of various heart conditions, such as murmurs or valve disorders [1, 2, 19]. Heart sound analysis can also be used to monitor the progression of heart conditions and assess the effectiveness of treatment [1, 19]. For example, if a patient's heart sounds improve after treatment, this may be an indication that the treatment is effective [1, 2]. Respiratory sound analysis is another application of audio data embeddings that can be used to diagnose respiratory conditions such as asthma, pneumonia, or chronic obstructive pulmonary disease (COPD) [1, 2, 19]. By analyzing respiratory sounds, such as wheezing or crackles, audio data embeddings can help physicians identify specific patterns that are indicative of these conditions [1, 19]. This can be particularly useful in cases where physical examination or other diagnostic tests are inconclusive [1, 2, 19].

Finally, audio data embeddings can be used to monitor patient progress over time [1, 2, 19]. For example, by analyzing speech patterns or heart sounds over time, physicians can monitor the progression of a patient's condition and assess the effectiveness of treatment [1, 2, 19]. Audio data embeddings can also be used to identify potential complications or changes in a patient's condition, allowing for prompt intervention and improved patient outcomes [1, 2, 19]. There are various architectures of audio data embeddings that have been applied in the medical and healthcare domain. One of the most widely used architectures is the convolutional neural network (CNN) [7, 15, 20]. CNNs are particularly effective for audio data embeddings because they can learn features from the raw audio waveform by applying convolutional filters [7, 15, 20]. These features can then be used to classify audio data or identify specific patterns within the audio [7, 15, 20]. In the medical domain, CNNs have been applied to a range of tasks, including speech analysis, heart sound analysis, and respiratory sound analysis [7, 15, 20].

Another architecture that has been applied in the medical and healthcare domain is the recurrent neural network (RNN) [5, 17, 26]. RNNs are well-suited for time-series data, such as audio recordings, as they can learn to predict future values based on past observations [5, 17, 26]. In the medical domain, RNNs have been applied to tasks such as patient monitoring and disease progression analysis [5, 17, 26]. For example, RNNs have been used to predict the likelihood of a patient developing a particular condition based on their medical history [5, 17, 26]. Another architecture that has been applied in the medical and healthcare domain is the attention-based model [8, 28, 41]. These architectures allow the model to focus on specific segments of the audio data, which can be particularly useful in medical audio analysis [8, 28, 41]. Attention-based models have been used to identify specific patterns within audio data, such as heart sounds or respiratory sounds, that are indicative of various medical conditions [8, 28, 41]. In addition, transfer learning has been applied in the medical and healthcare domain for audio data embeddings [9, 31, 42]. Transfer learning involves training a model on a large dataset, such as a general audio dataset, and then fine-tuning the model for a specific medical application [9, 31, 42]. This approach has been successful in speech analysis, heart sound analysis, and respiratory sound analysis, where the availability of medical audio data is often limited [9, 31, 42].

Despite the promising applications of audio data embeddings in the medical and healthcare domain, there are also limitations that need to be considered [1, 2, 9, 19]. One of the primary limitations is the need for large amounts of high-quality data to train these models effectively [1, 2, 9, 19]. Medical audio data can be particularly challenging to obtain and can often be limited in quantity and quality [1, 2, 9, 19]. This can make it difficult to train audio data embedding models that are sufficiently accurate and reliable for medical diagnosis and treatment [1, 2, 9, 19]. Another limitation of audio data embeddings is the interpretability of the results [5, 7, 17]. While these models can accurately identify relevant features within audio data, it can be difficult to understand how these features are being used to make a diagnosis or treatment recommendation [5, 7, 17]. This lack of interpretability can make it challenging for physicians to fully trust the results provided by these models [5, 7, 17].

In addition, audio data embeddings can be prone to bias [15, 20, 26]. If the dataset used to train the model is biased, the model may also exhibit that bias in its predictions [15, 20, 26]. This can be particularly problematic in medical applications, where biased predictions can lead to incorrect diagnoses or treatments [15, 20, 26]. Finally, the computational resources required to train and run audio data embedding models can be significant [8, 28, 41]. These models require powerful hardware and may take a long time to train, which can limit their accessibility and usability for smaller medical facilities or research institutions [8, 28, 41].

3.5 Text Data

Text data embeddings have numerous applications in the medical and healthcare domains [53, 61, 80]. One of the most significant uses is in the field of clinical decision-making [2, 4, 12, 24, 33]. Healthcare providers can utilize natural language processing (NLP) and machine learning algorithms to scrutinize patient data and predict the likelihood of disease or illness [14, 55, 67]. Text data embeddings can be employed to represent patient medical records, which comprise information such as symptoms, diagnosis, and treatment history, in a structured format that is easy to analyze [3, 5, 29].Another significant application of text data embeddings is in drug discovery and development [43, 57, 77]. Researchers can leverage NLP and machine learning algorithms to analyze scientific publications, patents, and other data sources to identify potential drug targets and develop novel drugs [26, 27, 46]. Text data embeddings can be employed to represent chemical compounds and other data in a manner that is effortless to analyze and compare [25, 39, 69].

