Multimedia Systems (2022) 28:1465–1479
https://doi.org/10.1007/s00530-022-00948-0

SPECIAL ISSUE ARTICLE

Deep learning in multimedia healthcare applications: a review

Diana P. Tobón V[1] · M. Shamim Hossain[2] · Ghulam Muhammad[3] · Josu Bilbao[4] · Abdulmotaleb El Saddik[5,6]

Received: 10 February 2021 / Accepted: 22 April 2022 / Published online: 24 May 2022

Abstract
The increase in chronic diseases has affected the countries' health system and economy. With the recent COVID-19 virus, humanity has experienced a great challenge, which has led to make efforts to detect it and prevent its spread. Hence, it is necessary to develop new solutions that are based on technology and low cost, to satisfy the citizens' needs. Deep learning techniques is a technological solution that has been used in healthcare lately. Nowadays, with the increase in chips processing capabilities, increase size of data, and the progress in deep learning research, healthcare applications have been proposed to provide citizens' health needs. In addition, a big amount of data is generated every day. Development in Internet of Things, gadgets, and phones has allowed the access to multimedia data. Data such as images, video, audio and text are used as input of applications based on deep learning methods to support healthcare system to diagnose, predict, or treat patients. This review pretends to give an overview of proposed healthcare solutions based on deep learning techniques using multimedia data. We show the use of deep learning in healthcare, explain the different types of multimedia data, show some relevant deep learning multimedia applications in healthcare, and highlight some challenges in this research area.

Keywords Chronic disease · COVID-19 · Deep learning · Healthcare · Monomedia · Multimedia · Multimodal

1 Introduction

The World Health Organization (WHO) has reported that two-thirds of the world's death are due to chronic diseases such as diabetes, cancer, cardiovascular disease, and respiratory diseases [1]. Those chronic diseases have burdened the existing healthcare systems since patients go frequently to the hospitals for periodic checkups or urgencies [2]. Obesity, and the increase of elderly population, are also challenging those systems. It is expected by 2050 that 22% of the world population will be over 60 years old [3]. On the other hand, COVID-19 pandemic has challenged the healthcare system [4]. Initial reports of the new coronavirus were published in Wuhan, China in December 2019. At the present, there are more than 430 million infected people and more than 5 million deaths around the world [5]. The needs mentioned above have opened doors to do research in the healthcare industry to provide applications that improve citizens' well-being and quality of life.

To offer better health services, healthcare research has been increased in recent years. Healthcare industry provides from basic to high health services to patients. Remote or

✉ Diana P. Tobón V
dtobon@udemedellin.edu.co

M. Shamim Hossain
mshossain@ksu.edu.sa

Ghulam Muhammad
ghulam@ksu.edu.sa

Josu Bilbao
jbilbao@ikerlan.es

Abdulmotaleb El Saddik
elsaddik@uottawa.ca

[1] Department of Telecommunications Engineering, Universidad de Medellín, Medellín, Colombia

[2] Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

[3] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

[4] Head of Research Department - ICT (IoT Digital Platforms, Data Analytics & Artificial Intelligence) IKERLAN, Arrasate, Spain

[5] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

[6] University of Ottawa, Ottawa, Canada

mobile healthcare applications allow to overcome patients' needs such as deficit in the number of specialist doctors, remote areas with lack of health services, high traffic congestion that avoid reaching a hospital on time, patients extremely occupied, and patients that do not like to see a doctor [6]. Moreover, the increase of elderly people has reduced the ratio of specialized doctors to patients [7]. The aforementioned factors have created the need for smart healthcare systems, where a patient can be monitored by specialized doctors from a hospital (i.e., medical applications) or the patient can by himself/herself be monitored at home (i.e., non-medical applications) [8]. In both cases, the patient's physiological signals are measured using wearable devices and sent to a personal computer or smartphone or to the cloud to be analyzed, by avoiding the need of clinic visitation [6]. Thus, healthcare applications let make corresponding interventions in time [9].

The growth of technology has allowed the development of research in medical field [10]. The advancement of technologies such as Internet of Things (IoT), data processing, wireless communications, machine learning algorithms, artificial intelligence (AI), 5G, big data, cloud and edge computing, have opened doors to new healthcare applications [6]. Preventive healthcare using mobile applications have been developed for diabetes [11, 12], obesity [13], mental health [14], cardiac diseases [15–17], and chronic diseases [18, 19]. Applications such as monitoring real-time human activities for elderly people are used for medical rehabilitation and elderly care. Daily activities are recorded to ensure safety, hence avoiding falls and abnormal behaviors [20]. To this end, wearable devices and mobile sensors are used to monitor activities of daily living (ADL). Those wearables are placed on-body (e.g., inertial sensors or smartphones) at different body locations (e.g., chest, head, waist, wrist hand). They collect and transfer data about body postures through a wireless sensor network [21]. Other important application is early cancer detection in different organs such as breast, lung, brain, and liver. Development in image processing technology has let to achieve this earlier detection [10]. Additionally, analysis of neurodegenerative diseases such as Alzheimer are also possible thanks to image processing [22], as well as image biomarkers are used in clinical practice for diagnosis. Therefore, emerging technologies in AI and IoT allow taking better decisions of patient's diagnoses, which improves healthcare services [23].

The advance of IoT technologies, wearable sensors, and wireless sensor Networks (WSN) is opportune to analyze multimedia data from human body anywhere and anytime [9]. The benefits of multimedia data were listed by Oracle [24]. They are less prone to loss, are stored digitally, can be used by computers without human intervention, speed up to recover media files, let data exchange among servers and archives, and allow sharing media content in different sources [25]. Multimedia community have raised their attention to health research (e.g., wellbeing and every day healthy living). The main interest is how the produced multimedia data can be used efficiently in healthcare systems [26]. That collected data from several sensors stand the importance of multimodal and multimedia data sensing and fusion, where the analysis of large amounts of multimedia data is an active research area [27]. There is a tremendous amount of generated multimedia data every day, which complicates their analysis and storage. Traditional machine learning techniques use only one sensing modality, but a robust healthcare application contains several modalities [28]. Deep learning approaches have helped the emergence of new healthcare solutions on domains such as computer vision, speech recognition, and natural language processing [29], which are multimedia-based clinical decision support applications [30]. Those approaches allow predicting and detecting different chronic diseases, so deriving insights from data [31]. They provide tools to extract image characteristics without the need to know about the underlying process [22]. However, deep learning techniques require a significant knowledge about the data, and high computational resources [32]. Recently, it has been an increase of computational parallel power of computer using Graphics Processing Unit (GPU) and clusters, which lets the development of deep learning models, specifically, Convolutional Neural Networks (CNN) for image classification. GPU architecture performs well for matrix operations in parallel, thus GPU alongside with Central Processing Unit (CPU) are well suited for acceleration in deep learning applications [25]. CNN models outperform other classification tasks [33]. Nowadays, with the COVID-19 pandemic, chest computed tomography (CT) data analysis and classification using deep learning have been developed for monitoring, screening, and predicting the COVID-19 virus [34–38].

Therefore, this review pretends give an overview of how deep learning has been used in healthcare applications using multimedia data, since deep learning has arisen as a solution for multimodal, multimedia, unstructured, large, and heterogeneous data, where computers can extract relevant features from data without the need of human intervention, by learning data representations automatically [29]. The rest of this paper is organized as follows as shown in Fig. 1. Section 2 describes the use of deep learning methods in healthcare applications, thus showing the advantages of those techniques to improve performance. Section 3 explains the different types of multimedia data used in healthcare applications. Section 4 highlights relevant multimedia and/or monomedia healthcare applications that use deep learning methods. Section 5 describes challenges in this research area. Finally, Sect. 6 presents the conclusion of the review.

