Application of Artificial Intelligence in Healthcare

Ethan Garnier[1] Matthew Tidd[2] Minh Nguyen[2]
ethan.garnier78@unb.ca mtidd2@unb.ca mnguyen6@unb.ca

[1]Department of Electrical and Computer Engineering, UNB
[2]Department of Mechanical Engineering, UNB

November 8, 2024

Abstract

Artificial intelligence and machine learning are revolutionary tools that enable rapid innovation in many different fields and industries. Healthcare and the medical field are foundational pillars of human society that protect the well-being of people around the world. Recent developments in artificial intelligence and machine learning have shown promise of successfully being applied in the field of medicine to aid physicians and improve quality of care. This paper aims to investigate artificial intelligence's roots in the field of medicine and follow some of its major developments up until today. Some such developments include the use of convolutional neural networks for classifying tumors, machine learning classification of antimicrobial resistance in bacteria, and the early identification of diseases. Due to the nature of artificial intelligence and the stringent regulations surrounding the field of medicine, the implications of applying artificial intelligence in healthcare is also discussed.

1 Introduction

2 History of Machine Learning in Medicine

The medical industry revolves around the collection of data and using that data to make diagnoses to save lives. As such, the application of a tool like machine learning, where learning from data is the focus, is clear in such an industry where data collection is so incredibly important. The history of artificial intelligence and machine learning in medicine spans many decades. Periods of rapid innovation followed by periods of AI ”Winters” define this history; however, it begins with the centralization and digitization of medical data that would help propel the research and development of the first prototype medical AI systems.

2.1 A Medical Library for Machines

During the beginning of research into AI’s potential application in the field of medicine, many issues were raised in regards to how the AI models could be used effectively. The two primary issues were 1) how to connect electronic health records and clinical documentation from around the world, and 2) how can a machine understand the underlying knowledge required for operation in a field as complex and critical as medicine [1]. Work towards solving these problems began in the 1960’s at the National Library of Medicine (NLM), where the Medical Literature Analysis and Retrieval System (MEDLARS), and its web-based search engine, PubMed, would be developed and provide access to the world’s largest repository of biomedical literature [1]. This, alongside the NLM’s support of capturing and computationally representing medical terminologies and vocabularies [1], essentially solved both of the aforementioned problems. There was now access to the medical data required to train an AI system, and there now existed a computational representation of medical vocabulary and knowledge required for an AI system to confidently operate in the field. It should be noted that the development of these tools was not specifically done for the advancement of AI in medicine; nonetheless, they would prove to be essential in the decades to come.

2.2 The First AI Winter

What followed in the 1970’s is what many refer to as the first ”AI Winter”, where there was a significant decrease in interest and investment into the field of AI, not just AI in medicine. Many believe the cause of this first winter to be due to the perceived limitations of artificial intelligence at the time [2], forcing people to disregard the field when in reality it was still in its infancy. Despite this deceleration in research, the 1970’s still represented a major stepping stone in the use of AI in medicine. This was accomplished through collaboration between universities across the United States and facilitated by an inter-university time-shared computer system based at Stanford called SUMEX-AIM (Stanford University Medical Experimental - AI in Medicine) developed in 1973 [1]. This collaboration allowed for discussions on many of the concerns that AI’s involvement in medicine presented at the time, as well as provided a breeding ground for innovative approaches to the newly developing field.

2.3 The First Successful Prototype Systems

Throughout and following the collaborative innovation of the 1970’s was the development of three majorly successful prototype AI systems used for medical diagnosis. These three systems, which will be explained and discussed in this section, provided an idea of what was possible with AI and laid the groundwork for AI’s future involvement in the medical field.

2.3.1 MYCIN

MYCIN, developed in the 1970’s, was a rule-based AI system that used a knowledge-base of roughly 600 rules [2] to identify infectious diseases a patient may have and then suggest antibiotic treatment tailored to the patient. Physicians would interface directly with the system by inputting the patient’s information and MYCIN would output a list of potential infections, as well as their respective treatments based on information provided about the patient. The list produced by MYCIN was achieved through backward chaining, where MYCIN would start with a given infection and work backwards to try to match it to the patient’s symptoms, and used a confidence factor representation to measure clinical uncertainty between the potential diagnoses [1]. Although this inference method and confidence factor is akin to calculating a probability for a given diagnosis being the correct one, the confidence-factor representation was actually found to psychologically aid in accepting MYCIN consultations [1]. MYCIN demonstrated that early AI systems had the potential to seriously improve the diagnosis process in medicine, even if only confined to hyper-specific subdomains of the field. Dr. Casimir Kulikowski, one of the authors of CASNET, went on to state that MYCIN ”was the most influential expert system that demonstrated the power of modularized rules for representing decision-making” [1]. The work done on MYCIN and the results it demonstrated would go on to encourage researchers in the field to create more AI prototype systems to aid in diagnosis within different subdomains of medicine, Dr. Kulikowski and his system CASNET being the next one.