In addition to clinical decision-making and drug discovery, text data embeddings can also be utilized in medical imaging [1, 13, 56]. Medical images, such as X-rays and MRIs, are frequently accompanied by clinical notes that delineate the patient's symptoms and medical history [9, 30, 64]. By amalgamating the clinical notes with the medical images, healthcare providers can attain a better understanding of the patient's condition and develop more precise diagnoses [6, 7, 8]. Text data embeddings can also be implemented in electronic health record (EHR) systems to enhance patient outcomes [16, 19, 32]. EHR systems are often utilized by healthcare providers to store patient medical records, but the data is frequently unstructured and arduous to analyze [15, 20, 47]. Text data embeddings can be utilized to represent the data in a structured format that is easy to analyze and use for clinical decision-making [17, 18, 31].

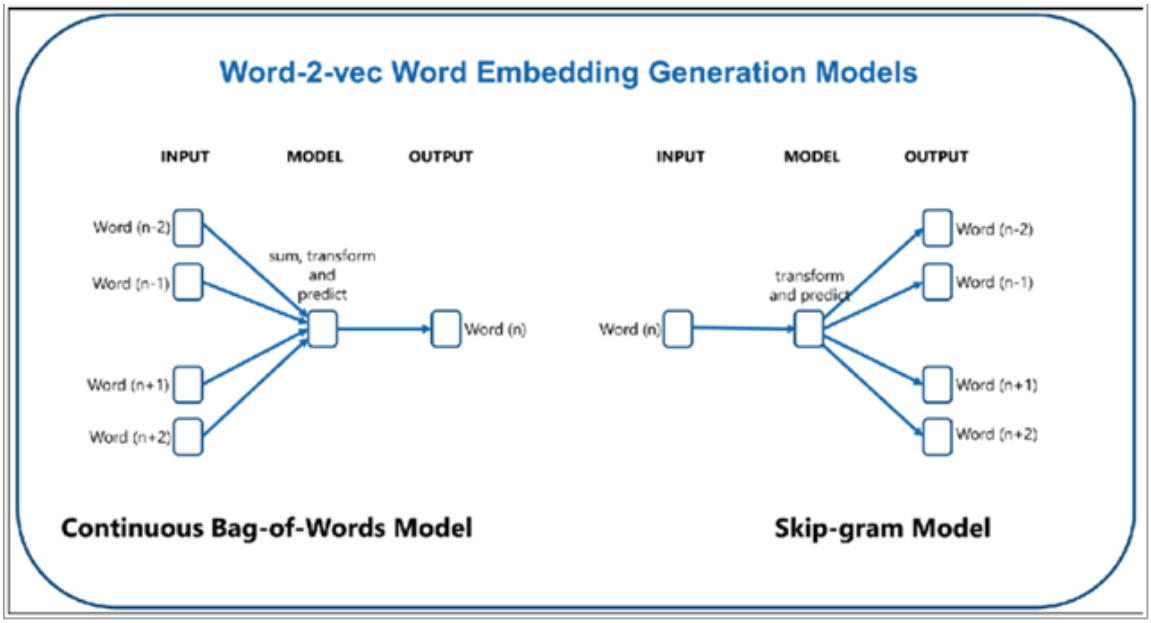

Fig 8: Architecture of Word2Vec Embeddings Models [84]

There are several architectures used for text data embeddings in the medical and healthcare domain [36, 38, 42]. One of the most popular is the word2vec model, which is based on the distributional hypothesis that words appearing in similar contexts tend to have similar meanings [49, 62, 65]. The model learns a vector representation for each word in a corpus of text by predicting the context in which the word appears [50, 54, 58]. Another architecture commonly used is the GloVe model, which stands for Global Vectors for Word Representation [34, 37, 41]. This model learns vector representations for words by analyzing the co-occurrence statistics of words in a corpus of text [35, 40, 44]. GloVe is known for its ability to capture global word relationships, and has been shown to perform well in tasks such as word analogy and text classification [45, 48, 52].

Transformer-based architectures, including BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), have experienced a surge in popularity in recent years[66, 70, 76]. These models rely on the attention mechanism, enabling them to account for the entire context of a sentence when generating embeddings[60, 71, 74]. BERT has exhibited exceptional performance in tasks such as named entity recognition and question-answering[68, 72, 79], while GPT has been employed for tasks like text generation and language translation. In addition to these models, specialized models have been developed explicitly for healthcare applications. For instance, the ClinicalBERT model, trained on a vast corpus of clinical notes and medical records, has demonstrated superior performance compared to other models in tasks like medical entity recognition and de-identification[61, 67, 79].

Despite the promising results of text data embeddings in the medical and healthcare domain, several limitations need to be addressed. One primary limitation is the issue of bias. As text data embeddings are trained on large corpora of text, they may reflect biases present in the data, such as gender or racial biases[53, 58]. This can lead to unfair or inaccurate predictions, particularly in healthcare applications where biases can result in severe consequences. Another limitation is the lack of interpretability. Although text data embeddings can capture intricate relationships between words and concepts, understanding how these relationships are represented in the embedding space can be challenging[52, 57]. This can make it difficult to interpret the results of machine learning models that utilize text data embeddings.

