Informatics in Medicine Unlocked 32 (2022) 101017

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Building an artificial intelligence-powered medical image recognition smartphone application: What medical practitioners need to know

Anindya Pradipta Susanto [a,b], Hariyono Winarto [b,c,*], Alessa Fahira [b], Harits Abdurrohman [d], Arief Purnama Muharram [b,d], Ucca Ratulangi Widitha [b], Gilang Edi Warman Efirianti [b], Yehezkiel Alexander Eduard George [b], Kevin Tjoa [b]

[a] *Cluster of Medical Technology, Indonesian Medical Education and Research Institute, Central Jakarta, Indonesia*
[b] *Faculty of Medicine, Universitas Indonesia, Central Jakarta, Indonesia*
[c] *Division of Gynecologic Oncology, Obstetrics and Gynecology Department, Dr. Cipto Mangunkusumo Hospital, Central Jakarta, Indonesia*
[d] *School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung, West Java, Indonesia*

ARTICLE INFO

Keywords:
Artificial intelligence
Deep learning
Medical image recognition
Smartphone applications

ABSTRACT

Emerging technologies powered by artificial intelligence (AI) have sparked hope of achieving better clinical outcomes among patients. One of the trends is the use of medical image recognition systems to screen, diagnose, or stratify risks of diseases. This technology may enhance sensitivity and specificity and thus, improve the accuracy and efficiency of disease diagnosis. Therefore, it is important and beneficial for healthcare providers to understand the basic concepts of AI so that they can develop and provide their own AI-powered technology. The purpose of this literature review is to provide (1) a simplified introduction to AI, (2) a brief review of studies on medical image recognition systems powered by AI, and (3) discuss some challenging aspects in this field. While there are various AI-powered medical image recognition systems, this paper mainly discusses those integrated in smartphone apps. Medical fields that have implemented image recognition models in smartphones include dermatology, ophthalmology, nutrition, neurology, respiratology, hematology, gynecology, and dentistry. Albeit promising, AI technology may raise challenges from the technical and social aspects of its application. Notable technical issues are limited dataset access and small datasets, especially for rare diseases. In a social context, the perspectives of all involved parties (physicians, patients, and engineers) must be considered.

1. Introduction

The 21st century is the time of accelerated growth of information and technology, the era of unprecedented big data, and the age of effortless information acquisition. Many of these changes are driven by the advancement of digital technology in many fields of life, including healthcare, creating the world that we know today. These changes have enabled us to create and develop technologies powered by artificial intelligence (AI), which emerged through the power of big data. The use of AI in clinical settings has sparked hopes of achieving better, faster, and more efficient clinical decision making for patients [1,2]. One of the most broadly studied aspects of AI in clinical settings is the application of the medical image recognition system using a deep learning model.

Deep learning is an algorithm that learns from enormous data by extracting the patterns within the data as features to produce a model. When trained with data such as images, sounds, and texts, deep learning models can potentially perform specific tasks—for example, medical image recognition, when it is trained by collections of certain medical images. The advantages of deep learning algorithms come from the automatic process of feature extraction, which lets the computer learn the data by itself. The trend toward the study of deep learning models for medical image recognition is highly influenced by the significant increase in the availability of medical image data. Moreover, the advances in information technologies have also led to the enormous growth of digital information in healthcare systems—the doubling time of medical knowledge was estimated to have increased from seven years in 1980 to 73 days in 2020 [3,4].

In the beginning, a highly computing resource was needed to train and serve a deep learning model. However, with smartphone software and hardware becoming tremendously powerful, deep learning models

* Corresponding author. Faculty of Medicine, Universitas Indonesia, Central Jakarta, Indonesia.
E-mail address: hariyono.winarto@ui.ac.id (H. Winarto).

https://doi.org/10.1016/j.imu.2022.101017
Received 27 October 2021; Received in revised form 10 July 2022; Accepted 10 July 2022
Available online 7 August 2022

can now also be served in smartphone applications in several ways. This enables the medical image recognition feature to be available on smartphones that are within the reach of anyone who has access to such devices and applications.

Considering that there are currently more than three billion smartphone users worldwide [6,7], the utilization of smartphone platforms to increase public engagement with healthcare systems is a strategy that must not be overlooked. In addition, the coronavirus disease 2019 (COVID-19) pandemic has hastened the emergence of telemedicine [8]. Therefore, the development and integration of AI technology into smartphones may serve as a bridge to the future of healthcare, thereby making it important for healthcare providers to be literate on this subject.

This literature review collected and summarized research in the field of AI technology in medicine, focusing on medical image recognition models integrated into smartphones. The purposes of this literature review are: (1) to provide a simplified introduction to AI; (2) to briefly review studies on AI-powered medical image recognition technology, especially that integrated into smartphone applications; and (3) to discuss some challenging aspects of this field.

1.1. The evolution of smartphones: a glimpse

Smartphones started out as devices merely for calling or messaging people. Then, various other capabilities were integrated into them, such as to take photos and videos, power social networking (e.g., via Facebook, LinkedIn, and Twitter), and stream entertainment (e.g., via Spotify, Netflix, and YouTube). Market demand has forced the development of even more powerful hardware and software in recent years that have tremendously increased the computing power of smartphones and have turned them practically into personal pocket computers and have even made it possible for them to perform as lightweight AI models with the integration into them of various internet of things (IoT) infrastructures such as smart homes and smart cities [9].

1.2. The big picture of deep learning and image recognition

AI is a broad term describing a field in computer science that studies

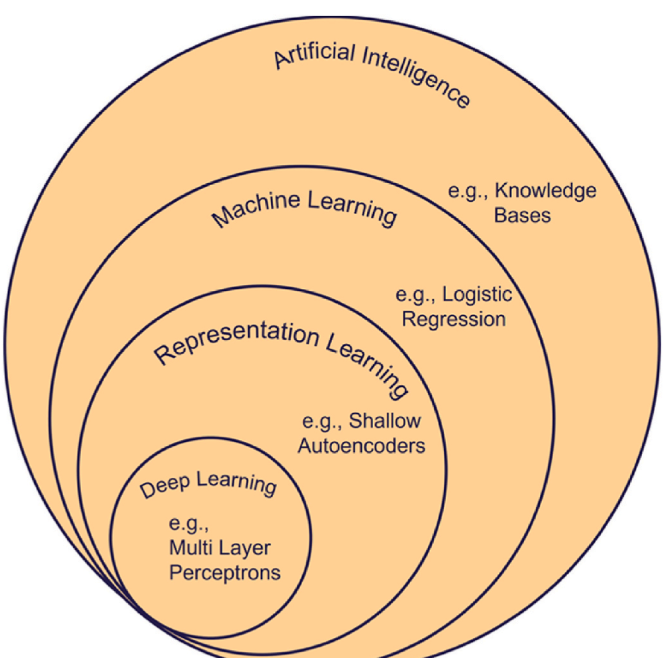

Fig. 1. Venn diagram schematic representation of the hierarchy of terms used in artificial intelligence, machine learning, and deep learning [14].