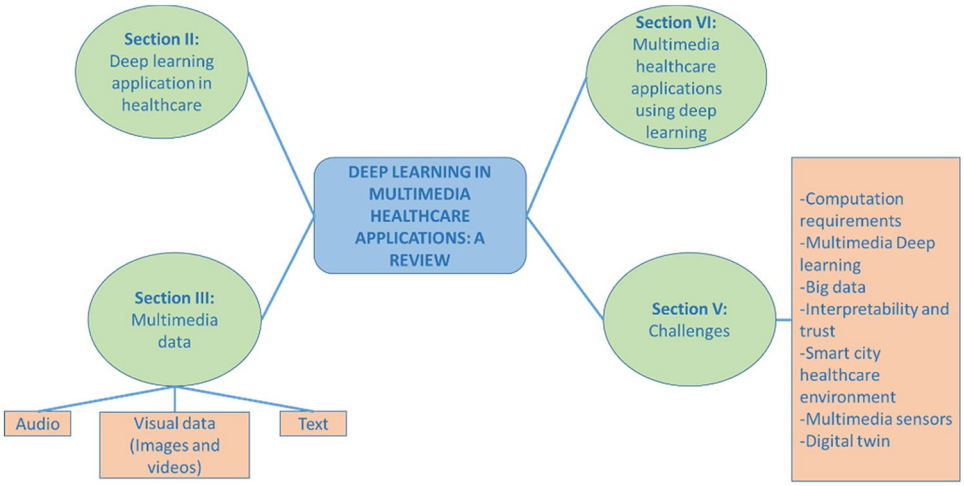

Fig. 1 Taxonomy of this manuscript

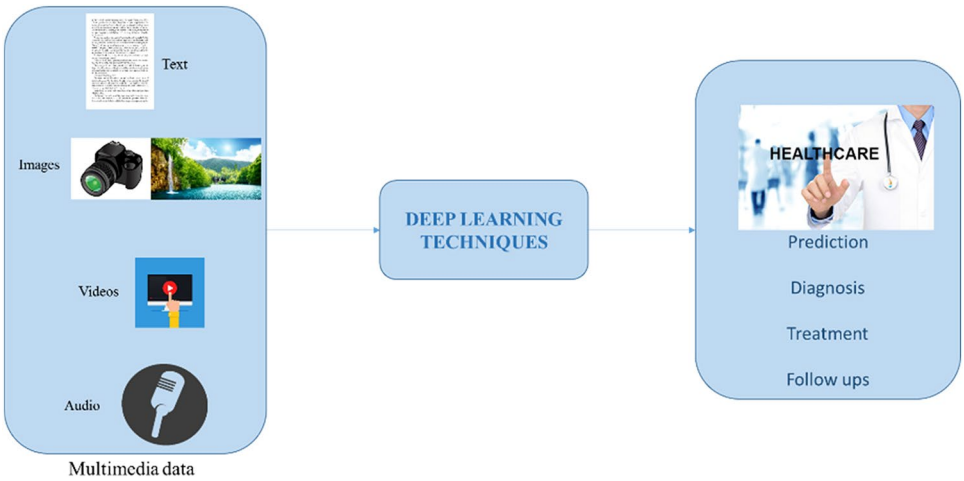

Fig. 2 Deep learning applied in healthcare applications

2 Deep learning aplication in healthcare

The traditional healthcare system uses expert's knowledge to diagnose and predict medical issues. However, the experts can miss data new patterns and the definition of feature space. As a solution, machine and deep learning approaches can predict and extract those patterns and feature space automatically [39]. The burgeoning of digital healthcare data has driven the data medical research based on machine and deep learning techniques. Those techniques are powerful for big data and perform well in feature representation and pattern recognition [40]. Figure 2 shows a representation of how deep learning is used in healthcare applications. Multimedia data is used to feed a neural network, which delivers a prediction, a diagnosis, a treatment or a follow up in the healthcare system. For example, deep learning approaches have been used in applications related to visual feature extraction, speech identification, and textual analysis. Healthcare monitoring involves the use of wearable devices and smartphones, which transfer data to mobile apps or a medical center. That data can be complex and noisy, where deep learning model offers a novel and efficient solutions for healthcare data. Thus, the goal of using deep learning is representation, where input data features can be represented by training multilayer neural network effectively [40].

Deep learning builds neural networks based on the structure of the brain by being inspired from neuroscience. Deep neural networks (DNN) approaches can learn classification and feature representation from raw data [41]. A deep architecture is accomplished by adding more hidden layers to the neural network, which allows capturing nonlinear relationships [42]. Traditional deep learning applications have been implemented using only one modality (i.e., or multimedia data) such as text, image, audio or video. In multimodal applications, deep learning solutions analyze all the information from different modalities to find logical connections between all of them to make decisions that are more accurate [29]. Data coming from those modalities are larger and

complex [43]. Multimodal data consist of data from different sensors, where deep learning approaches use it to learn a complex task. In each modality, an automatic hierarchical representation is learned. Several solutions use a fusion that integrates different modalities, where smart healthcare applications are possible with deep learning techniques and cloud technologies, thus providing real-time services and fast response. Hence, complex signals are transmitted and processed with low delay to medical personnel can realize an initial analysis [44]. As a result, hierarchical representations are learned from raw data, which helps to control how to fuse the learned representations. Henceforth, a shared representation or fusion layer is built to merge all the modalities to the network learn a joint representation. A fusion layer can be a layer with several hidden units, where the number of hidden units will depend on the number of modalities [39].

Deep learning approaches have influenced medical applications. They are useful to handle multi-modality data, learn relationship among data, extract predictive models, and analyze data patterns in real time. For example, deep learning can capture patterns of patients from clinical data to make predictions or to diagnose diseases. By monitoring a patient for a period, accumulated clinical data can be analyzed by specialized personnel to make prognosis. DNN allow identification of clinical notes, where it can extract information from unstructured text. In addition, early warning can be done since DNN will help to detect and diagnose a disease early as well as to do preventive treatment, consequently reducing healthcare cost and saving lives [40].

Deep learning combined with large volume of data and GPU enable to explore medical data to achieve precision medicine [40]. In medical image, it is hard to extract handcrafted features, whereas deep learning simplifies that extraction and improves the precision in diagnosis. Used deep learning structures in healthcare are CNN, recurrent neural networks (RNNs), and Auto Encoders (AEs) [39]. CNNs are based on their shared-weights convolutional kernels as well as characteristics of translation invariance [45]. This is the deep learning model more used in clinical applications (e.g., image diagnosis, electrocardiogram monitoring, image segmentation, and diagnose from clinical notes [46, 47]). It detects anomalous findings in different modality images such as histology or tomography scans in large-scale images from thousands of patients. In addition, low-level features are learned and fine-tuned using different images and networks [48]. Using CNN has helped to reduce the vanishing gradient problem, fortify feature propagation, allow feature reuse, and reduce the number of parameters [49]. RNNs, in turn, are a powerful tool for varying length sequences such as audios, videos, and sentences [50]. It has been used for speech recognition, sentence classification, and machine translation. For example, RNN are used for analyzing clinical data to predict and diagnose diseases, thus helping to make treatments timely. On the other hand, deep auto encoders are used for feature extraction or dimensionality reduction to reconstruct the inputs. Thus, auto encoders are used in clinical applications to extract features [40].

3 Multimedia data

Multimodal or multimedia data are a combination of various types of data from different media or multiple sensors such as texts, images, videos, or audios. By extracting multimodal data, complementary information can be extracted from each of the modalities to improve the performance application [43]. Modality refers to a specific way to encode information. Different viewpoints of a physiological object using multimodal data, give more information that is complementary to the analysis. In speech recognition, for example, not only the audio signals give valuable information but also visual modality about lip and mouth motion [50]. Multimedia representation learns features from different modalities in an automatically way, where those modalities have correlations and associations among them [51].

With the burgeoning online services and wireless technologies, multimedia data are generated every day [25]. The availability of these data has enabled the development of several healthcare applications. Nevertheless, that amount of data is difficult to process, collect, manage, and store, thus demanding new research directions in multimedia big data analytics [29]. Advancements in multimedia data research are related to computing clusters, new hardware developments, and new algorithms for huge amount of data. Multimedia research analytics studies the understanding and visualization of different types of data to solve challenges in real-world applications [25].

Multimedia data are characterized as heterogeneous, unstructured, and large. It involves data coming from different sources such as sensors, cameras, and social networks, to name a few. That heterogeneity nature imposes the need to transform that data into a singular format to be analyzed [27]. In healthcare applications, there is a variety of data such as patient records, medical images, physician notes, and radiographic films [25]. Deep learning approaches are a good solution for multimedia data since features can be extracted automatically from data without human intervention and can represent multiple levels of abstraction [29]. They also address the challenges about modeling several variables simultaneously by integrating multimodal heterogeneous data to improve accuracy [41].