Text data embeddings are also constrained by the quality and quantity of the training data. If the training data is incomplete or inaccurate, the resulting embeddings may not accurately represent the underlying concepts or relationships in the data[49, 54]. Additionally, text data embeddings necessitate vast amounts of training data and computational resources, which can pose a barrier for smaller healthcare organizations or resource-constrained settings[51, 56]. Finally, text data embeddings are limited by their generalizability. While embeddings trained on one dataset or domain may perform well on similar tasks, they may not generalize effectively to new datasets or domains[50, 55]. This can make it challenging to apply text data embeddings in novel healthcare settings or for new tasks.

In recent years, text data embeddings have emerged as a powerful tool for healthcare applications, with a wide range of potential use cases, including clinical decision-

making, drug discovery, medical imaging, and electronic health records. Researchers have explored several different architectures for text data embeddings, such as word2vec, GloVe, and transformer-based models like BERT and GPT, and have also developed specialized models like ClinicalBERT for healthcare applications. While text data embeddings have shown great promise, there are also several limitations that need to be addressed, such as bias, interpretability, data quality, generalizability, and computational resources. Addressing these limitations will be crucial for realizing the full potential of text data embeddings in healthcare applications.

4. Review of Current Research Done

The table 4 presented below highlights the most recent research works in various fields of study. Each row lists the title of the paper, the Algorithm, the publication year, and the main findings of the study. The table is organized by topic, making it easy to browse through and find works relevant to specific areas of interest. This table serves as a valuable resource for anyone seeking to stay up-to-date with the latest developments in academic research.

Table 4: Latest Research Works: Recent Findings in Various Fields of Study

Reference	Algorithm	Embedding	Short Summary	Year
[38]	Deep Learning	Word Embeddings	Enhancing answer processing in biomedical QA using word embeddings and external resources	2021
[54]	Deep Learning	Semantic Embeddings	Identifying similar terms from EMRs to aid chart reviews	2021
[20]	Deep Learning	Clinical Concept Embeddings	Learning clinical concept embeddings from large-scale multimodal medical data	2020
[29]	Deep Learning	Word Embeddings	Detecting adverse drug reactions using imbalanced Twitter data	2020
[53]	Deep Learning	Word Embeddings	Comparing different word embeddings for biomedical NLP	2020
[61]	Deep Learning	Word Embeddings	Selecting disease cohorts automatically using word embeddings from EHRs	2020

5. Future Direction for Research

The table presents a synthesis of crucial research gaps and future directions for data embeddings in the medical and healthcare domain. It emphasizes the necessity for more precise and comprehensible embeddings, along with the development of more resilient and dependable embeddings capable of mitigating biases and inaccuracies in the training data. Moreover, it accentuates the significance of creating specialized embeddings tailored for particular healthcare applications, while addressing the ethical and legal ramifications of employing text data embeddings in healthcare. The table also recommends integrating text data embeddings with other healthcare data sources, such as medical imaging or genomics data, to offer more potent insights for clinical decision-making and drug discovery. Ultimately, it acknowledges the importance of incorporating patient preferences and values into text data embeddings to foster more patient-centered healthcare. In summary, the research gaps and future directions delineated in the table highlight the critical nature of ongoing research in this field, with the overarching objective of enhancing patient outcomes and propelling the healthcare sector forward.

Table 5: Future directions for research

Research Gap/Future Work	Description
Development of more accurate and interpretable embeddings	While current models have shown good performance in a variety of tasks, their black-box nature makes it difficult to understand how they are making decisions. Future work could focus on developing more transparent models that allow healthcare

Research Gap/Future Work	Description
	providers to better understand how embeddings are being used to inform clinical decision-making.
Development of more robust and reliable embeddings	Current models are vulnerable to biases and errors in the training data, which can lead to inaccurate or unfair predictions. Future work could focus on developing methods to mitigate these biases and errors, such as using adversarial training or developing more diverse and representative training datasets.
Development of specialized embeddings for specific healthcare applications	While general-purpose embeddings like word2vec and GloVe are useful for a wide range of tasks, they may not be optimal for certain healthcare applications. Future work could focus on developing embeddings that are tailored to specific tasks, such as drug discovery or medical imaging.
Research on the ethical and legal implications of using text data embeddings in healthcare	As with any technology, there are risks and potential harms associated with the use of text data embeddings, such as the potential for biased or inaccurate predictions. Future work could focus on developing guidelines and best practices for the ethical use of text data embeddings in healthcare, as well as addressing legal and privacy concerns.
Integration of text data embeddings with other sources of healthcare data	Text data embeddings are just one type of healthcare data, and integrating them with other sources of data, such as medical imaging or genomics data, could provide even more powerful insights for clinical decision-making and drug discovery. Future work could focus on developing methods to integrate different types of healthcare data in a way that is efficient and effective.
Development of methods to incorporate patient preferences and values into text data embeddings	While text data embeddings can provide valuable insights into patient health, they do not capture patient preferences and values. Future work could focus on developing methods to incorporate this information into text data embeddings, in order to develop more patient-centered healthcare.