algorithms to solve problems that usually require human intelligence (Fig. 1). It has many subsets, among which is machine learning (ML). ML is a technique for designing and training algorithms to learn and act upon data and therefore, to achieve an AI model [10,11]. Artificial neural networks (ANNs) are a subset of ML inspired by the connectivity and functions of neurons in the human brain. They consist of an input layer, followed by one or many hidden layers that are interconnected to each other, and an output layer. In each hidden layer, information within the input data is extracted and preserved as a weight. Data enter as inputs through neurons via their dendrites and are processed in the neuron body, and the results are outputted to the axon terminals and passed along to other neurons (Fig. 2). In the early development of ANNs, researchers had to define the features within each class of labeled data (Fig. 3) in a process called *feature extraction*. In ML, features represent unique, invariant traits of an object or phenomenon that can distinguish it from other instances. For example, in images, edges and corners are considered invariant properties that can be treated as features for ML. These handcrafted features will then be identified and analyzed in each input to determine the input's class or label [12].

In 1980, LeCun [13] proposed a novel ANN approach that could extract features from raw inputs called *convolutional layers* using convolution kernels. Convolution is a mathematical operation that produces a new matrix by multiplying and summing up kernels using a sliding window. That is, as the kernel is relatively smaller than the input image, the stated operations are done by a sliding window that is constrained by stride and padding. A stride is a number that indicates how far the sliding window can move, usually in pixel units. Padding is an extended area of an image. The kernels act as weights and represent the abstraction extracted from the input. These convolutional neural networks (CNNs) pioneered deep learning algorithms.

Nowadays, in deep learning, CNNs are widely used for image-based analysis. CNNs consist of stacked convolution functions that extract coarse-to-fine information (Fig. 4). CNNs extract the features from the input data, process them, and then use them to interpret other data with the same features (Fig. 3). CNNs in deep learning have rapidly evolved and now consist of many layers and modifications that have resulted in numerous architectures. VGGNet, DenseNet, and ResNet are well-known algorithms in deep learning [14].

Visual Geometry Group (VGG) is one of several models that adopt CNN in deep learning architecture [15]. Consisting of six CNN blocks, VGG significantly outperformed other CNN models in both the 2012 and 2013 ImageNet Large Scale Visual Recognition Challenge, an annual competition for different mechanisms of automated image classification. However, training the VGG model is challenging, as it requires a longer training time and has a considerably large weight. The challenge may be resolved by ResNet through the introduction of a residual network concept that uses fewer layers and less resources than VGG nets. The main contributions of ResNet are its introduction of residual learning and identity mapping with skip connections or shortcuts [16,17]. On the other hand, DenseNet has the advantages of having fewer parameters and having cross-layer information flow, a mechanism for sharing features with other upper layers [17,18].

2. Results and discussion

2.1. Developing an AI-powered medical image recognition smartphone application: an introduction

2.1.1. Data preparation process in image recognition

In the medical field, image recognition is typically used to aid screening or diagnostic processes. The development of an AI-powered image recognition application begins with data preparation, which includes data collection, cleaning, and labeling (Fig. 5) [19]. Data collection is a process of collecting data from many sources to create a dataset. Data cleaning removes and fixes all incorrect or ambiguous data within the dataset. Data labeling assigns labels to the data. This process

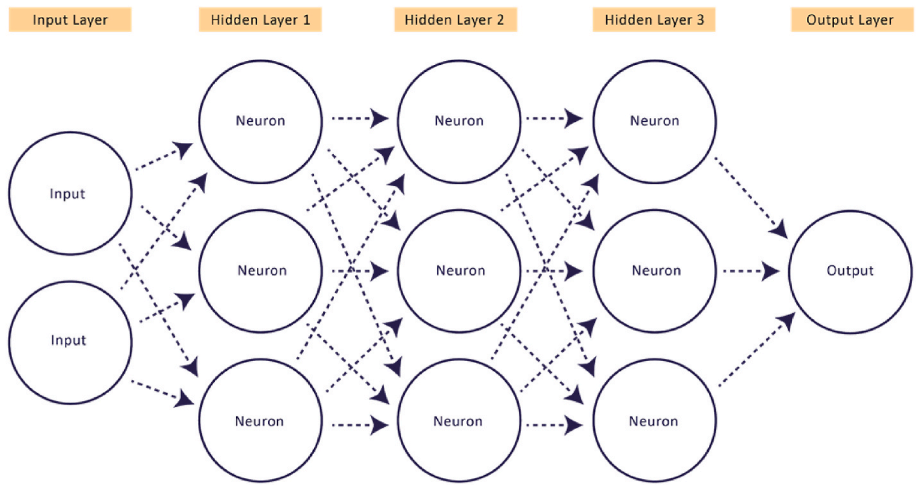

Fig. 2. Schematic representation of a simple artificial neural network with three hidden layers (adopted from Klang) [11].

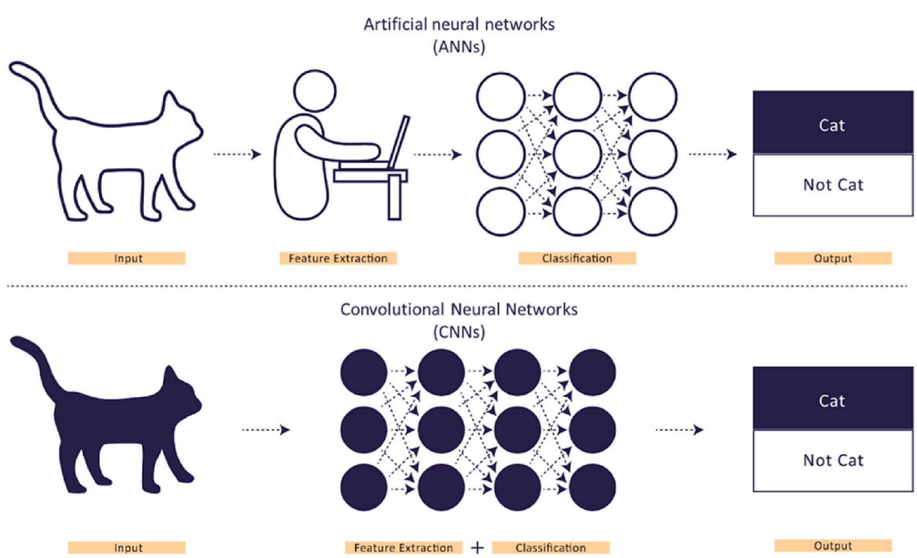

Fig. 3. Illustrations of conventional artificial neural networks compared to the modern architecture of convolutional neural networks.