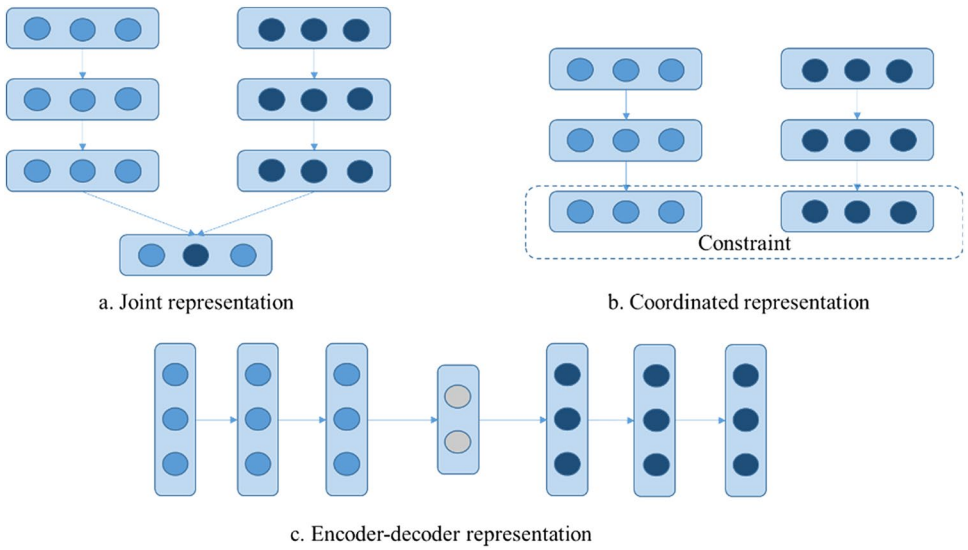

Fig. 3 Three types for multimodal representation. **a** Joint representation, **b** Coordinated representation, and **c** Encoder–decoder representation [50]

The main objective of multimodal representation is to reduce the distribution gap in a common subspace, hence keeping modality specific characteristics. Multimodal representation methods can be categorized into three types such as joint representation, coordinated representation, and encoder–decoder as shown in Fig. 3 [50]. Joint representation projects all the unimodal representations into a global shared subspace, by fusing all the multimodal features. Each modality is encoded through an individual neural network. Then, they are mapped to a shared subspace to extract and fuse similarities into a single vector. Coordinated representation, in turn, uses cross-modal similarity models and maximizes similarities or correlations. Under some constraint, it learns coordinated representations for each modality, thus helping to preserve characteristics that are specific for each modality. Finally, encoder–decoder method learns an intermediate presentation, so mapping one modality into another. It is composed of an encoder that maps source modality into a vector, and a decoder that takes that vector to generate a novel sample [43].

3.1 Audio

In healthcare, for medical devices, real-time audio (e.g., speech) analytical applications are required. Audio analytical consist in extracting valuable information from the unstructured raw audio data [27]. With the help of wearable technology, human body sounds such as heart rate, voice, breathe sound, or digestive system can be recorded [25]. For example, Opensmile [52] is used to extract acoustic features such as pitch, voice intensity, and Mel-frequency cepstral coefficients.

3.2 Visual data

Visual data includes images and videos (i.e., sequence of images). A big challenge of visual data is the huge amount of information, which requires big data solutions [25]. CNNs are the most popular models used for image feature learning. Existing CNN models are LeNet [53], AlexNet [54], GoogleNet [55], VGGNet [56], and ResNet [57]. Those pre-trained versions of CNN are a good solution when it is required high amount of training data and computation resources [50].

Video data is used in applications that include fraud detection, surveillance, healthcare, and crime detection, to name a few. It consists of a sequence of images to be analyzed one by one to extract important information [27]. In healthcare applications, multiple imaging modalities are used such as CT, Magnetic Resonance Imaging (MRI), ultrasound imaging, Positron Emission Tomography (PET), functional MRI (fMRI), X-ray, Optical Coherence Tomography (OCT), and microscopy image [58]. Those modalities are being used by medical experts in diagnosis and treatments [43]. The used techniques for images can be used for videos since the input of each time step is an image. Handcrafted features are also used in addition to deep learning features. For example, OpenFace allows extracting facial landmarks, head pose, and eye gaze [59]. However, due to the size of visual data, it is required high-performance computation and technologies such as cloud computing for doing analytics research in the mentioned applications [27].

Medical images (i.e., first-hand information) reflect patient's status and allow detecting pathologies in organs. For example, eyes are analyzed using Ophthalmic imaging [60], brain, cardiac system, bone, and joints pathologies using MRI [61], chest and abdominal organs using CT [62],

and for chest and breast using X-ray [63]. Medical images have a vector format, 2D or 3D pixel values that can be used in deep learning methods such as multilayer neural networks and CNN to detect pathologies, for image segmentation, and disease prediction [40].

3.3 Text

Text multimedia data can be in the form of metadata, social media feeds, or web pages. Text data can be structured and unstructured. Structured data are analyzed using database techniques of query retrieval. Unstructured data, in turn, needs to be transformed into structured data to be analyzed [64]. Emotion analysis from text multimedia data is one of the applications in healthcare [27]. Among text analytic techniques are Information Extraction (IE), sentiment analysis, summarization, and Question Answering (QA) [65]. SparkText [66] is a text-mining framework for big data text analytics in biomedical data [25]. Electronic Health Records (EHRs) can be analyzed with deep learning techniques to improve healthcare quality, by prediction diseases or complications. EHRs contain several modalities of clinical data, thus requiring an adequate fusion technique. EHRs can be discharge summaries (e.g., lab tests results, physician diagnoses, medical history, and treatments, to name a few), measurement reports, and death certificates [31, 41].

Table 1 shows a comparison between deep multimedia and traditional learning methods. Approaches for analyzing multimedia data using deep learning have many advantages compared to traditional machine learning techniques [43]. In addition, existing multimedia datasets are found in [67–70].

4 Multimedia healthcare applications using deep learning

A large branch of big data is from medical applications. The number of these applications has increased fast in lately years, thus imposing the need of healthcare research methods. Medical data come from sensors, records, images, and videos from patients. Traditional diagnosis relies on doctors' empirical knowledge based on symptoms, antecedents, and consultation information. However, that diagnostic method can caused misdiagnosis due to human error. In addition, medical experts or patients may not use efficiently the collected data [26]. Hence, science and technology developments in big data analysis have contributed with efficient diagnosis and treatments [23]. The goal is to design a reliable system to study the relationship between symptoms and diseases, explore treatments, diseases trends, and risk factors [40]. Over the last decade, images and videos have become important for medical imagining in different specialties. However, this explosion of data requires advance multimedia technology for diagnosis decision support, examination, surgery, and real-time support to help medical professionals [26]. Existing medical multimedia applications are shown to follow and summarized in Table 2.

Applications that use *text data* have been developed in healthcare. An application that analyzes electronic medical records (EMR) with CNN to predict future risk automatically was proposed in [71]. The authors present Deepr, which is an end-to-end learning system. The system transforms a record into a discrete sequence with coded time gaps and hospital transfers. A multilayer architecture based on CNN (i.e., embedding, convolution, pooling, and classifier) is used to detect and combines local clinical motifs to predict risk. A large private hospital dataset collected in Australia was used for evaluation, where an Area Under Curve (AUC) of 0.819 was reported. In [72], the history of patients in EHR is used to observe medical conditions and determine medications for the next visit. The authors used RNNs, where the inputs are encounter records such as diagnosis codes, medication or procedure codes. Then, the RNN predicts diagnosis and medications for the following visit. The used dataset was care patients from Sutter Health Palo Alto Medical Foundation. The performance reported was a recall up to 79%. Another approach is a fusion of multimodal signal to analyze different motion patterns [21]. The authors integrate on-body sensor data, context sensor data, and personal profile data (i.e., *text data*) such as pure acceleration data (i.e., nine

Table 1 Comparison between deep multimedia and traditional learning methods [43]

Deep multimedia learning methods	Traditional learning methods
Features are learned from the fused data	Prior knowledge is required to extract features manually
Little preprocessing of the input data is required	Preprocessing is needed for early fusion
Dimensionality reduction is done by the architecture	It requires perform dimensionality reduction
Performs early, intermediate, or late fusion	Performs early or late fusion
Fusion architecture is learned during training	Fusion architecture is designed manually
Requires many data for training	Not a lot of training data is required
GPUs are need for training time	Use of GPUs are not critical
Hyper parameter tuning is needed	Hyper parameter tuning is not needed