6. Conclusion

In conclusion, data embeddings have emerged as a powerful tool for healthcare applications in recent years, with a wide range of potential use cases, including clinical decision-making, drug discovery, medical imaging, and electronic health records. Several different architectures, such as word2vec, GloVe, and transformer-based models, have been explored for data embeddings in healthcare, with promising results. However, there are still several research gaps and future works that need to be addressed, such as the development of more accurate and interpretable embeddings, more robust and reliable embeddings, and guidelines for the ethical use of data embeddings in healthcare. Despite these research gaps, significant progress has been made in the application of data embeddings in healthcare, with the potential to improve patient outcomes and advance the field of healthcare. As the field continues to evolve, we can expect to see even more applications of data embeddings in the future, as well as continued research on improving the accuracy, reliability, and ethical use of these tools.

References

[1] Keidar, D., Yaron, D., Goldstein, E., Shachar, Y., Blass, A., Charbinsky, L., Aharony, I., Lifshitz, L., Lumelsky, D., Neeman, Z., Mizrachi, M., Hajouj, M., Eizenbach, N., Sela, E., Weiss, C. S., Levin, P., Benjaminov, O., Bachar, G. N., Tamir, S., … Eldar, Y. C. (2021). COVID-19 classification of X-ray images using deep neural networks. *European Radiology*, *31*(12), 9654–9663. https://doi.org/10.1007/s00330-021-08050-1

[2] Despotovic, V., Ismael, M., Cornil, M., Call, R. M., & Fagherazzi, G. (2021). Detection of COVID-19 from voice, cough and breathing patterns: Dataset and preliminary results. *Computers in Biology and Medicine*, *138*(104944), 104944. https://doi.org/10.1016/j.compbiomed.2021.104944

[3] Kumar, V., Recupero, D. R., Riboni, D., & Helaoui, R. (2021). Ensembling Classical Machine Learning and Deep Learning Approaches for Morbidity Identification from Clinical Notes. *IEEE Access*, *9*, 7107–7126. https://doi.org/10.1109/ACCESS.2020.3043221

[4] Ye, J., Yao, L., Shen, J., Janarthanam, R., & Luo, Y. (2020). Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. *BMC Medical Informatics and Decision Making*, *20*(Suppl 11), 295. https://doi.org/10.1186/s12911-020-01318-4

[5] Heo, T. S., Kim, Y. S., Choi, J. M., Jeong, Y. S., Seo, S. Y., Lee, J. H., Jeon, J. P., & Kim, C. (2020). Prediction of stroke outcome using natural language processing-based machine learning of radiology report of brain MRI. *Journal of Personalized Medicine*, *10*(4), 1–11. https://doi.org/10.3390/jpm10040286

[6] Tian, Y., Shen, W., Song, Y., Xia, F., He, M., & Li, K. (2020). Improving biomedical named entity recognition with syntactic information. *BMC Bioinformatics*, *21*(1), 539. https://doi.org/10.1186/s12859-020-03834-6

[7] Landi, I., Glicksberg, B. S., Lee, H. C., Cherng, S., Landi, G., Danieletto, M., Dudley, J. T., Furlanello, C., & Miotto, R. (2020). Deep representation learning of electronic health records to unlock patient stratification at scale. *Npj Digital Medicine*, *3*(1), 96. https://doi.org/10.1038/s41746-020-0301-z

[8] Wang, H., Li, Y., Khan, S. A., & Luo, Y. (2020). Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. *Artificial Intelligence in Medicine*, *110*(101977), 101977. https://doi.org/10.1016/j.artmed.2020.101977

[9] Luo, J. W., & Chong, J. J. R. (2020). Review of Natural Language Processing in Radiology. *Neuroimaging Clinics of North America*, *30*(4), 447–458. https://doi.org/10.1016/j.nic.2020.08.001

[10] Afzal, M., Alam, F., Malik, K. M., & Malik, G. M. (2020). Clinical Context–Aware Biomedical Text Summarization Using Deep Neural Network: Model Development and Validation. *Journal of Medical Internet Research*, *22*(10), e19810. https://doi.org/10.2196/19810

[11] Barash, Y., Guralnik, G., Tau, N., Soffer, S., Levy, T., Shimon, O., Zimlichman, E., Konen, E., & Klang, E. (2020). Comparison of deep learning models for natural language processing-based classification of non-English head CT reports. *Neuroradiology*, *62*(10), 1247–1256. https://doi.org/10.1007/s00234-020-02420-0

[12] Obeid, J. S., Davis, M., Turner, M., Meystre, S. M., Heider, P. M., O'Bryan, E. C., & Lenert, L. A. (2020). An artificial intelligence approach to COVID-19 infection risk assessment in virtual visits: A case report. *Journal of the American Medical Informatics Association*, *27*(8), 1321–1325. https://doi.org/10.1093/jamia/ocaa105

[13] Chen, B., Li, J., Lu, G., Yu, H., & Zhang, D. (2020). Label co-occurrence learning with graph convolutional networks for multi-label chest X-ray image classification. *IEEE Journal of Biomedical and Health Informatics*, *24*(8), 2292–2302. https://doi.org/10.1109/JBHI.2020.2967084

[14] Chen, C. H., Hsieh, J. G., Cheng, S. L., Lin, Y. L., Lin, P. H., & Jeng, J. H. (2020). Emergency department disposition prediction using a deep neural network with integrated clinical narratives and structured data. *International Journal of Medical Informatics*, *139*(104146), 104146. https://doi.org/10.1016/j.ijmedinf.2020.104146