2.3.2 CASNET

The Causal Associational Network (CASNET) was a consultation program designed specifically for glaucoma patients developed at Rutgers University in the 1970’s. CASNET used causal explanations of disease, alongside empirical knowledge of presumptive diagnoses, prognoses, and treatments to advise glaucoma patients [1]. The CASNET model was divided into three programs [2]: model-building, consultation, and a maintained database. By applying disease specific information to a patient of interest, CASNET could provide physicians with patient specific care advice. CASNET was unveiled at the Academy of Ophthalmology meeting in Las Vegas, Nevada in 1976 [2].

2.3.3 INTERNIST-I

Whereas MYCIN and CASNET were designed to be applied in only small, hyper-specific subdomains of medicine, INTERNIST-I was a system developed to cover a broad range of different medical diagnoses within internal medicine. Developed at the University of Pittsburgh by AI researcher Harry Pople and medicine specialist Dr. Jack Myers [1], INTERNIST-I was a prototype AI system designed to assist in general internal medicine diagnosis by modeling the expertise of Dr. Myers [3]. Due to the estimated size of the internal medicine field and the number of diseases it contains, designing a system that could confidently classify symptoms and provide correct diagnoses was a monumental task. Development of INTERNIST-I and its accompanying knowledge-base began in the 1970’s, and by 1988 the knowledge-base had grown to contain roughly 600 diseases and represented about 25 person-years of effort [3]. This massive knowledge-base would mean INTERNIST-I would have difficulty with complex diagnosis, when there wasn’t a single, clear answer. Despite this drawback, INTERNIST-I would demonstrate that the diagnosis process could be broken down into a pure classification problem, instead of the probabilistic models used by MYCIN and CASNET. As such, INTERNIST-I is considered a crucial step in the development of early AI systems in medicine.

2.4 The First AI Algorithms

The advent of AI in medicine provided a training ground to validate the usefulness of the different AI algorithms that had already been established at the time. By observing the impact a training algorithm has in a complicated field such as medicine, researchers would be able to judge how useful the algorithm could be in the greater AI landscape. Focusing only on medical diagnosis at the time of the first AI prototype systems discussed above, there existed two primary AI algorithms being employed: decision trees and Bayesian classification.

2.4.1 Decision Trees

In a medical diagnosis, it is up to the diagnostician to use the patient’s symptoms to iteratively rule out possible causes until the final diagnosis is reached. As such, the use of decision trees and decision rules to narrow down a medical diagnosis makes intuitive sense, and early researchers agreed with this. Decision trees were considered to be the most promising area for medical data analysis within a potential AI system [4]. This promise was further amplified with the invention of the Iterative Dichotomizer 3 (ID3) decision tree algorithm and its application in oncological diagnosis [4]. Decision trees continued to be a popular option for medical diagnosis due to their more easily interpretable structure. This allowed for the algorithm to better explain its inductive reasoning, re-assuring the physician and the patient on the diagnosis [4].

2.4.2 Bayesian Classification

Bayesian classification is a probabilistic approach to inference that uses Bayes’ theorem to calculate how likely a given hypothesis is. This naturally maps to the world of medical diagnosis, as given a list of patient symptoms, a Bayesian model assigns probabilities to each potential diagnosis, similar to a physician. Although the transparency, or ability to explain its reasoning through its representation, of Bayesian classification was a cause for concern early on, it was found that using Bayesian classifiers far outperformed other AI algorithms in medical diagnostic tests [4]. With improvements to the issue of transparency addressed in the 1990’s [4], Bayesian classification became a commonly used approach in the realm of AI medical diagnosis systems.

3 Recent Developments in Healthcare AI

Ever since the age of big data, there has been a surge in AI developments. This acceleration originates from several factors, including the wealth of data collected by big tech, the decreased cost of computational power, developments of more efficient machine learning techniques, and the availability of open-source machine learning packages The healthcare and medical fields are no exclusion from this wave of AI developments. In fact, AI and ML application is an active field of research, receiving attention from researchers, medical stakeholders, and policy makers. The immersion of the technology in the medical field is evident in the growth of FDA-approved AI/ML-enabled medical devices in Figure 1 [5]