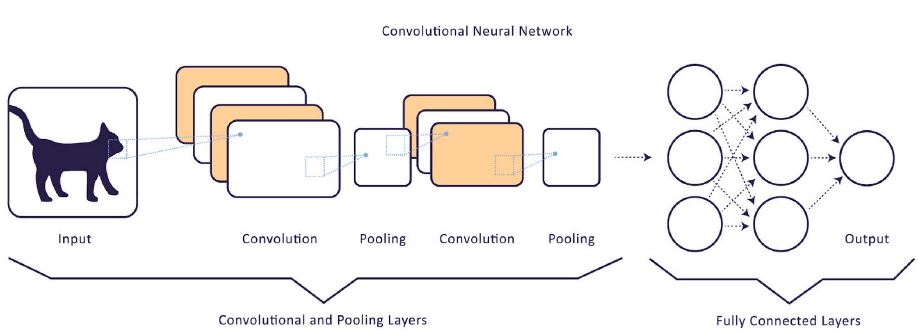

Fig. 4. The convolutional neural network model.

can only be done by experts. The label itself is a tag that provides better information in context. Clear specifications for data preparation should be articulated prior to data collection. Following the data preparation, data preprocessing is performed on all datasets according to statistical method criteria. This involves generalizing all the characteristics of the images. For example, all the images may be cropped and resized to a specific size, or the color channel may be changed from RGB to black and white. After the preprocessing, data augmentation is applied to the dataset to improve the robustness of the model [20,21]. For example, if a class of cervical cancer images that requires 1000 images only has 500 images, data augmentation can be used to rotate the original images by 30°, 60°, 90°, 120°, and 150° and/or to flip the images, thus allowing the model to build a robust feature representation although the number of items in dataset is slightly imbalanced. The images are then fed into the deep learning model [19,21].

Fig. 5. Data preparation process.

2.1.2. Building a smartphone application

Building a smartphone application involves three dimensions of requirements engineering. Requirements engineering is the process of transforming a business concern or problem into information system requirements through a systematic approach. This approach is an iterative process that starts from requirements elicitation, requirements specification, and requirements validation and verification (Fig. 6) [22]. The first step, requirements elicitation, begins with analyzing and understanding the problem. Then, specification requirements are created to define the (technical) specifications of the product to be delivered. The proposed specifications are then carefully validated and verified by all parties involved in the project. Once the product's specifications are established, the developer can start writing the product code [22].

2.1.3. Integrating AI into smartphone applications

Integrating AI functionality and capability into a smartphone application is a current trend. There are two main approaches to integrating AI into smartphone applications: (1) "in-app-based" and (2) Application Programming Interface (API) service-based. An in-app-based model is built within a computer and deployed into the smartphone bundled with the application. The model is then run by the application through the provided software development kit (SDK), for example, TensorFlow Lite in Android and CoreML in iOS. Because the model is bundled and run simultaneously with the application, the result can be instantaneous; however, more complex models may be limited by the capabilities of the SDK and hardware used and may not be able to run.

In contrast, the API service-based approach integrates AI functionality via an API protocol (Fig. 7). This approach allows more complex models to run and has better results. However, the maintenance costs associated with the API service-based approach are higher than those of the in-app-based approach [6,22].

2.2. Current medical image-recognition AI-powered smartphone applications

The use of AI in medical image recognition has been widely studied. Recent advances in the integration of AI in the biomedical field through image processing or smartphone-based imaging devices (SIDs) have happened rapidly, thanks to the development of advanced hardware and software for image processing, including an adjustable camera function or structure (e.g., by adding more lenses or other supporting tools) that enhances image capture. Some SID applications are for routine diagnosis and microbiological detection (e.g., bacterial and viral detection). According to Hunt et al. [23], smartphone-based biomedical imaging applications can be classified into four clinical workflows: *ex vivo* diagnosis, monitoring, in vivo diagnosis, and treatment guidance. To further study and investigate current developments in AI-powered smartphone medical image-recognition applications, we conducted a systematic search of PubMed, Scopus, ScienceDirect, and the Cochrane Central Register of Controlled Trials (CENTRAL) up to December 2020 using the following keywords: "artificial intelligence AND (machine learning OR deep learning) AND (image recognition) AND (smartphone OR smartphone application OR application)." In addition, the reference lists in relevant published articles were searched. Only studies that used smartphone image recognition AI in the medical field and that were written in English were included. Articles that had not yet been published and peer-reviewed, that contained non-original research, and that reported nonmobile applications were excluded. Table 1 shows some of the published studies that have reported on medical image recognition smartphone applications in several disciplines.

Smartphone usage for AI-powered medical image recognition ranges from screening to risk stratification to diagnosis. In most studies, a CNN is the preferred algorithm. Moreover, most smartphone applications are designed for healthcare professionals. While there are many applications in the market, only a few have been validated through meticulous studies. Unfortunately, some applications have been discarded [24].

2.2.1. Dermatology

AI-powered smartphone applications have been introduced in the field of dermatology as self-monitoring screening tools in the evaluation of melanoma, a form of skin cancer that arises from melanin pigment-producing cells [24,25]. Skin lesions are categorized by an image-recognition feature in the application as high-, medium-, and low-risk, followed by the appropriate clinical recommendations. The performance of AI-powered melanoma detection ranges from 73 to 80% sensitivity and 78–83% specificity. Despite having consistently lower sensitivity and specificity compared to dermatologists' diagnosis, melanoma screening tools are useful for early detection and for patients' consideration to seek further evaluation with a dermatologist [24,25]. However, some issues that have arisen in image recognition in dermatology are a small sample size, unstandardized methodological evaluations between studies (different gold standards: histopathology vs. expert assessment), and limited peer-reviewed validated applications

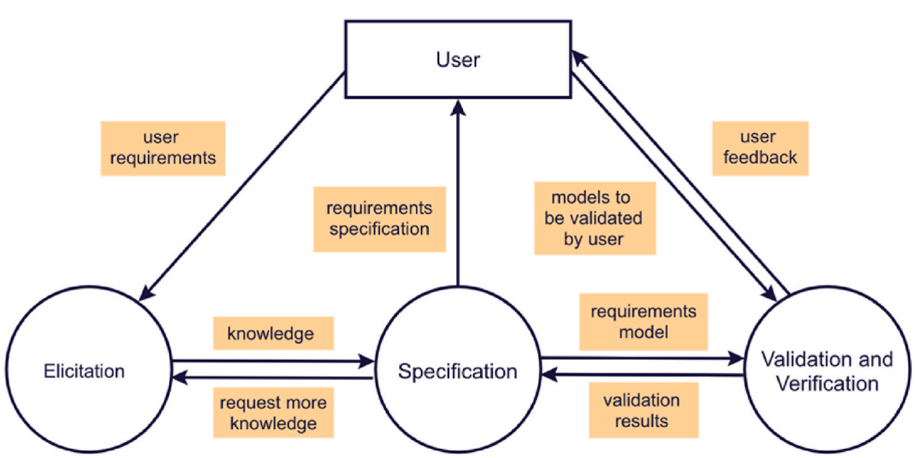

Fig. 6. Requirements engineering process (adopted from Wahono) [22].