[15] Blanco, A., Perez-de-Viñaspre, O., Pérez, A., & Casillas, A. (2020). Boosting ICD multi-label classification of health records with contextual embeddings and label-granularity. *Computer Methods and Programs in Biomedicine*, *188*(105264), 105264. https://doi.org/10.1016/j.cmpb.2019.105264

[16] Malik, K. M., Krishnamurthy, M., Alobaidi, M., Hussain, M., Alam, F., & Malik, G. (2020). Automated domain-specific healthcare knowledge graph curation framework: Subarachnoid hemorrhage as phenotype. *Expert Systems with Applications*, *145*(113120), 113120. https://doi.org/10.1016/j.eswa.2019.113120

[17] Korach, Z. T., Yang, J., Rossetti, S. C., Cato, K. D., Kang, M. J., Knaplund, C., Schnock, K. O., Garcia, J. P., Jia, H., Schwartz, J. M., & Zhou, L. (2020). Mining clinical phrases from nursing notes to discover risk factors of patient deterioration. *International Journal of Medical Informatics*, *135*(104053), 104053. https://doi.org/10.1016/j.ijmedinf.2019.104053

[18] Akhtyamova, L., Martínez, P., Verspoor, K., & Cardiff, J. (2020). Testing contextualized word embeddings to improve NER in Spanish clinical case narratives. *IEEE Access*, *8*, 164717–164726. https://doi.org/10.1109/ACCESS.2020.3018688

[19] Dai, H. J., Su, C. H., & Wu, C. S. (2020). Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. *Journal of the American Medical Informatics Association*, *27*(1), 47–55. https://doi.org/10.1093/jamia/ocz120

[20] Dai, H. J., Su, C. H., & Wu, C. S. (2020). Adverse drug event and medication extraction in electronic health records via a cascading architecture with different sequence labeling models and word embeddings. *Journal of the American Medical Informatics Association*, *27*(1), 47–55. https://doi.org/10.1093/jamia/ocz120

[21] Gligic, L., Kormilitzin, A., Goldberg, P., & Nevado-Holgado, A. (2020). Named entity recognition in electronic health records using transfer learning bootstrapped Neural Networks. *Neural Networks*, *121*, 132–139. https://doi.org/10.1016/j.neunet.2019.08.032

[22] Yang, X., Lyu, T., Li, Q., Lee, C. Y., Bian, J., Hogan, W. R., & Wu, Y. (2019). A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, *19*(Suppl 5), 232. https://doi.org/10.1186/s12911-019-0935-4

[23] Yang, X., Lyu, T., Li, Q., Lee, C. Y., Bian, J., Hogan, W. R., & Wu, Y. (2019). A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC Medical Informatics and Decision Making*, *19*(Suppl 5), 232. https://doi.org/10.1186/s12911-019-0935-4

[24] Gao, C., Osmundson, S., Velez Edwards, D. R., Jackson, G. P., Malin, B. A., & Chen, Y. (2019). Deep learning predicts extreme preterm birth from electronic health records. *Journal of Biomedical Informatics*, *100*(103334), 103334. https://doi.org/10.1016/j.jbi.2019.103334

[25] Topaz, M., Murga, L., Bar-Bachar, O., McDonald, M., & Bowles, K. (2019). NimbleMiner: An Open-Source Nursing-Sensitive Natural Language Processing System Based on Word Embedding. *CIN - Computers Informatics Nursing*, *37*(11), 583–590. https://doi.org/10.1097/CIN.0000000000000557

[26] Zhang, Y., Lin, H., Yang, Z., Wang, J., Sun, Y., Xu, B., & Zhao, Z. (2019). Neural network-based approaches for biomedical relation classification: A review. *Journal of Biomedical Informatics*, *99*(103294), 103294. https://doi.org/10.1016/j.jbi.2019.103294

[27] Suárez-Paniagua, V., Rivera Zavala, R. M., Segura-Bedmar, I., & Martínez, P. (2019). A two-stage deep learning approach for extracting entities and relationships from medical texts. *Journal of Biomedical Informatics*, *99*(103285), 103285. https://doi.org/10.1016/j.jbi.2019.103285

[28] Batbaatar, E., & Ryu, K. H. (2019). Ontology-based healthcare named entity recognition from twitter messages using a recurrent neural network approach. *International Journal of Environmental Research and Public Health*, *16*(19), 3628. https://doi.org/10.3390/ijerph16193628

[29] Dai, H. J., & Wang, C. K. (2019). Classifying adverse drug reactions from imbalanced twitter data. *International Journal of Medical Informatics*, *129*, 122–132. https://doi.org/10.1016/j.ijmedinf.2019.05.017

[30] Trivedi, G., Hong, C., Dadashzadeh, E. R., Handzel, R. M., Hochheiser, H., & Visweswaran, S. (2019). Identifying incidental findings from radiology reports of trauma patients: An evaluation of automated feature representation methods. *International Journal of Medical Informatics*, *129*, 81–87. https://doi.org/10.1016/j.ijmedinf.2019.05.021