Table 2 Deep Learning in multimedia healthcare applications

Reference	Application	Multimedia data	Deep learning method	Database
[71]	Predict future risk	Electronic Medical Records	A multilayer architecture based on CNN (i.e., embedding, convolution, pooling, and classifier)	Private hospital dataset collected in Australia
[72]	Diagnose and determine medications for the next visit	Electronic Health Record	Recurrent Neural Networks	Patients from Sutter Health Palo Alto Medical Foundation
[21]	Analyze different motion patterns	Personal profile data, sound level	Long-short-term memory-based	Two free data sets using smartphones and on-body wearable devices
[10]	Lung cancer detection	Images	CNN with seven layers and trained in transfer learning	Data collected in K1 hospital located in Kirkuk city, Iraq
[22]	Detection of AD and in the diagnosis of dementia	MRI Images	Deep convolutional auto encoder (CAE) architecture	ADNI database
[33]	Diagnose AD	MRI Images	Multi-projection fusion with CNN	ADNI database
[73]	Detection of AD	Images and text	Deep network of auto-encoders	ADNI database
[37]	Diagnosis and prognosis of COVID-19	Chest computed tomography (CT) images	CNN	Patients from the participating hospitals between September 2016 and January 2020
[76]	COVID-19 classification	Chest X-rays and computerized tomography images of the lungs	CNN and feed-forward neural network	Cohen's Database and Kermany's Database
[78]	COVID-19 classification	Chest X-rays images	Proposed CNN model	Constructed dataset containing 180 COVID-19 and 200 normal
[79]	COVID-19 classification	Chest CT (CCT) images	Pre-trained CNN	Dataset from local hospitals
[80]	Diagnosis of COVID-19 patients	Chest X-Ray (CXR) images from COVID-19, normal, and other pneumonia categories	A deep meta learning framework based on Siamese neural network	Open access CXR dataset from multiple sources
[81]	COVID-19 screening and mass surveillance	Chest CT scan	ResNet50, Inception V3, Deep tree	Kaggle and GitHub repositories
[82]	Estimate food attributes such as ingredients and nutritional values	Images	CNN	Food-101 and Image-net
[6]	Voice pathology detection	Voice signals	CNN	Saarbrucken voice disorder database
[7]	Voice pathology detection	Voice signals	CNN and Multilayer Perceptron (MPL)	Saarbrucken voice disorder database
[87]	Automated egocentric human action and activity recognition	Videos	CNN and Long-Short-Term Memory	Multimodal Insulin Self-Injection (ISI) dataset
[20]	Action recognition	Videos	3D convolutional neural network and LSTM	UCF101 dataset
[89]	Human fall detection	Videos	RNN and LSTM	NTU RGB+D Action Recognition Dataset
[41]	Predicting Gastrointestinal Bleeding Events	Text: Electronic Health Record	1-, 3-, and 5-layer neural networks	EHRs ofTaichung Veterans General Hospital
[30]	Early diagnosis of AD	R-fMRI image data	Auto-encoder network	ADNI database

types of activities in daily living and eight types of falls), Global Positioning System (GPS), orientation, light, and sound level. A deep learning framework was designed using auto-labeling module and long-short-term memory (LSTM)-based classification module. The auto-labeling module is based on Q-network (DQN) and use a distance-based reward rule to overcome inadequately labeled sample to improve efficiency. The LSTM-based classification module fuses the multimodal data to manage sequential motion data to detect movement patterns. The authors used Receiver Operating Characteristic (ROC) curve as a performance metric, thus finding an average value for all activities of 0.95.

Image data have been used in healthcare applications for cancer detection, Alzheimer's disease (AD) diagnosis, and COVID-19 classification. Image processing techniques enhance diseases analysis, diagnosis and treatments. For example, a recognition system based on transfer learning and CNN was proposed in [10] to identify early detection of lung cancer. They classify benign tumor (non-cancerous) and malignant (cancerous) such as small cell lung cancer, adenocarcinoma, squamous cell cancer, large cell carcinoma, and undifferentiated non-small cell lung cancer. The used dataset was collected in K1 hospital located in Kirkuk city, Iraq. Experimental results showed an accuracy of 93.33%.

AD is a common neurodegenerative brain disease in elderly people. The statistics shows that in developed countries, 5% after 65 years old and 30% after 85 years old are prevalent to suffer this disease. It is estimated by 2050 that 0.64 billion people in the world will undergo AD. Nowadays, AD diagnosis is invasive, painful and dangerous, thus imposing the need to study non-invasive methods [33]. Deep learning techniques are used in the study of AD. High-level abstract features can be extracted from MRI *images*, which describe data distribution in low dimension. In [22], the authors propose a deep convolutional auto encoder (CAE) architecture to do an automatic non-linear decomposition in large datasets. Features extracted using this technique are highly correlated with other clinical and neuropsychological variables. The affected regions by disease progression as well as the relation with cognitive decline are detected, using visualization of influence areas of each neuron and correlations with clinical variables. CAE architecture can also be applied in diagnosis of dementia. The authors fuse information of neuropsychological test outcomes, diagnoses, and clinical data with the imaging features extracted from MRI. Hence, they found associations between cognitive symptoms and the neurodegeneration process. Regression and classification techniques are used to analyze the distribution of extracted features in different combinations. Then, they estimate the influence of each coordinate of the auto encoder over the brain, thus achieving an accuracy of 80% to diagnose AD. Another approach consisting in multi-projection fusion with CNN to diagnose AD is proposed in [33]. The authors use different projection of the brain such as sagittal, coronal, and axial, instead of using the whole brain volume. They focus on the hippocampal region, which is a stronger biomarker of AD. By binary classification, patients with AD, Mild Cognitive Impairment (MCI), and Normal Control subject (NC) are differentiated. To this end, the most discriminative projection is identified from MRI data using Alzheimer's Disease Neuroimaging Initiative (ADNI) database. After, a fusion framework with CNNs is proposed. Two different fusion strategies were implemented. The first one consists in training three CNNs, thus concatenating features in a fully connected layer. The second strategy applies algebraic late fusion, where the winner is the one with majority vote. Experimental results reported an accuracy of 91%. Other solution for early detection of AD is proposed in [73]. A resting-state fMRI-based framework is developed using DNN based on stacked auto-encoders (i.e., remove high-dynamic data) and medical data. A network focused on each resting-state operations are built based on the connectivity of the brain network in some brain regions, thus measuring the loss of brain function between AD and healthy patients. Data analysis and interpretation is done using deep learning techniques. fMRI *images* and *texts* such as age, sex and genetics are used for training and data classification, using the ADNI database. Functional intellectual networks are built based on time series R-fMRI signal correlation. Experimental results showed an accuracy increase by 25% compared to traditional approaches for AD.

A supervised deep learning strategy using chest CT *images* for both diagnosis and prognosis of COVID-19 patients is proposed in [37]. This approach can distinguish between COVID-19 from non-COVID-19 cases, thus minimizing the manual labeling. CT is a non-invasive technique that detect bilateral patchy shadows or ground glass opacity, which are common manifested symptoms in COVID-19 infected lung [74, 75]. The framework detects COVID-19 infected regions automatically, using chest CT data taken from multiple scanners. Based on the CT radiological features, it can classify COVID-19 cases from pneumonia and non-pneumonia. As dataset, the authors used 150 3D volumetric chest CT exams of COVID-19, community acquired pneumonia (CAP) and non-pneumonia (NP) patients, respectively. Between September 2016 and March 2020, 450 patient scans were acquired from participating hospitals. The authors leverage the representation learning on multiple feature levels to identify the exact position of the lesion caused by the virus. The high-level representation (i.e., Conv5) takes patch-like lesions but discard small local lesions. Hence, the mid-level representation (i.e., Conv4 and Conv5) complements the high-level, by detecting infections located in peripheral lungs, inferior lobe of the lungs, and in the posterior segment. The framework allows giving advice on patient severity to guide triage and treatment. Experimental

results show the proposed framework obtained high accuracy, precision, AUC, and qualitative visualization of the lesion detections. Another approach for classifying COVID-19 using chest X-rays and computerized tomography images of the lungs was proposed in [76]. The authors employed *image* features, feed-forward, and CNNs. They used the texture features technique [77] to extract the best features from X-ray images. That technique takes groups of pixels of the same intensity, thus quantifying and localizing them. Experimental results showed the CNN obtained an accuracy of 83.02% with an AUC of 0.907. Other experiment with feed-forward neural network showed a 100% of accuracy. Other solution using chest x-ray images was proposed in [78]. The authors developed a CNN model with end-to-end training to classify between COVID-19 and healthy patients. They used pre-trained CNN networks such as ResNet18, ResNet50, ResNet101, VGG16, and VGG19 for deep feature extraction. Then, the classification is done using support vector machine (SVM). Experimental results showed an end-to-end training accuracy of the developed model of 91.6%. A different approach using chest CT (CCT) images was proposed in [79]. The authors used pre-trained models (i.e., AlexNet, DenseNet201, ResNet50, ResNet101, VGG16, and VGG19) for feature learning and proposed a new model to extract features (i.e., transfer feature learning algorithm). These models explain the removal of layers by testing pre-trained networks with different configurations. Then, they fuse the two models using discriminant correlation analysis, thus finding a model named CCSHNet. They use datasets from local hospitals with 284 COVID-19 images. Experimental results showed a F1 score of 97.04%. A deep meta-learning framework based on Siamese neural networks to diagnose coronavirus infections from chest X-Ray images was proposed in [80]. The framework uses a fine-tuned base CNN encoder to extract features from input images and applies contrastive loss function to train the Siamese network for n-shot classification of new images. The evaluation results showed the proposed framework attained state-of-the-art performance even with a limited size of the dataset. A framework to combat COVID-19 was proposed in [81]. In the framework, there were two systems, one for the screening of COVID-19 using chest CT scan videos and the other for the mass surveillance. For the screening, three CNN models in the form of ResNet50, Inception V3, and deep tree, were investigated. For the mass surveillance, modules of mask detection, body temperature detection, and social distance measurement were used. Edge and cloud computing was integrated into the framework and the systems performed well with low latency and less waiting time.