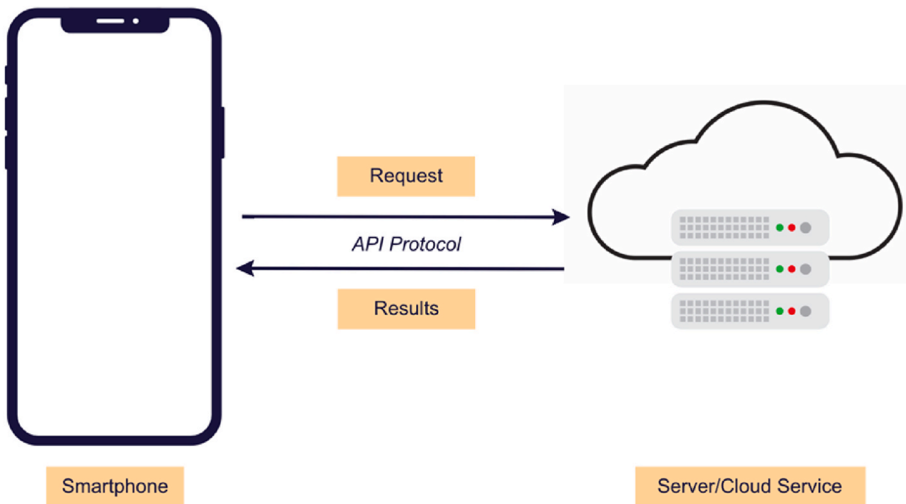

Fig. 7. Integrating artificial intelligence functionality into a smartphone through an Application Programming Interface (API) protocol.

Table 1
Studies that have reported on the use of smartphone-based medical image recognition.

Reference	Discipline	Usage	User	AI algorithm	Gold standard	Sen/Spe (%)	Acc (%)
[1]	Dermatology	Melanoma diagnosis	HP	Box Counting	Histopathology	73/83	81
[2]	Dermatology	Malignant skin lesion diagnosis	HP	CNN	Histopathology	80/78	N/A
[3]	Ophthalmology	Visual anomaly (strabismus, myopia, anisometropia) screening	HP	CNN	Expert	80–83/ 98–100	94–99
[4]	Ophthalmology	Amblyopia risk factors screening	HP	CNN	Expert	88/75	79
[5]	Ophthalmology	Diabetic retinopathy diagnosis	HP	CNN: AlexNet, GoogLeNet, ResNet	Expert	98/99	98
[6]	Ophthalmology	Glaucoma diagnosis	HP	CNN: ResNet	Expert	95/87	99
[7]	Nutrition	Food and drink recognition	HP, patient	CNN: NutriNet, AlexNet, GoogLeNet, ResNet	Expert	N/A	55–87
[8]	Neurology	Determination of the degree of bradykinesia	HP	Bayesian, SVM	Expert	58–78/ 40–73	61–69
[9]	Respirology	Severe pharyngitis diagnosis	HP	CNN: ReNet, Inception, MobileNet	Expert	97/94	95
[10]	Hematology	Malaria parasite in blood smear diagnosis	HP	CNN	Expert	92/94	93
[11]	Hematology	Sickle cell disease screening	HP	CNN	Expert	N/A	98
[12]	Gynecology	Cervical precancer screening	HP	CNN: ResNet, DenseNet, MobileNet	Expert	N/A	96
[13]	Gynecology	Cervical precancer screening	HP	CNN	Expert	91–93/ 78–90	83–91
[14]	Dentistry	Oral cancer lesion screening	HP	HRNet	Expert	83.0/96.6	N/A

CNN: convolutional neural network; ICP: iterative closest point; CLA: clustering algorithm; SVM: support vector machine; Sen/Spe: sensitivity/specificity; Acc: accuracy; HP: healthcare provider.

available for download (only two out of six apps) [24]. More efforts to control disturbing factors such as ulcerated lesions or abnormal surrounding skin are also considered [25]. Although further development is necessary to improve the accuracy of melanoma detection using AI-powered smartphone applications, the study highlighted the potential use of image recognition-based screening in the future for the evaluation of melanocytic nevi.

2.2.2. *Ophthalmology*

Image-recognition applications have been proven useful as children's vision screening solutions for areas with limited resources. Ma et al. [26] utilized a smartphone's flash and camera as light source and image sensor, respectively, following the automated Hirschberg test and photorefractory principles. The examinee's head gestures and eye location were estimated using a deep learning model. Then, the limbus areas, interpupillary distance, corneal luminous reflection spots, and red reflex areas were identified to derive the risk levels of strabismus, myopia, and anisometropia, resulting in a claimed accuracy of 94–99%. It should be considered that the AI model was not validated by an independent examiner, and thus, there is a possibility of overclaiming of accuracy. Moreover, the model relies on generalized limbus detection, which requires a contrasting conjunctiva and cornea border that is apparent in the Chinese samples. Therefore, inaccuracies may occur if the examinee has a lighter iris color or is wearing contact lenses, which means that replication of the study may be harder in Western countries. The study aimed to be able to quickly screen visual anomalies in children in rural schools when no optometrist is present. Unfortunately, the study was tested in urban schools with an optometrist supervising, reflecting a study design that was the opposite of the research target.

Another AI-based smartphone application in the field of ophthalmology was developed to screen for amblyopia risk factors in children using photoscreeners and autorefractors in smartphone cameras. The combination of low and ambient light was needed for amblyopia risk factor (ARF) screening, after which the data were processed with deep learning and image processing models that required only a smartphone app without any customized hardware component, which made the app superior to its competitors. Murali et al. also claimed that the app has a high sensitivity of 88.2% in detecting ARF, deeming it useable for early screening before referral to a tertiary healthcare facility. Despite the seemingly superior outcome of the study's use of AI models, limitations persist in the app. For instance, false negatives occur due to the app's inability to detect ARF in children with a very large or small pupil size as they are undilated in a dimly lit room, whereas the model requires detecting the eye's red reflex to work. False positives also occurred in the study, wherein a larger refractive error was detected when the algorithm failed to differentiate the corneal light reflex that overlapped with the crescent zone, which caused blurred margins to the actual crescent width [27].

Smartphone-captured images combined with deep learning frameworks have been adopted as tools for vision-threatening diabetic retinopathy and glaucoma detection. Hacisoftaoglu et al. [28] proposed a deep learning system to detect diabetic retinopathy and to validate it using publicly available retina image datasets. Despite achieving a high accuracy of 98.6%, the study used five different datasets with images taken without a standardized method, which resulted in inconsistent labeling and different image qualities. Furthermore, the accuracy dropped during the cross-testing between databases. To overcome these problems, Hacisoftaoglu et al. [28] suggested a standardized method of retinal image collection using an additional device to a smartphone's camera. On the other hand, Li et al. [29] collected their own visual field data and achieved an accuracy of 99% in glaucoma detection using a deep learning system. However, as the system is trained solely using visual field data, if the patient presents with functional or structural defects, false negatives may occur.