[31] Obeid, J. S., Weeda, E. R., Matuskowitz, A. J., Gagnon, K., Crawford, T., Carr, C. M., & Frey, L. J. (2019). Automated detection of altered mental status in emergency department clinical notes: A deep learning approach. *BMC Medical Informatics and Decision Making*, *19*(1), 164. https://doi.org/10.1186/s12911-019-0894-9

[32] Yang, Y., Wang, X., Huang, Y., Chen, N., Shi, J., & Chen, T. (2019). Ontology-based venous thromboembolism risk assessment model developing from medical records. *BMC Medical Informatics and Decision Making*, *19*(Suppl 4), 151. https://doi.org/10.1186/s12911-019-0856-2

[33] Oleynik, M., Kugic, A., Kasáč, Z., & Kreuzthaler, M. (2019). Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. *Journal of the American Medical Informatics Association*, *26*(11), 1247–1254. https://doi.org/10.1093/jamia/ocz149

[34] Si, Y., Wang, J., Xu, H., & Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, *26*(11), 1297–1304. https://doi.org/10.1093/jamia/ocz096

[35] Gargiulo, F., Silvestri, S., Ciampi, M., & De Pietro, G. (2019). Deep neural network for hierarchical extreme multi-label text classification. *Applied Soft Computing Journal*, *79*, 125–138. https://doi.org/10.1016/j.asoc.2019.03.041

[36] Yao, L., Mao, C., & Luo, Y. (2019). Clinical text classification with rule-based features and knowledge-guided convolutional neural networks.

BMC Medical Informatics and Decision Making, *19*(Suppl 3), 71. https://doi.org/10.1186/s12911-019-0781-4

[37] Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Medical Informatics*, *7*(2), e12239. https://doi.org/10.2196/12239

[38] Dimitriadis, D., & Tsoumakas, G. (2019). Word embeddings and external resources for answer processing in biomedical factoid question answering. *Journal of Biomedical Informatics*, *92*(103118), 103118. https://doi.org/10.1016/j.jbi.2019.103118

[39] Ning, W., Chan, S., Beam, A., Yu, M., Geva, A., Liao, K., Mullen, M., Mandl, K. D., Kohane, I., Cai, T., & Yu, S. (2019). Feature extraction for phenotyping from semantic and knowledge resources. *Journal of Biomedical Informatics*, *91*(103122), 103122. https://doi.org/10.1016/j.jbi.2019.103122

[40] Li, Y., Jin, R., & Luo, Y. (2019). Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks (Seg-GCRNs). *Journal of the American Medical Informatics Association*, *26*(3), 262–268. https://doi.org/10.1093/jamia/ocy157

[41] Li, X., Wang, H., He, H., Du, J., Chen, J., & Wu, J. (2019). Intelligent diagnosis with Chinese electronic medical records based on convolutional neural networks. *BMC Bioinformatics*, *20*(1), 62. https://doi.org/10.1186/s12859-019-2617-8

[42] Topaz, M., Murga, L., Gaddis, K. M., McDonald, M. V., Bar-Bachar, O., Goldberg, Y., & Bowles, K. H. (2019). Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches. *Journal of Biomedical Informatics*, *90*(103103), 103103. https://doi.org/10.1016/j.jbi.2019.103103

[43] Wunnava, S., Qin, X., Kakar, T., Sen, C., Rundensteiner, E. A., & Kong, X. (2019). Adverse Drug Event Detection from Electronic Health Records Using Hierarchical Recurrent Neural Networks with Dual-Level Embedding. *Drug Safety*, *42*(1), 113–122. https://doi.org/10.1007/s40264-018-0765-9

[44] Chapman, A. B., Peterson, K. S., Alba, P. R., DuVall, S. L., & Patterson, O. V. (2019). Detecting Adverse Drug Events with Rapidly Trained Classification Models. *Drug Safety*, *42*(1), 147–156. https://doi.org/10.1007/s40264-018-0763-y

[45] Guan, M., Cho, S., Petro, R., Zhang, W., Pasche, B., & Topaloglu, U. (2019). Natural language processing and recurrent network models for identifying genomic mutation-associated cancer treatment change from patient progress notes. *JAMIA Open*, *2*(1), 139–149. https://doi.org/10.1093/jamiaopen/ooy061

[46] Zhao, B. (2019). Clinical Data Extraction and Normalization of Cyrillic Electronic Health Records Via Deep-Learning Natural Language Processing. *JCO Clinical Cancer Informatics*, *3*(3), 1–9. https://doi.org/10.1200/cci.19.00057

[47] Fan, Y., Pakhomov, S., McEwan, R., Zhao, W., Lindemann, E., & Zhang, R. (2019). Using word embeddings to expand terminology of dietary supplements on clinical notes. *JAMIA Open*, *2*(2), 246–253. https://doi.org/10.1093/jamiaopen/ooz007

[48] Chowdhury, S., Dong, X., Qian, L., Li, X., Guan, Y., Yang, J., & Yu, Q. (2018). A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records. *BMC Bioinformatics*, *19*(17), 75–84. https://doi.org/10.1186/s12859-018-2467-9