A mobile-system application that use an input *image* of food to estimate food attributes such as ingredients and nutritional values is proposed in [82]. It is a real-time system for analyzing nutritional content in the images to recommend patient's diet and their healthy habits. They estimate attributes (e.g., protein, calcium, or vitamins) and ingredients by extracting related words (*text*) over the internet. The system consists of CNNs to recognize the food item in an image and a component to estimate food attributes using text retrieval from internet (i.e., vector space embedding [83]). Transfer learning (i.e., CNN models such as Inception-v3 [84], Inception-v4 [85]) was used for training, using publicly available datasets such as Food-101 and Image-net. The authors removed the last fully connected layer and joined with dropout, ReLU activations, and softmax layers. A two-layer neural network is used for training to calculate probabilities of ingredients in particular food items. Experimental results showed an accuracy about 85%.

Audio data can be used to detect voice pathology in a non-invasive way, where an early voice pathology detection can reduce a permanent voice problem. To this end, a mobile multimedia healthcare framework is designed in [6]. Voice data is recorded using mobile devices (e.g., microphone, a smart phone, or any voice recorder) and processed to be fed into a CNN. The voice signal from the patient is the input of the proposed system and the output has two classes corresponding to pathological or normal voice. The voice signals are 1-s long, which are divided in 40 ms overlapping frames to capture pitch periods and voice breaks. The framed signal is converted to a frequency-domain representation using Fast Fourier Transform (FFT). A spectrogram is obtained when all the frequency-domain representations of all frames are concatenated. Hence, the spectrograms are processed as images and then filtered. After, first and second-order derivatives are applied to the images, thus obtaining three inputs to the CNNs, namely, octave spectrogram, and first and second-order derivatives. Transfer learning using the existing robust CNNs such as VGG-16 and CaffeNet is employed. In the proposed system, the authors replaced the softmax layer of the CNN models by another softmax layer with two classes such as pathological and normal voice. A SVM is used before the softmax layer, where the last fully connected layer was fed. The authors used the Saarbrucken voice disorder (SVD) database and obtained a voice-pathology detection accuracy of 98.77%. The same authors in [7], for voice pathology detection, use the spectrograms with three CNN pre-trained models in parallel to exploit the temporal aspect, which outputs are fused in a Multilayer Perceptron (MPL) with three fully connected layers followed by an output layer. The first layer contains 4096 neurons each and the third one contains 2048 neurons. In this work, the authors use the pre-trained AlexNet model for the three CNN models and showed experimental results with an accuracy of 95.5%.

Research advances in multimodal data and automated *video* have made possible to process supervision tasks. Action classification from multimodal data is a burgeoning research area [86]. Wearable cameras such as Microsoft

HoloLens, Google Glass, and Taser body allow monitoring user's activities. The behavior and actions of the user can be described with wearable video acquisition devices as well as other variety of wearable sensors (e.g., Apple and Samsung smart watches have integrated accelerometers, gyroscopes, and compasses). An algorithm for activity recognition and egocentric human action to assist a user in a medical procedure is proposed in [87]. The authors proposed a supervised deep multimodal fusion framework that process motion data from wearable sensors and *video data* from egocentric or body-mounted camera. They use high temporal dependencies across time-varying sequences for all data modalities. An early fusion is done before assigning class labels to minimize correlation between features and captures temporal sequence behaviors in each data modality. A CNN is used for each individually optimized unimodal representation, which are temporally fused with a LSTM. The multimodal Insulin Self-Injection (ISI) dataset [88] was used in the experiments. The dataset includes action related to an insulin self-injection activity recorded using video data acquired with a Google Glass wearable camera and motion data with an InvenSense motion wrist sensor. Experimental results show that the multimodal fusion approach outperform approaches that rely on one only modality. On the other hand, physically impaired people or elderly people can be monitored remotely by human action monitoring techniques in intelligent healthcare applications [28]. An approach to extract spatial–temporal features using *video* streaming for action recognition is proposed in [20]. The authors proposed a deep learning architecture, 3D convolutional neural network (R3D) to obtain short-term spatial–temporal features and then aggregates the 3D convolutional network entries as inputs to the LSTM architecture to capture long-range temporal information. LSTM output features represent high-level abstraction of the human actions. Experimental results showed an accuracy of 86.8%, which is a good performance of the proposed approach that can be used for remote monitoring since the extracted features capture the motion in the video and recognize the actions. An approach for human fall detection based on *videos* is proposed in [89]. The authors proposed a deep-learning-based approach, using short-term memory neural network. The deep neural network (i.e., RNN) extracts the most important features based on training data, thus covering real-life scenarios. RNN is selected since there is sequential information in the states of the body skeleton at different time steps. LSTM is used for the long sequence of frames. Transfer learning is used, and the approach is trained and tested on "NTU RGB + D Action Recognition Dataset" [90]. According to depth map sequences, the proposed approach detects if there is a falling incident and send an alarm to medical staff, and patient's family and friends. Experimental results show the proposed approach outperforms existing methods based on handcrafted features, thus showing an area under the ROC curve of 0.99.

DNN with multimodal fusion is used in [41] to derive analytics for 1-year gastrointestinal (GI) bleeding hospitalizations prediction using a large in-hospital EHR database (i.e., *text data*) with three different modalities such as disease diagnoses, medications usage, and laboratory testing measurements, which are complementary to each other. The patients have been treated with anticoagulants or antiplatelet drugs. The authors compare neural networks to random forest, gradient boosting decision tree, and logistic regression, thus finding a better performance with neural networks in both single modal and multimodal condition. They found DNNs leverage the EHRs correlations among features in different modalities. They also demonstrate the deep multimodal neural network with early fusion obtain the best GI bleeding predictive power with an area under the receiver operator curve of 0.876.

A multimedia approach for early diagnosis of AD and clinical decision support system is proposed in [30]. The authors use functional connectivity of brain regions using R-fMRI *image* data and clinical *text* information such as age, gender, and genetic information. They compare different machine learning models such as Linear Discriminant Analysis (LDA), Linear Regression (LR), SVM, and auto-encoder network, where R-fMRI time series data and their correlation coefficient data are the inputs. It was found the auto-encoder performs better. The method extracts reliable discriminative brain network features and physical condition in different scales to detect AD. An early stage of AD is distinguished form normal aging with a targeted deep auto-encoder network. The authors construct a customized auto-encoder architecture to classify mild cognitive Impairment (MCIs) and use the ADNI database. That database contains MRI, fMRI, PET, genetic data, and clinical examinations. Experimental results show the comparison with traditional classifiers based on R-fMRI time series data, so making an improvement of 31.21% of prediction accuracy with the proposed method.

The works mentioned above show several research efforts in image analysis (i.e., monomedia applications) using CNNs. Others, in turn, show the combination of two or three modalities (i.e., multimedia) for healthcare solutions. However, there is a need for further research on the use of multimedia data in healthcare applications that covers images, videos, audios, and text data. To accomplish those multimedia healthcare applications, there are some challenges that need to be addressed as mentioned in the section to follow.

5 Challenges

There is a high demand of healthcare services that need to be in real time, accurate, and remote, to provide smart healthcare monitoring to citizens. To this end, citizens' signals (e.g., physiological signals, voice, ECG, video, to name a few) need to be collected, transmitted and processed for assessment. In medical assessment, for example, there is a manual evaluation where the doctor has to be physically and the outcomes depend on the doctor, there are invasive sensors and the medical devices are expensive. On the other hand, the automatic assessment requires an accurate algorithm to predict, diagnose, or treat, as well as it needs low errors, and high response, where its outcome must be validated with a doctor [6]. However, the implementation of automatic assessment healthcare systems that handle several multimedia data have several challenges that need to be addressed as mentioned to follow.