2.2.3. Nutrition and food science

Daily nutritional assessment has been made easier by AI-powered image-recognition models that can be used to quickly and accurately recognize food items. A model developed by Mezgec et al. [30] called *NutriNet* utilizes deep learning systems and smartphone applications to encourage Parkinson's disease patients to assess their diet. Using a dataset of 130,517 images with a 512 × 512-pixel resolution (54,564 food or drink images and 75,953 images of other objects), they achieved an accuracy of 94.47%. However, when a real-world test was performed using 200 food images self-captured by Parkinson's disease patients, the highest accuracy was 55%. A proposed explanation of this accuracy gap is that the majority of the patients' diets were a Central European region diet, whereas the dataset that was used for the model training was not specific to a certain region's diet. Mezgec et al. proposed a solution— to continually update the deep learning model on a weekly basis [30,31].

2.2.4. Neurology

Also in relation to Parkinson's disease, Williams et al. developed an AI-based smartphone application to determine the presence of bradykinesia (slowness of movement) by analyzing videos of finger-tapping behavior, classified using the Unified Parkinson's Disease Rating Scale [32]. With a dataset of 70 videos (40 of Parkinson's disease patients and 30 of control subjects), prediction of bradykinesia reached an accuracy of 80% and of Parkinson's disease, 67%. The results showed that this method is comparable to conventional examination by human experts, although the approach is still highly affected by the camera position and the lighting conditions. Furthermore, the results of bradykinesia from AI were only binary (UPDRS 0–1 vs. Higher) which is a low classification resolution. As for the Parkinson's disease group, with a small sample size and varying levels of clinical manifestation, the mentioned result did not reach an acceptable level of confidence [32].

2.2.5. Respirology

Another AI-based smartphone application was developed to detect severe pharyngitis using self-taken throat images. Using ResNet50, the highest detection accuracy achieved was 95.3%, which shows that this deep learning model is a promising rapid identification and contactless diagnostic method for pharyngitis in every setting, especially in the COVID-19 pandemic. Unfortunately, the algorithm was built from self-taken pharyngeal images from open social platforms, without metadata, and had a low resolution, which might lead to misdiagnosis. Moreover, the study used experts' opinions rather than bacterial culture, the gold standard for pharyngitis, for the diagnostic comparison. Furthermore, some clinically relevant components such as the Mallampati score (a simple grading system based on visualizing pharyngeal structures) and oral cavity anatomical variations were not included in the algorithm [33].

2.2.6. Hematology

The diagnosis of malaria is still challenging, especially in areas that lack laboratory facilities and parasitologists. Automated detection of malaria-causing parasites in thick blood smears was developed by combining intensity-based iterative global minimum screening and a CNN, which has resulted in a promising alternative to conventional parasite counting for malaria diagnosis. Yang et al. [34] developed an algorithm based on 1819 datasets of thick smear images. Although a high accuracy of 93% was achieved, the system cannot differentiate malaria subspecies. Moreover, the study did not include non-malaria parasite infections that can mimic malaria.

Another deep learning framework by De Haan et al. [35] can perform automatic screening of sickle cell disease (SCD). First, the model standardizes the blood smear image taken with a smartphone to match the image quality of a laboratory microscope. Then, the model distinguishes sickle-shaped cells from healthy red blood cells to determine the SCD status of the patient. The results are promising, with high accuracy for SCD screening, especially in resource-limited settings. However, the small sample size and datasets (96 patients, 32 with SCD) should be considered when interpreting the results. Another critical aspect is the need for image enhancement networks to overcome the aberrations and lower-resolution problems of smear images [35].

2.2.7. Gynecology

In gynecology, cervical cancer continues to be the most common cancer among women of reproductive age. A traditional cervical cancer assessment method involves visual inspection of the cervix with acetic acid to detect the acetowhite lesion associated with the human papillomavirus (HPV). However, this method is highly subjective and dependent on the clinical experience of healthcare workers. To improve this method, a new deep learning approach called *automated visual evaluation* was developed to analyze cervigram images for automatic detection of cervical precancer lesions. Hu et al. showed that the algorithm can be run in 30 s on a smartphone [36]. While it demonstrated 96% accuracy, there is no detailed information on the number of images used from the cervigram dataset. Furthermore, the algorithm can only be presented with high-quality images that show the entire squamocolumnar transformation zone, as the algorithm was not trained with real smartphone images.

A similar approach was developed by Guo et al. [37] using four databases: MobileODT, Kaggle, SEVIA, and COCO2017, with a total of 31,967 images (30,151 cervix images and 1816 noncervix images). Their novel algorithm identifies cervix and noncervix images. The results showed high accuracy with images captured using smartphones and less time for analysis. Although some cervix images were still misclassified as non-cervix and vice versa, this model is important as a method of image acquisition control.

2.2.8. Dentistry

Oral cancer is a common type of cancer that occurs inside the mouth. It has high mortality rates, especially oral squamous cell carcinoma. More than 90% of all oral cancers are squamous cell carcinomas [38]. Early screening is the key to a better survival rate. Current screening is through simple visual detection to search for lesions, but challenges in early detection arise due to the similarity of clinical cancerous lesions to noncancer lesions. Experts in gynecology and oncology are limited, and thus, early cervical cancer detection supported by AI would be beneficial for early detection.

Lin et al. [39] developed an image-recognition with CNN-based system for identifying oral cancer (HRNet). The system was trained with oral datasets divided into five categories: aphthous ulcer, low-risk Oral Potentially Malignant Disorder (OPMD), high-risk OPMD, and oral cancer. HRNet was proven to be superior to VGG16, ResNet50, and DenseNet169, as it achieved a sensitivity of 83.0%, a specificity of 96.6%, and a precision of 84.3%. These results demonstrate that the use of AI models is effective in oral cancer detection.

Therefore, care must be taken in interpreting the high performance of AI-powered image-recognition smartphone applications. Overclaiming of accuracy may be traced to low-quality studies with few samples and small training datasets, which, in addition, might have been obtained not with a smartphone (thus showing unstandardized image quality and acquisition) and omission of relevant clinical aspects.

2.3. Challenges in technical aspects

2.3.1. Artificial neural networks in healthcare are highly specialized and multidisciplinary

The increasing trend toward ANN research in medicine and its seemingly limitless possibilities may provide a false impression, in which the complexities of AI development are overlooked. To put this in perspective, expertise in ANN is comparable to specialization in medicine—it takes considerable experience in data science to comprehend the mechanisms, let alone to implement ANN as an assistive tool in healthcare. Attempting to develop ANN without recognizing the steep requirements may lead to the risk of lower-quality projects with false accuracy, which could result in heavy bias and potential harm to patients [40]. Thus, a collaborative approach to developing ANN in medicine should be adopted, wherein clinical, computational, and legal experts work together in a multidisciplinary team [41].