[49] Xu, K., Zhou, Z., Gong, T., Hao, T., & Liu, W. (2018). SBLC: A hybrid model for disease named entity recognition based on semantic bidirectional LSTMs and conditional random fields. *BMC Medical Informatics and Decision Making*, *18*(Suppl 5), 114. https://doi.org/10.1186/s12911-018-0690-y

[50] Catling, F., Spithourakis, G. P., & Riedel, S. (2018). Towards automated clinical coding. *International Journal of Medical Informatics*, *120*, 50–61. https://doi.org/10.1016/j.ijmedinf.2018.09.021

[51] Song, H. J., Jo, B. C., Park, C. Y., Kim, J. D., & Kim, Y. S. (2018). Comparison of named entity recognition methodologies in biomedical documents. *BioMedical Engineering Online*, *17*(2), 1–14. https://doi.org/10.1186/s12938-018-0573-6

[52] Segura-Bedmar, I., Colón-Ruíz, C., Tejedor-Alonso, M. Á., & Moro-Moro, M. (2018). Predicting of anaphylaxis in big data EMR by exploring machine learning approaches. *Journal of Biomedical Informatics*, *87*, 50–59. https://doi.org/10.1016/j.jbi.2018.09.012

[53] Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., & Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of*

Biomedical Informatics, *87*, 12–20. https://doi.org/10.1016/j.jbi.2018.09.008

[54] Ye, C., & Fabbri, D. (2018). Extracting similar terms from multiple EMR-based semantic embeddings to support chart reviews. *Journal of Biomedical Informatics*, *83*, 63–72. https://doi.org/10.1016/j.jbi.2018.05.014

[55] Menger, V., Scheepers, F., & Spruit, M. (2018). Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Applied Sciences (Switzerland)*, *8*(6), 981. https://doi.org/10.3390/app8060981

[56] Zech, J., Pain, M., Titano, J., Badgeley, M., Schefflein, J., Su, A., Costa, A., Bederson, J., Lehar, J., & Oermann, E. K. (2018). Natural language-based machine learning models for the annotation of clinical radiology reports. *Radiology*, *287*(2), 570–580. https://doi.org/10.1148/radiol.2018171093

[57] Zhou, D., Miao, L., & He, Y. (2018). Position-aware deep multi-task learning for drug–drug interaction extraction. *Artificial Intelligence in Medicine*, *87*, 1–8. https://doi.org/10.1016/j.artmed.2018.03.001

[58] Koola, J. D., Davis, S. E., Al-Nimri, O., Parr, S. K., Fabbri, D., Malin, B. A., Ho, S. B., & Matheny, M. E. (2018). Development of an automated phenotyping algorithm for hepatorenal syndrome. *Journal of Biomedical Informatics*, *80*, 87–95. https://doi.org/10.1016/j.jbi.2018.03.001

[59] Jadhav, S. B. ., & Kodavade, D. V. . (2023). Enhancing Flight Delay Prediction through Feature Engineering in Machine Learning Classifiers: A Real Time Data Streams Case Study. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2s), 212–218. https://doi.org/10.17762/ijritcc.v11i2s.6064

[60] Duarte, F., Martins, B., Pinto, C. S., & Silva, M. J. (2018). Deep neural models for ICD-10 coding of death certificates and autopsy reports in free-text. *Journal of Biomedical Informatics*, *80*, 64–77. https://doi.org/10.1016/j.jbi.2018.02.011

[61] Si, Y., & Roberts, K. (2018). A Frame-Based NLP System for Cancer-Related Information Extraction. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2018*, 1524–1533. /pmc/articles/PMC6371330/

[62] Glicksberg, B. S., Miotto, R., Johnson, K. W., Shameer, K., Li, L., Chen, R., & Dudley, J. T. (2018). Automated disease cohort selection using word embeddings from Electronic Health Records. *Pacific Symposium on Biocomputing*, *0*(212669), 145–156. https://doi.org/10.1142/9789813235533_0014

[63] Luo, Y., Cheng, Y., Uzuner, Ö., Szolovits, P., & Starren, J. (2018). Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. *Journal of the American Medical Informatics Association*, *25*(1), 93–98. https://doi.org/10.1093/jamia/ocx090

[64] Xie, J., Liu, X., & Zeng, D. D. (2018). Mining e-cigarette adverse events in social media using Bi-LSTM recurrent neural network with word embedding representation. *Journal of the American Medical Informatics Association*, *25*(1), 72–80. https://doi.org/10.1093/jamia/ocx045

[65] Banerjee, I., Chen, M. C., Lungren, M. P., & Rubin, D. L. (2018). Radiology report annotation using intelligent word embeddings: Applied to multi-institutional chest CT cohort. *Journal of Biomedical Informatics*, *77*, 11–20. https://doi.org/10.1016/j.jbi.2017.11.012

[66] Wang, A., Wang, J., Lin, H., Zhang, J., Yang, Z., & Xu, K. (2017). A multiple distributed representation method based on neural network for biomedical event extraction. *BMC Medical Informatics and Decision Making*, *17*(Suppl 3). https://doi.org/10.1186/s12911-017-0563-9