5.1 Computation requirements

Computation efficiency is one of the major challenges in multimedia analytics. The emerging deep learning techniques require high computation requirements to allow the development and scalability of those new solutions [32]. They are computationally expensive due to the hardware requirements such as GPU and Random-Access Memory (RAM) [91]. DNN requires several days on powerful CPUs and GPUs clusters for training [29]. Tensorflow [92] is one solution, by allowing numerical computation with data flow graphs on CPUs and GPUs. In addition, cloud computing provides on-demand services for data computing and database storage, thus employing several GPU computations through several nodes. Cloud technology offers cluster-computing framework with support for machine and deep learning, which has built-in standard operation libraries [32]. In addition, researches are using parallel and scalable deep learning models, are building low-power models, and are using deep learning accelerators through field-programmable gate array (FPGA) [29]. FPGAs allow a high degree of parallelism and optimization as well as low-power computations [25]. Parallel computing in multimedia analytics deal with the non-stationary problem. However, scalable computational methods are still a challenge in multimedia applications [27].

5.2 Multimedia deep learning

Multimedia deep learning applications have open challenges that need to be addressed such as unstructured nature, data volume, nonstationary, variety, and decision-making in real time [27]. There is a heterogeneity gap in multimedia data, where different sources are involved such as mobile devices, social media, sensors, cameras, and virtual worlds, to name a few. They require an efficient multimedia data management, merge different data modalities, and understand DNN decision-making process [29]. All the information from the different data modalities must be analyzed and the logical connections among them must be extracted. Development models can take decisions that are more accurate using the information from the different data modalities. Hence, the importance of fusion techniques to integrate those modalities and find the best configuration in terms of correlations, redundancy, and complementarity among the different modalities [29]. Multimodal data fusion can be made in data space (i.e., raw data fusion or early data fusion) as well as in feature space (i.e., feature fusion or intermediate data fusion) [51]. Multimodal heterogeneous features can be projected into a common subspace, where vectors can represent data with similar characteristics [93]. In this direction, there is ongoing research where transfer learning is being applied. The knowledge of one modality can be transferred to another one, named cross-modal transfer learning [50]. Recent studies have shown that transfer learning has performed well when the learned knowledge from images is transferred to videos [94]. Therefore, treatment heterogeneous, high-dimensional, and unlabeled multimedia data is a future research work in deep learning applications [25], where new techniques and technologies are required to analyze the changing nature of multimedia data [27].

5.3 Big data

The use of deep multimedia learning methods has been growing, thus imposing the need of more training data to increase accuracy [50]. Deep learning methods are highly dependent of large-scale datasets with high quality, where their generation are time-consuming. Label data is required, which incurs in high cost. In addition, it is laborious to create a dataset sufficiently large and diverse from several sources [26], as well as large amount of storage for analytics is needed [27]. As a solution, transfer learning, deep reinforcement learning, and variational auto encoders have been proposed [29]. Transfer learning allow transferring general knowledge from on domain with a large dataset to another domain with no enough data [95]. For example, pre-trained CNN networks such as VGGNet [56], ResNet [96] can be used for image-feature extracting, and word2vec [83], Glove [97] for feature-word embedding. However, for audio and video modality, there is not an effective transfer learning strategy due to limited number of training data. Hence, effective big data techniques are still required in the multimedia learning area to handle big,

unstructured, heterogeneous, and noisy data [25, 50]. On the other hand, even though healthcare applications are emerging in recent days, there is still few healthcare datasets available for clinical studies due to patient privacy. Most of the datasets are in some hospitals and institutions, which can be noisy, with errors, and with missing values, thus making hard to obtain useful datasets for healthcare applications research [40].

5.4 Interpretability and trust

Features discovered by DNNs are complicated to be understood by humans. This is why there is an increasing effort to design interpretable DNNs [43]. Generalization is another big challenge in the medical field [29]. Moreover, many patients do not feel confident about automated systems' outcomes. Hence, the trust in those outcomes relies in the interpretation of specialists and physicians [6]. For example, in the image-medical application field, the diagnosis still requires highly trained human experts. Manually designed features are needed in computer vision approaches. However, this is a challenging task since that not only requires interpretation of visual indicators and abnormalities (this can take years of training), but also to identify complementary and conflicting information from all multimedia data to apply fusion. This is why deep learning can overcome this situation due to it has the ability to learn multimedia features from examples [43].

5.5 Smart city healthcare environment

Healthcare multimedia applications are producing many real-time data that come from Internet of Things devices and sensors. Hence, a smart healthcare framework that provides high quality, high accuracy, and low-cost services is needed [44]. In the medical field, there are many applications focused on classification tasks to diagnose some kind of diseases. Therefore, healthcare applications should be a dynamic process, where not only diagnosis applications are involved [9]. They must provide support for examination, diagnosis, surgery using all the available information to assist medical personnel. Even though, DNN-based systems are automatic, medical personnel are need to interpret and act the results [98]. Despite of success of deep learning in the medical field, there is still a prevention to implement these kind of solutions in the real world.

5.6 Multimedia sensors

Low-cost wearable devices can measure different multimedia data from different modalities (e.g., audio, video, images, and physiological signals, to name a few). However, low-cost wearable sensors can have the problem that can be considered intrusive and has high noise to signal ratio. Video cameras, in turn, are considered intrusive and generate privacy issues [89]. Sensor data is collected from different kind of sensors such as wearable devices, light, GPS, audio, and video, to name a few. Features that are more complex can be extracted to analyze a specific application from a multimodal view. Thus, a synchronization strategy is required to integrate all those sensors, using a unified data fusion strategy. However, synchronization of video and audio in multimedia analytics is challenging [21].

5.7 Digital twin application

Digital twins are now thought as a digital copy (i.e., virtual representation) of a living entity (e.g., a user), which communicates with that living entity and creates a virtual world. It monitors the living entity to understand and give some feedback to improve the user's quality of life and wellbeing. That monitoring consists of collecting physiological, physical, and context information about the living entity. However, digital twin's evolution requires the development of machine and deep learning algorithms that use multimedia data to characterize a user's condition, detect patterns, give suggestions, and make predictions and recommendations. In addition, accurate data collection and signal processing methods are needed, as well as the implementation of effective communication protocols. On the other hand, proposed digital twin solutions must be accurate, thus gaining users' trust and confidence and have to guarantee security issues to protect user's privacy [99].

6 Conclusion

Increasing multimedia data usage (i.e., text, audio, video, and images) in the last years has allowed the development of new research solutions in healthcare applications. Those applications need to address a series of challenges to do an automatic assessment about citizens' health condition. In this review, we show a summary of proposed healthcare applications that use deep learning techniques and multimedia data. We found that few solutions implement different type of multimedia data, thus being mostly monomedia, which highlights the need of research efforts in the

construction of a global solution that can process different type of multimedia data at the same time. We also show the use of deep learning techniques in healthcare applications as well as stand out some challenges that still need to be addressed in the domain of multimedia healthcare applications based on deep learning techniques.