2.3.2. The dataset size is essential

Although small datasets may be adequate for a small population, training a complex ANN model with a larger dataset is necessary for superior accuracy and scalability [42]. Dataset collection may be more difficult for rare illnesses, diseases with a high variability of manifestations, or diseases that require manual data extraction. To optimize the accuracy of the results of using AI technology, it is beneficial to use multi-institutional data. However, subtle variations in the case definitions used by medical experts from different institutions may cause unintentional issues and have a detrimental impact on the accuracy of the ANN results. This problem can be addressed by agreeing on only one set of clinical definitions and guidelines [1,43].

2.3.3. Apparatus requirements

The requirements for performing the aforementioned research can be generally divided into the following processes, each requiring careful preparation and handling (Fig. 8) [44,45].

1. Data collection: The data can be collected with any device or tool for as long as the data have the same modalities. In our case, we used images as our modality.
2. Data warehousing: All the data sources must be integrated into a single data source. This data warehouse stores images, labels, and annotations. The data labeling must be done by experts.
3. Training: This is the phase where the AI model is created. Besides training, in this phase, validation and testing are also carried out. Thus, the collected data are split into those for training, validation, and testing. We store each best model with the best validation score and test it with the test dataset. If its score is above our threshold, we put it on the best model record and process for the model serving. A mandatory peripheral in this stage is a computer with a high number of processing units such a GPUs or TPUs.
4. Serving and monitoring: Model serving is done with two options. For a low-resource model, the model serving can be embedded into the smartphone device; but for a high-demand resource model, the model serving is done as the API calls from the server. The server might use a specific computer processing unit such as a GPU or a TPU. Model monitoring is done by collecting eligible user data with the user's informed consent in writing. The collected data can be a taken image, the model results, or user input feedback. This feedback could solve several problems such as training–serving skew, model drift, and an underperforming system. Model re-evaluation might be needed periodically.

2.3.4. Challenges in medical image preparation and the machine learning process

The chief obstacles in the development of medical imaging screening or detection lie in the preparation of a medical imaging dataset, which requires the availability of a sufficiently large and representative dataset. The preparation of an adequate dataset may require subjects or samples from large geographic areas and thus, may be costly and time-intensive. Challenges may also arise during each step in the data preparation, as follows [1,43].

- Acquiring the dataset itself may be challenging, as AI algorithm developers are not usually located within the hospital itself and hence, easy and direct access to data may be limited. Thus, good collaboration between clinicians and AI developers is vital.
- Good data management and storage may raise questions about availability and safety. As mentioned, AI algorithm developers may require direct, safe, and fast access to data, which makes cloud-based data storage a preferable option as it improves data sharing, backup, and security. However, it may be more expensive and requires an adequate internet connection. Data management should strive to follow the FAIR (findable, accessible, interoperable, and reusable) principle, as described by Wilkinson et al. [1,43,46].
- Medical data training usually requires a supervised learning approach. Labeling by medical experts (e.g., radiologists and pathologists) will inevitably result in errors and misinterpretations over

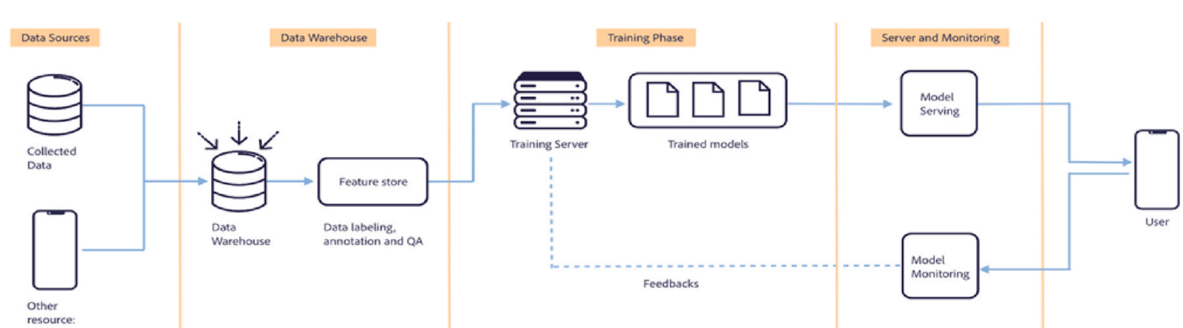

Fig. 8. Apparatus requirements for artificial intelligence model development.

time. Evaluating the impact of a defective sample may be very difficult due to the nature of deep learning, and the feasibility of debugging an ANN with millions of data samples for both the AI developer and the medical expert may be debated [1,43].

2.4. Challenges in social aspects

2.4.1. Artificial neural network in legal terms

The continuous progress of AI in the academic domain, which is currently oblivious to legal concerns, may hit a few legal walls when the technology starts to occupy a larger social space. Legal aspects may concern safety (How valid, reliable, or safe to use are the data employed in the AI?) and effectiveness (Does the software or do the data need regular updates? Who is in charge of those?). Regulations have already been developed in some countries. For example, in the U.S., the Food and Drug Administration regulates AI. Gerke et al. [47] emphasized the need for regulators to also focus on the development of a procedure for continuously monitoring, managing, and identifying risks associated with features linked to the reliability of AI or ML systems. In Europe, AI regulations have been formulated by the European Union (EU) documented in Medical Device Regulation EU 2017/745 and the In Vitro Diagnostic Medical Devices Regulation EU 2017/746 [47]. New rules are embedded in the regulations regarding classes of risks of AI software in aiding decision making for diagnostic or therapeutic purposes. Class I is “low-risk”; Class II, “medium-risk”; Class IIb, “higher-risk”; and Class III, “highest-risk.” [47].

Liability may be a topic that should also be discussed. When an AI-based tool makes an incorrect recommendation that may harm patients, who will be at fault? If AI is regarded as a tool for supporting clinical decisions of clinicians, it may appear that the clinician should be held liable. Other views proposed liability of the AI manufacturer, but such proposal has not yet been adopted considering the higher premise that AI is merely a support tool for decision making. Another topic that may be considered is the implication of compensation of patients harmed by AI-related diagnoses and treatment recommendations, as applied to vaccine manufacturers in the U.S [47].

2.5. Interpretability and explainability

Before an AI-based tool in medicine can be used extensively in clinical settings, clinicians may first need to be able to understand, interpret, and explain the outcomes of AI models, even of very complex models. The obstacles in the universal acceptance of AI models in medical practice may be the lack of precise interpretability of the results. If a model cannot be comprehensively explained both to experts and patients in wide settings, its outcomes and recommendations could not be reasonably trusted [48].

2.5.1. Data safety, privacy, and anonymity

Failure to protect patients’ data from use outside the doctor—patient relationship may deleteriously affect patients in various cases such as in influencing insurance qualification or premiums, job opportunities, and even personal relationships [49]. A pivotal discussion regarding data safety is on informed consent in the field of AI. In what circumstances does informed consent need to be solicited in an AI study? Can we in the future use a ‘broad’ and ‘blanket’ consent statement that pre-authorizes future secondary analysis? However, what about cases of previously collected data in medical records, for which obtaining consent to use years after a trial was conducted can be very difficult, if not impossible? [50] Will it be enough and safe for us to use de-identified data?