[67] Weng, W. H., Wagholikar, K. B., McCray, A. T., Szolovits, P., & Chueh, H. C. (2017). Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. *BMC Medical Informatics and Decision Making*, *17*(1). https://doi.org/10.1186/s12911-017-0556-8

[68] Jauregi Unanue, I., Zare Borzeshi, E., & Piccardi, M. (2017). Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *Journal of Biomedical Informatics*, *76*, 102–109. https://doi.org/10.1016/j.jbi.2017.11.007

[69] Lin, C., Hsu, C. J., Lou, Y. S., Yeh, S. J., Lee, C. C., Su, S. L., & Chen, H. C. (2017). Artificial intelligence learning semantics via external resources for classifying diagnosis codes in discharge notes. *Journal of Medical Internet Research*, *19*(11). https://doi.org/10.2196/jmir.8344

[70] Lyu, C., Chen, B., Ren, Y., & Ji, D. (2017). Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinformatics*, *18*(1), 462. https://doi.org/10.1186/s12859-017-1868-5

[71] Cho, H., Choi, W., & Lee, H. (2017). A method for named entity normalization in biomedical articles: Application to diseases and plants. *BMC Bioinformatics*, *18*(1). https://doi.org/10.1186/s12859-017-1857-8

[72] Sulieman, L., Gilmore, D., French, C., Cronin, R. M., Jackson, G. P., Russell, M., & Fabbri, D. (2017). Classifying patient portal messages using Convolutional Neural Networks. *Journal of Biomedical Informatics*, *74*, 59–70. https://doi.org/10.1016/j.jbi.2017.08.014

[73] Mr. Anish Dhabliya. (2013). Ultra Wide Band Pulse Generation Using Advanced Design System Software . International Journal of New Practices in Management and Engineering, 2(02), 01 - 07. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/14

[74] Jimeno Yepes, A. (2017). Word embeddings and recurrent neural networks based on Long-Short Term Memory nodes in supervised biomedical word sense disambiguation. *Journal of Biomedical Informatics*, *73*, 137–147. https://doi.org/10.1016/j.jbi.2017.08.001

[75] Kuang, S., & Davison, B. D. (2017). Learning word embeddings with chi-square weights for healthcare tweet classification. *Applied Sciences (Switzerland)*, *7*(8), 846. https://doi.org/10.3390/app7080846

[76] Luo, Y. (2017). Recurrent neural networks for classifying relations in clinical notes. *Journal of Biomedical Informatics*, *72*, 85–95. https://doi.org/10.1016/j.jbi.2017.07.006

[77] Tao, C., Filannino, M., & Uzuner, Ö. (2017). Prescription extraction using CRFs and word embeddings. *Journal of Biomedical Informatics*, *72*, 60–66. https://doi.org/10.1016/j.jbi.2017.07.002

[78] Gridach, M. (2017). Character-level neural network for biomedical named entity recognition. *Journal of Biomedical Informatics*, *70*, 85–91. https://doi.org/10.1016/j.jbi.2017.05.002

[79] Nguyen, T., Larsen, M. E., O'Dea, B., Phung, D., Venkatesh, S., & Christensen, H. (2017). Estimation of the prevalence of adverse drug reactions from social media. *International Journal of Medical Informatics*, *102*, 130–137. https://doi.org/10.1016/j.ijmedinf.2017.03.013

[80] Kang, T., Zhang, S., Xu, N., Wen, D., Zhang, X., & Lei, J. (2017). Detecting negation and scope in Chinese clinical notes using character and word embedding. *Computer Methods and Programs in Biomedicine*, *140*, 53–59. https://doi.org/10.1016/j.cmpb.2016.11.009

[81] Wu, Y., Jiang, M., Xu, J., Zhi, D., & Xu, H. (2017). Clinical Named Entity Recognition Using Deep Learning Models. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2017*, 1812–1819.

[82] Banerjee, I., Madhavan, S., Goldman, R. E., & Rubin, D. L. (2017). Intelligent Word Embeddings of Free-Text Radiology Reports. *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, *2017*, 411–420. https://arxiv.org/abs/1711.06968v1

[83] Zhang, S., Kang, T., Zhang, X., Wen, D., Elhadad, N., & Lei, J. (2016). Speculation detection for Chinese clinical notes: Impacts of word segmentation and embedding models. *Journal of Biomedical Informatics*, *60*, 334–341. https://doi.org/10.1016/j.jbi.2016.02.011

[84] Chen, C., Zhao, X., Wang, J., Li, D., Guan, Y., & Hong, J. (2022). Dynamic graph convolutional network for assembly behavior recognition based on attention mechanism and multi-scale feature fusion. *Scientific Reports*, *12*(1), 1–13. https://doi.org/10.1038/s41598-022-11206-8

[85] Raza, A., Mehmood, A., Ullah, S., Ahmad, M., Choi, G. S., & On, B. W. (2019). Heartbeat sound signal classification using deep learning. *Sensors (Switzerland)*, *19*(21). https://doi.org/10.3390/s19214819

[86] Varghese, A., Agyeman-Badu, G., & Cawley, M. (2020). Deep learning in automated text classification: a case study using toxicological abstracts. *Environment Systems and Decisions*, *40*(4), 465–479. https://doi.org/10.1007/s10669-020-09763-2