References

1. W. H. Organization, "World Health Organization," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases. [Accessed 4 December 2020].
2. Yach, D., Hawkes, C., Gould, C.L., Hofman, K.J.: The global burden of chronic diseases: overcoming impediments to prevention and control. J. Amer. Med. Assoc. **291**(21), 2616–2622 (2004)
3. W. H. Organization, "World Health Organization," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ageing-and-health. [Accessed 4 December 2020].
4. KashifNaseer, Q., et al.: Self-assessment and deep learning-based coronavirus detection and medical diagnosis systems for healthcare. Multimed Syst (2021). https://doi.org/10.1007/s00530-021-00839-w
5. W. H. Organization, "World Health Organization," [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. [Accessed 8 December 2020].
6. Alhussein, M., Muhammad, G.: Voice pathology detection using deep learning on mobile healthcare framework. IEEE Access **6**, 41034–41041 (2018)
7. Alhussein, M., Muhammad, G.: Automatic voice pathology monitoring using parallel deep models for smart healthcare. IEEE Access **7**, 46474–46479 (2019)
8. Tobón, D.P., Falk, T.H., Maier, M.: Context awareness in WBANs: a survey on medical and non-medical applications. IEEE Wirel. Commun. **20**(4), 30–37 (2013)
9. Dai, Y., Wang, G., Muhammad, K., Liu, S.: A closed-loop healthcare processing approach based on deep reinforcement learning. Multimed. Tools Appl. **81**, 3107–3129 (2022)
10. Anwer, DN., Ozbay, S.: "Lung Cancer Classification and Detection Using Convolutional Neural Networks." Proceedings of the 6th International Conference on Engineering & MIS. (2020)
11. Sieverdes, J.C., Treiber, F., Jenkins, C., Hermayer, K.: Improving diabetes management with mobile health technology. Am. J. Med. Sci. **345**(4), 289–295 (2013)
12. Kirwan, M., Vandelanotte, C., Fenning, A., Duncan, M.J.: Diabetes self-management smartphone application for adults with type 1 diabetes: randomized controlled trial. J. Med. Internet Res. **15**(11), e235 (2013)
13. Maamar, H.R., Boukerche, A., Petriu, E.M.: 3-D streaming supplying partner protocols for mobile collaborative exergaming for health. IEEE Trans. Inf. Technol. Biomed. **16**(6), 1079–1095 (2012)
14. Zhang, Y., Qiu, M., Tsai, C.W., Hassan, M.M., Alamri, A.: Health-CPS: healthcare cyber-physical system assisted by cloud and big data. IEEE Syst. J. **11**(1), 88–95 (2017)
15. Martínez-Pérez, B., de la la TorreDíez, I., López-Coronado, M., Herreros-González, J.: Mobile apps in cardiology: review. JMIR Mhealth Uhealth **1**(2), e15 (2013)
16. Bisio, I., Lavagetto, F., Marchese, M., Sciarrone, A.: A smartphone centric platform for remote health monitoring of heart failure. Int. J. Commun. Syst. **28**(11), 1753–1771 (2014)
17. Fayn, J., Rubel, P.: Toward a personal health society in cardiology. IEEE Trans. Inf. Technol. Biomed. **14**(2), 401–409 (2010)
18. Fontecha, J., Hervás, R., Bravo, J., Navarro, J.F.: A mobile and ubiquitous approach for supporting frailty assessment in elderly people. J. Med. Internet. Res. **15**(9), e197 (2013)
19. Chiarini, G., Ray, P., Akter, S., Masella, C., Ganz, A.: mhealth technologies for chronic diseases and elders: a systematic review. IEEE J. Sel. Areas Commun. **31**(9), 6–18 (2013)
20. Gao, Y., Xiang, X., Xiong, N., Huang, B., Lee, H.J., Alrifai, R., Jiang, X., Fang, Z.: Human action monitoring for healthcare based on deep learning. IEEE Access **6**, 52277–52285 (2018)
21. Zhou, X., Liang, W., Wang, K.I.-K., Wang, H., Yang, L.T., Jin, Q.: Deep-learning-enhanced human activity recognition for internet of healthcare things. IEEE Internet Things J. **7**(7), 6429–6438 (2020)
22. Martinez-Murcia, F.J., Ortiz, A., Gorriz, J.-M., Ramirez, J., Castillo-Barnes, D.: Studying the manifold structure of Alzheimer's disease: a deep learning approach using convolutional autoencoders. IEEE J. Biomed. Health Inf. **24**(1), 17–26 (2020)
23. Wu, C., Luo, C., Xiong, N., Zhang, W., Kim, T.-H.: A greedy deep learning method for medical disease analysis. IEEE Access **6**, 20021–20030 (2018)
24. Dijcks, JP.: "Oracle: Big data for the enterprise," 2012. [Online]. Available: http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf. [Accessed 1 December 2020].
25. Pouyanfar, S., Yang, Y., Chen, S.-C., Shyu, M.-L., Iyengar, S.S.: Multimedia Big data analytics: a survey. ACM Comput. Surv. **51**(1), 1–34 (2018)
26. Halvorsen, P., Riegler, M.A., Schoeffmann, K.: "Medical Multimedia Systems and Applications." 27th ACM International Conference on Multimedia. (2019)
27. Hiriyannaiah, S., Akanksh, B.S., Koushik, A.S., Siddesh, G.M., Srinivasa, K.G.: "Deep learning for multimedia data in IoT." Multimed. Big Data Comput. IoT Appl. pp. 101–129. (2019)
28. Gumaei, A., Hassan, M.M., Alelaiwi, A., Alsalman, H.: A hybrid deep learning model for human activity recognition using multimodal body sensing data. IEEE Access **7**, 99152–99160 (2019)
29. Chen, S.-C.: Multimedia deep learning. IEEE Multimed **26**(1), 5–7 (2019)
30. Ju, R., Hu, C., Zhou, P., Li, Q.: Early diagnosis of Alzheimer's disease based on resting-state brain networks and deep learning. IEEE/ACM Trans. Comput. Biol. Bioinf. **16**(1), 244–257 (2017)
31. Muhammed, T., Mehmood, R., Albeshri, A., Katib, I.: UbeHealth: a personalized ubiquitous cloud and edge-enabled networked healthcare system for smart cities. IEEE Access **6**, 32258–32285 (2018)
32. Sierra-Sosa, D., Garcia-Zapirain, B., Castillo, C., Oleagordia, I., Nuño-Solinis, R., Urtaran-Laresgoiti, M., Elmaghraby, A.: Scalable healthcare assessment for diabetic patients using deep learning on multiple GPUs. IEEE Trans. Ind. Inf. **15**(10), 5682–5689 (2019)
33. Aderghal, K., Benois-Pineau, J., Afdel, K., Gwenaëlle, C.: "FuseMe: classification of sMRI images by fusion of Deep CNNs in 2D+ε projections." 15th International Workshop on Content-Based Multimedia Indexing. (2017).
34. Shan, F., Gao, Y., Wang, J., Shi, W., Shi N., Han, M., et. al., "Lung infection quantification of COVID-19 in CT images with deep learning." arXiv:2003.04655. (2020).
35. Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., et al.: Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology **296**(2), 65–67 (2020)
36. Song, Y., Zheng, S., Li, L., Zhang, X., Zhang, X., Huang, Z.: Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. IEEE/ACM Trans. Comput. Biol. and Bioinf. **18**(6), 2775–2780 (2020)