2.5.2. Data synthesis

On several occasions, collected datasets do not reach the minimum requirements for AI development [51] and thus, several researchers opt for synthetic data. However, synthetic data are subject to ethical issues on account of the risk of false decisions caused by the lack of information on the generated data. To the best of our knowledge, using data augmentation is still tolerable for as long as it does not remove valuable information [52]. Despite having issues, synthetic data have a potential role in overcoming the lack of available datasets. The ethics of synthetic data usage should be further discussed to develop a better protocol.

2.5.3. Cost of development

Correct data labeling is crucial for achieving an accurate AI model, especially in the medical field. Thus, extensive human resources for data collection and labeling are required in AI model development. This may incur a significant cost due to the need to hire large numbers of related medical experts. Thus, the initiation phase for developing medicine-related AI models can be expensive. A more efficient approach, such as semi-supervised or self-supervised learning, could be a way to optimize the sample number, which may help reduce costs.

2.5.4. Possible resistance to adopting AI technology in clinical settings

Replacing physicians with AI is a big issue—if not the main issue—despite most studies reporting that AI can complement medical practice but not replace it [1]. A survey among the staff of the Royal Free London National Health Service (NHS) Foundation in the U.K. found that only 10% of the healthcare workers who participated were worried about being replaced by AI. Moreover, 79% of the respondents perceived AI as beneficial [53]. A study paradigm that shifts from an “AI vs. human” comparison to a “human with AI vs. human without AI” comparison should be initiated.

Resistance of doctors may arise from their lack of knowledge about emerging technologies, especially AI [1,53]. Moreover, there is a lack of involvement of physicians as experts in technology development. These problems lead to physicians questioning AI results [54]. Furthermore, due to the nature of AI, any conflict between the AI result and doctor’s interpretation will be difficult to explain. This can lead to confusion and aggravate patient’s distrust especially in difficult cases which require decision making [1,54]. At a broader level, implementation of AI in healthcare has been found to be hampered by technical constraints that create an administrative burden in the early phase [1,54].

Resistance by patients is also a factor. The lack of consideration of the unique aspects of an individual patient’s circumstances is a concern in the application of AI in medical practice. Some patients may perceive that the use of AI neglects their individuality, and they may prefer doctors who care rather than merely consider objective performance (i.e., accuracy) [47,55].

Performing more collaborative studies that involve clinicians, programmers, and, of course, patients may be the bridge required to achieve the implementation of novel technologies in medical practice [41]. To reach this goal, each stakeholder group must become more familiar with the motivations and needs of the other groups so that they can better understand and reach agreement with each other.

3. Conclusion

AI-powered medical image recognition smartphone applications are currently being developed for use in many disciplines of medicine. CNN is the most common algorithm being used. Applications are not limited to clinical images but are being adapted to microscopic images for screening, risk assessment, and diagnosis. The rapid global increase in the use of such technologies in medicine requires developers and all other stakeholders to direct their attention toward ensuring that each step and aspect of the ML technologies and methodologies are sufficiently robust, responsive, and relevant for use in medicine. Challenges that have been identified in data preparation, algorithm development, training, testing, validation, and implementation can be overcome with further research and better study design.

Funding details

This work was supported by the *Direktorat Inovasi dan Science Techno Park* (DISTP) [grant number PKS-20/UN.2.INV/HKP.05/2020].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the Division of Gynecologic Oncology, Obstetrics, and Gynecology Department of the Dr. Cipto Mangunkusumo Hospital, the Faculty of Medicine of Universitas Indonesia, and the Cluster of Medical Technology of the Indonesian Medical Education and Research Institute for their support throughout this study.