37. Hu, S., Gao, Y., Niu, Z., Jiang, Y., Li, L., Xiao, X., Wang, M., Fang, E.F., Ye, H.: Weakly supervised deep learning for COVID-19 infection detection and classification from CT images. IEEE Access **8**, 118869–118883 (2020)
38. Shankar K, Eswaran P, Prayag T, et al.: "Deep learning and evolutionary intelligence with fusion-based feature extraction for detection of COVID-19 from chest X-ray images." Multimedia Systems. (2021)
39. Yazhini, K., Loganathan, D.: "A state of art approaches on deep learning models in healthcare: an application perspective." 3rd International Conference on Trends in Electronics and Informatics (ICOEI), India. (2019)
40. Yu, Y., Li, M., Liu, L., Li, Y., Wang, J.: Clinical big data and deep learning: applications, challenges, and future outlooks. Big Data Min. Anal. **2**(4), 288–305 (2019)
41. Hung, C.Y., Lin, C.H., Chang, C.S., Li, J.L., Lee, C.C.: "Predicting gastrointestinal bleeding events from multimodal in-hospital electronic health records using deep fusion networks." 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Germany. (2019)
42. Ravi, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., Yang, G.Z.: Deep learning for health informatics. IEEE Biomed. Health Inf. **21**(1), 4–21 (2017)
43. Ramachandram, D., Taylor, G.W.: Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process. Mag. **34**(6), 96–108 (2017)
44. Amin, S.U., Hossain, M.S., Muhammad, G., Alhussein, M., Rahman, M.A.: Cognitive smart healthcare for pathology detection and monitoring. IEEE Access **7**, 10745–10753 (2019)
45. LeCun Y., and Bengio, Y.: Convolutional networks for images, speech, and time series, in Handbook of Brain Theory and Neural Networks, USA: M. A. Arbib, ed. Cambridge, MA. (1995)
46. Li, M., Fei, Z., Zeng, M., Wu, F.-X., Li, Y., Pan, Y., Wang, J.: Automated ICD-9 coding via a deep learning approach. IEEE/ACM Trans. Comput. Biol. Bioinf. **16**(4), 1193–1202 (2019)
47. Yin, W., Yang, X., Zhang, L., Oki, E.: ECG monitoring system integrated with IR-UWB radar based on CNN. IEEE Access **4**, 6344–6351 (2016)
48. Lu, L., Harrison, A.P.: Deep medical image computing in preventive and precision medicine. IEEE Multimedia **25**(3), 109–113 (2018)
49. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: "Densely connected convolutional networks." 2017 IEEE Conf. Computer Vision and Pattern Recognition (CVPR), USA, (2017)
50. Guo, W., Wang, J., Wang, S.: Deep multimodal representation learning: a survey. IEEE Access **7**, 63373–63394 (2019)
51. Zhang, S.F., Zhai, J.H., Xie, B.J., Zhan Y., Wang, X.: "Multimodal representation learning: advances, trends and challenges." International Conference on Machine Learning and Cybernetics (ICMLC), Japan. (2019)
52. Eyben, F., Wöllmer, M., Schuller, B.: "Opensmile: the Munich versatile and fast open-source audio feature extractor." 18th ACM Int. Conf. Multimedia. (2010).
53. El-Sawy, A., Bakry, H.E., Loey, M.: "CNN for handwritten Arabic digits recognition based on LeNet-5." International Conference on Advanced Intelligent Systems and Informatics. (2016)
54. Minhas, R.A., Javed, A., Irtaza, A., et al.: Shot classification of field sports videos using AlexNet convolutional neural network. Appl. Sci. **9**(3), 483 (2019)
55. Balagourouchetty, L., Pragatheeswaran, J.K., Pottakkat, B., Ramkumar, G.: GoogLeNet-based ensemble FCNet classifier for focal liver lesion diagnosis. IEEE J. Biomed. Health Inf. **24**(6), 1686–1694 (2020)
56. Simonyan K., Zisserman, A.: "Very deep convolutional networks for large-scale image recognition." Computer Vision and Pattern Recognition. (2016)
57. Lu, Z., Jiang, X., Kot, A.: Deep coupled resnet for low-resolution face recognition. IEEE Signal Process. Lett. **25**(4), 526–530 (2018)
58. Yang, M., Zhang, L., Feng, X., Zhang, D., "Fisher discrimination dictionary learning for sparse representation." International Conference on Computer Vision, Spain. (2011)
59. Baltrušaitis, T., Robinson, P., Morency, L.P., "OpenFace: an open source facial behavior analysis toolkit." IEEE Winter Conference on Applications of Computer Vision (WACV). (2016)
60. Burlina, P., Freund, D.E., Joshi, N., Wolfson, Y., Bressler, N.M., "Detection of age-related macular degeneration via deep learning." IEEE 13th International Symposium on Biomedical Imaging (ISBI), Prague. (2016)
61. Liu, J., Pan, Y., Li, M., Chen, Z., Tang, L., Lu, C., Wang, J.: Applications of deep learning to MRI images: a survey. Big Data Min. Anal. **1**(1), 1–18 (2018)
62. Hu, P., Wu, F., Peng, J., Bao, Y., Chen, F., Kong, D.: Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. Int. J. Comput. Assist Radiol. Surg. **12**(3), 399–411 (2017)
63. Bar, Y., Diamant, I., Wolf L., Greenspan, H.:"Deep learning with non-medical training used for chest pathology identification." Medical Imaging: Computer-Aided Diagnosis. (2015)
64. Che, D., Safran, M., Peng, Z.: "From Big data to big data mining: challenges, issues, and opportunities." International Conference on Database Systems for Advanced Applications. (2013)
65. Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. Int. J. Inf. Manag. **35**(2), 133–144 (2015)
66. Ye, Z., Tafti, A.P., He, K.Y., Wang, K., He, M.M.: Sparktext: biomedical text mining on big data framework. PLoS ONE **11**(9), e0162721 (2016)
67. Leibetseder, A., Petscharnig, S., Primus M.J., et. Al.: "Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology." 9th ACM Multimedia Systems Conference. (2018)
68. Pogorelov, K., Randel, K.R., de Lange, T., et. al, "Nerthus: a bowel preparation quality video dataset." 8th ACM on Multimedia Systems Conference. (2017)
69. Pogorelov, K., Randel, K.R., Griwodz C., et Al.: "Kvasir: a multi-class image data set for computer aided gastrointestinal disease detection." ACM Multimedia Systems(MMSYS). (2017)
70. Schoeffmann, K., Taschwer, M., Sarny, S., et al., "Cataract-101--video dataset of101 cataract surgeries." ACM International Conference on Multimedia Retrieval (ICMR). (2018)
71. Nguyen, P., Tran, T., Wickramasinghe, N., Venkatesh, S.: Deepr: a convolutional net for medical records. IEEE J. Biomed. Health Inf. **21**(1), 22–30 (2017)
72. Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., Sun, J., "Doctor ai: Predicting clinical events via recurrent neural networks." 1st Mach. Learn. Healthcare Conf. (2016)
73. Guo, H., Zhang, Y.: Resting state fMRI and improved deep learning algorithm for earlier detection of Alzheimer's disease. IEEE Access **8**, 115383–115392 (2020)
74. Huang, C., et al.: Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet **395**(10223), 497–506 (2020)
75. Wang, D., Hu, B., Hu, C., et al.: Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA **323**(11), 1061 (2020)
76. Varela-Santos, S., Melin, P.: A new approach for classifying coronavirus COVID-19 based on its manifestation on chest X-rays using texture features and neural networks. Inf. Sci. **545**, 403–414 (2020)
77. Bankman, I.: Handbook of medical image processing and analysis, San Diego, CA, USA: second ed., Academic Press. (2008)

78. Ismael, A.M., Sengür, A.: Deep learning approaches for COVID-19 detection based on chest X-ray images. Exp Syst. Appl. **164**, 114054 (2020)
79. Wang, S.-H., Nayak, D.R., Guttery, D.S., et al.: COVID-19 classification by CCSHNet with deep fusion using transfer learning and discriminant correlation analysis. Inf. Fusio **68**, 131–148 (2020)
80. Shorfuzzaman, M., and Hossain, M.S.: MetaCOVID: a siamese neural network framework with contrastive loss for n-shot diagnosis of COVID-19 patients. Pattern Recognit. **113**, 107700 (2020)
81. Hossain, M.S., Muhammad, G., Guizani, N.: Explainable AI and mass surveillance system-based healthcare framework to combat COVID-i9 like pandemics. IEEE Netw. **34**(4), 126–132 (2020)
82. Yunus, R., Arif, O., Afzal, H., Amjad, M.F., Abbas, H., Bokhari, H.N., et al.: A framework to estimate the nutritional value of food in real time using deep learning techniques. IEEE Access **7**, 2643–2652 (2018)
83. Mikolov, T., Chen, K., Corrado G., Dean, J., "Efficient estimation of word representations in vector space." Computation and Language. (2013)
84. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., "Rethinking the inception architecture for computer vision." IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
85. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: "Inception-v4, Inception-ResNet and the impact of residual connections on learning." Computer Vision and Pattern Recognition. (2016)
86. Cheng, G., Wan, Y., Saudagar, A.N., Namuduri, K., Buckles, B.P.: "Advances in human action recognition: a survey." Computer Vision and Pattern Recognition. (2015)
87. Bernal, E.A., Yang, X., Li, Q., Kumar, J., Madhvanath, S., Ramesh, P., Bala, R.: Deep temporal multimodal fusion for medical procedure monitoring using wearable sensors. IEEE Trans. Multimed. **20**(1), 107–118 (2018)
88. Kumar, J., Li, Q., Kyal, S., Bernal, E.A., Bala, R.: "On-the-Fly Hand detection training with application in egocentric action recognition." IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (2015)
89. Shojaei-Hashemi, A., Nasiopoulos, P., Little, J.J., Pourazad, M.T., "Video-based human fall detection in smart homes using deep learning." IEEE International Symposium on Circuits and Systems (ISCAS), Italy. (2018)
90. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.:"NTU RGB+D: a large scale dataset for 3d human activity analysis." IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
91. Muhammad, K., Khan, S., Ser, J.D., de Albuquerque, VHC.: "Deep learning for multigrade brain tumor classification in smart healthcare systems: a prospective survey." IEEE Transactions on Neural Networks and Learning Systems. Early Access. pp. 1–16 (2020).
92. Abadi, M.: "TensorFlow: learning functions at scale." 21st ACM SIGPLAN International Conference on Functional. (2016)
93. Rasiwasia, N., Pereira, J.C., Coviello E., et. al: "A new approach to cross-modal multimedia retrieval." 18th ACM international conference on Multimedia. (2010)
94. Zhang, J., Han, Y., Tang, J., Hu, Q., Jiang, J.: Semi-supervised image-to-video adaptation for video action recognition. IEEE Trans. Cybern. **47**(4), 960–973 (2017)
95. Pan, S.J., Yang, Q.: A Survey on Transfer Learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
96. He K., Zhang, X., Ren, S., Sun, J.: "Deep residual learning for image recognition." IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
97. Pennington, J., Socher, R., Manning, C. D.: "GloVe: Global vectors for word representation." Conf. Empirical Methods Natural Lang. Process. (2014).
98. Riegler, M., Lux, M., Griwodz C., et. Al: "Multimedia and medicine: teammates for better disease detection and survival." 24th ACM international conference on Multimedia. (2016)
99. Saddik, A.E.: Digital twins: the convergence of multimedia technologies. IEEE Multimedia **25**(2), 87–92 (2018)