References

[1] Briganti G, Le Moine O. Artificial intelligence in medicine: today and tomorrow. Front Med 2020 Feb 5;7:27.
[2] null Amisha, Malik P, Pathania M, Rathaur VK. Overview of artificial intelligence in medicine. J Fam Med Prim Care 2019 Jul;8(7):2328–31.
[3] Kruskal JB, Berkowitz S, Geis JR, Kim W, Nagy P, Dreyer K. Big data and machine learning—strategies for driving this bus: a summary of the 2016 intersociety summer conference. J Am Coll Radiol 2017 Jun;14(6):811–7.
[4] Densen P. Challenges and opportunities facing medical education. Trans Am Clin Climatol Assoc 2011;122:48–58.
[5] Cao C, Liu F, Tan H, Song D, Shu W, Li W, et al. Deep learning and its applications in biomedicine. Dev Reprod Biol 2018 Feb;16(1):17–32.
[6] Indonesia smartphone users [Internet]. Statista. [cited 2021 Jul 8]. Available from: https://www.statista.com/statistics/266729/smartphone-users-in-indonesia/.
[7] Smartphone users 2021 [Internet]. Statista. [cited 2021 Jul 8]. Available from: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.
[8] Wosik J, Fudim M, Cameron B, Gellad ZF, Cho A, Phinney D, et al. Telehealth transformation: COVID-19 and the rise of virtual care. J Am Med Inf Assoc 2020 Jun 1;27(6):957–62.
[9] Sarker IH, Hoque MM, Uddin MdK, Alsanoosy T. Mobile data science and intelligent apps: concepts, AI-based modeling and research directions. Mobile Network Appl 2021 Feb 1;26(1):285–303.
[10] Toh TS, Dondelinger F, Wang D. Looking beyond the hype: applied AI and machine learning in translational medicine. EBioMedicine 2019 Sep;47:607–15.
[11] Klang E. Deep learning and medical imaging. J Thorac Dis 2018 Mar;10(3):1325–8.
[12] Russell SJ, Norvig P, Davis E. Artificial intelligence: a modern approach. third ed. Upper Saddle River: Prentice Hall; 2010. p. 1132 [Prentice Hall series in artificial intelligence)].
[13] LeCun Y, Boser B, Denker J, Henderson D, Howard R, Hubbard W, et al. Handwritten digit recognition with a back-propagation network. NIPS; 1989.
[14] Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, Massachusetts: The MIT Press; 2016. p. 775 [p. (Adaptive computation and machine learning)].
[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014 Sep 4 [cited 2022 Apr 29]; Available from: https://arxiv.org/abs/1409.1556v6.
[16] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR) [internet]. Las Vegas, NV, USA: IEEE; 2016 [cited 2022 May 21]. pp. 770–8. Available from: http://ieeexplore.ieee.org/document/7780459/.
[17] Khan A, Sohail A, Zahoora U, Qureshi AS. A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 2020 Dec;53(8):5455–516.
[18] Huang G, Liu Z, van der Maaten L, Weinberger KQ. Densely connected convolutional networks. ArXiv160806993 Cs [Internet]. 2018 Jan 28 [cited 2022 Apr 29]; Available from: http://arxiv.org/abs/1608.06993.
[19] Kudva V, Prasad K, Guruvare S. Hybrid transfer learning for classification of uterine cervix images for cervical cancer screening. J Digit Imag 2020 Jun;33(3):619–31.
[20] Gonzalez RC, Woods RE. Digital image processing. third ed. Upper Saddle River, N.J: Prentice Hall; 2008. p. 954.
[21] Pandey B, Kumar Pandey D, Pratap Mishra B, Rhmann W. A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: challenges and research directions. J King Saud Univ - Comput Inf Sci. 2021 Jan S1319157821000070.
[22] Wahono RS. Analyzing requirements engineering problems. Japan: IECI Japan Workshop; 2003. p. 55–8.
[23] Hunt B, Ruiz AJ, Pogue BW. Smartphone-based imaging systems for medical applications: a critical review. J Biomed Opt 2021 Apr;26(4):040902.
[24] Freeman K, Dinnes J, Chuchu N, Takwoingi Y, Bayliss SE, Matin RN, et al. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. BMJ 2020 Feb 10:m127.
[25] Maier T, Kulichova D, Schotten K, Astrid R, Ruzicka T, Berking C, et al. Accuracy of a smartphone application using fractal image analysis of pigmented moles compared to clinical diagnosis and histological result. J Eur Acad Dermatol Venereol 2015 Apr;29(4):663–7.
[26] Ma S, Guan Y, Yuan Y, Tai Y, Wang T. A one-step, streamlined children's vision screening solution based on smartphone imaging for resource-limited areas: design and preliminary field evaluation. JMIR MHealth UHealth 2020 Jul 13;8(7):e18226.
[27] Murali K, Krishna V, Krishna V, Kumari B. Application of deep learning and image processing analysis of photographs for amblyopia screening. Indian J Ophthalmol 2020;68(7):1407.
[28] Hacisoftaoglu RE, Karakaya M, Sallam AB. Deep learning frameworks for diabetic retinopathy detection with smartphone-based retinal imaging systems. Pattern Recogn Lett 2020 Jul;135:409–17.
[29] Li F, Song D, Chen H, Xiong J, Li X, Zhong H, et al. Development and clinical deployment of a smartphone-based visual field deep learning system for glaucoma detection. Npj Digit Med 2020 Sep 22;3(1):1–8.
[30] Mezgec S, Koroušić Seljak B. NutriNet: a deep learning food and drink image recognition system for dietary assessment. Nutrients 2017 Jun 27;9(7):E657.
[31] Ji Y, Plourde H, Bouzo V, Kilgour RD, Cohen TR. Validity and usability of a smartphone image-based dietary assessment app compared to 3-day food diaries in assessing dietary intake among Canadian adults: randomized controlled trial. JMIR MHealth UHealth 2020 Sep 9;8(9):e16953.
[32] Williams S, Relton SD, Fang H, Alty J, Qahwaji R, Graham CD, et al. Supervised classification of bradykinesia in Parkinson's disease from smartphone videos. Artif Intell Med 2020 Nov 1;110:101966.
[33] Yoo TK, Choi JY, Jang Y, Oh E, Ryu IH. Toward automated severe pharyngitis detection with smartphone camera using deep learning networks. Comput Biol Med 2020 Oct;125:103980.
[34] Yang F, Poostchi M, Yu H, Zhou Z, Silamut K, Yu J, et al. Deep learning for smartphone-based malaria parasite detection in thick blood smears. IEEE J Biomed Health Inform 2020 May;24(5):1427–38.
[35] de Haan K, Ceylan Koydemir H, Rivenson Y, Tseng D, Van Dyne E, Bakic L, et al. Automated screening of sickle cells using a smartphone-based microscope and deep learning. Npj Digit Med 2020 May 22;3(1):1–9.
[36] Hu L, Horning MP, Banik D, Ajenifuja OK, Adepiti CA, Yeates K, et al. Deep learning-based image evaluation for cervical precancer screening with a smartphone targeting low resource settings - engineering approach. Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf 2020 Jul;2020.1944–9.
[37] Guo P, Xue Z, Mtema Z, Yeates K, Ginsburg O, Demarco M, et al. Ensemble deep learning for cervix image selection toward improving reliability in automated cervical precancer screening. Diagn Basel Switz 2020 Jul 3;10(7):E451.
[38] Montero PH, Patel SG. Cancer of the oral cavity. Surg Oncol Clin 2015 Jul;24(3):491–508.
[39] Lin H, Chen H, Weng L, Shao J, Lin J. Automatic detection of oral cancer in smartphone-based images using deep learning for early diagnosis. J Biomed Opt 2021 Aug;26(8).
[40] Sunarti S, Fadzlul Rahman F, Naufal M, Risky M, Febriyanto K, Masnina R. Artificial intelligence in healthcare: opportunities and risk for future. Gac Sanit 2021 Jan 1;35:S67–70.
[41] Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med Inf Decis Making 2020 Nov 30;20:310.
[42] Panch T, Mattie H, Celi LA. The "inconvenient truth" about AI in healthcare. Npj Digit Med 2019 Aug 16;2(1):1–3.
[43] Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing medical imaging data for machine learning. Radiology 2020 Apr 1;295(1):4–15.
[44] Jouppi NP, Young C, Patil N, Patterson D, Agrawal G, Bajwa R, et al. In-datacenter performance analysis of a tensor processing UnitTM. :vol. 17.
[45] Dreyfus-Schmidt L. Introducing MLOps [internet]. S.l.: O'Reilly Media, Inc.; 2020 [cited 2022 May 22]. Available from: https://learning.oreilly.com/library/view/~/9781492083283/?ar?orpq&email=^u.
[46] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018.
[47] Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. Artif Intell Healthc 2020:295–336.
[48] Vellido A. Societal issues concerning the application of artificial intelligence in medicine. Kidney Dis 2019;5(1):11–7.
[49] Gerke S, Minssen T, Yu H, Cohen IG. Ethical and legal issues of ingestible electronic sensors. Nat Electron 2019 Aug;2(8):329–34.
[50] Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. Sci Eng Ethics 2016 Apr;22(2):303–41.
[51] Rees EM, Nightingale ES, Jafari Y, Waterlow NR, Clifford S, Pearson Ca B, et al. COVID-19 length of hospital stay: a systematic review and data synthesis. BMC Med 2020 Sep 3;18(1):270.
[52] IBM's Watson recommended "unsafe and incorrect" cancer treatments - STAT [Internet], [cited 2022 Apr 29]. Available from: https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/.
[53] Castagno S, Khalifa M. Perceptions of artificial intelligence among healthcare staff: a qualitative survey study. Front Artif Intell [Internet] 2020 [cited 2021 Jul 9];3.

Available from: https://www.frontiersin.org/articles/10.3389/frai.2020.578983/full.
[54] Safi S, Thiessen T, Schmailzl KJ. Acceptance and resistance of new digital technologies in medicine: qualitative study. JMIR Res Protoc 2018 Dec 4;7(12):e11072.
[55] Longoni C, Bonezzi A, Morewedge CK. Resistance to medical artificial intelligence. J Consum Res 2019 Dec 1;46(4):629–50.