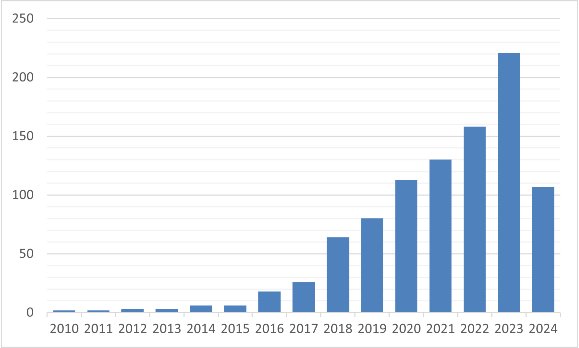

Figure 1: Number of FDA-approved AI/ML-enabled medical devices since 2010

3.1 Radiology

In radiology, machine learning is a powerful tool that can help physisicians with CT scan and MRI scan to diagnose diseases and extract latent insights from the scans. Some example work include

- Classification of triple negative breast canser using ultrasound images [6]
- Detection of pulmonary lung nodules from CT scans by convolutional neural network [7]
- Pneumonia detection from chest X-rays using modern deep learning [8]
- Detection of breast mass from mammography scans using convolutional neural networks [9]

3.2 Dermatology

In dermatology diagnosis, visual inspection remains the prominent method for determining the severity of a skin abnormalities or lesions. Machine learning emerges as an effective automation method as they learn the latent features that differentiate between benign and malignant lesions in skin melanoma. As early as 1994, the research in [10] applied neural network to automate the classification of malignant melanoma from melanoma-like benign tumors. The model was given the discriminant characteristics such as tumor shape and relative color, manually extracted from digital images of skin cancer. In [11], a convolutional neural network was trained on 100,000 clinical images of skin cancer and achieved an accuracy similar to that of a dermatologist. Once these models have been trained extensively on computationally efficient hardware, they can be packaged and deployed on mobile devices to make inferences on skin lesions. This can further increase the accessibility of machine learning functionalities to the wider population.

3.3 Hematology

In the case of hematological diseases, early prediction can help prevent progression and complication of blood disorders such as leukemia and lymphoma. Traditionally, practitioners diagnose hematological conditions by cytomorphologic phenotypic assessment of peripheral blood (PB) and bone marrow (BM) samples. Newer techniques for disease diagnosis includes flow cytometry and molecular genetic analyses. However, diagnostic ambiguity often occurs in such manual process depending on the operator experience and capabilities. The reproducibility of clinical results suffer from inter- and intravariations amongst skilled hematologists and pathologists [12, 13].

Therefore, there needs to be an automated process for reduced reliance on expert knowledge and increased consistency in data interpretation. Given its predictive power, machine learning implementation in disease diagnosis from blood samples is an active research topic.

In 2014, early developments of automated ML systems for cytomorphologic assessment had limited capabilities, performing tasks such as blood cell counting or classificiation of lymphoid cell types [14, 15, 16]. In [17], two ML models were developed to predict hematologic disease. One of them was trained on all available blood test parameters and the other trained on a reduced set of parameters usually obtained from patient admittance. When this input is paired with the knowledge of the five most likely diseases, the models achieve predictive performance on par with haematology specialists. In a research by *Wu et al.* [13], the researchers developed a BMSNet with the YOLO-v3 CNN architecture to assist in BM smear interpretation. The model performance was reported as comparable to that of expert hematologists and pathologists in various tests, except for the classification of myelodysplastic cases. *Matek et al.* [18] automated cytomorphologic examination of BM smears by training two CNN-based classifiers. The first model has the ResNeXt-50 architecture, previously used in [19] for recognition of acute myeloid leukemia blast cells from PB smears due to its low number of hyperparameters. The second model was a sequential CNN models with a simpler architecture for baseline performance comparison. The research in [20] employed a two-stage CNN architecture to classifies four types of white blood cells (leukocytes). The first layer of CNN leveraged a Faster R-CNN network to delineate the regions of interest and differentiate mononuclear cells from polymorphonuclear cells. The seconds layer consists of two parallels CNN with the MobileNet architectures and transfer learning to further categorize the subclasses of leukocytes. The lightweight architecture and parallelization in the second layer allows for faster inferencing time.

Flow cytometry or multiparameter flow cytometry (MFC) is a modern techniques for routine blood analysis and an enabling factor for AI integration in hematology. Modern MPC can analyze thousands of cells a second to generate a large dataset of high dimensional data [12]. Although this appears challenging for human interpretation and emphasizes the reliance on expert knowledge, a machine learning algorithm can be trained to interpret the latent representation and correlation in the flow cytometry data. The researchers in [21], leveraged machine learning by first converting MPC data into multicolor 2D images using a self-organizing map before feeding this data representation through a CNN architecture for classification. The developed model was capable of categorizing healthy cells from diseased ones, as well as the seven subclasses of mature B-cell neoplasm.

3.4 Ophthamology

In the field of ophthamology, a fundoscopy is a non invasive procedure for inestigating the patient's fundus, or the back of their eyes, to diagnose vision conditions and risk factors leading to vision loss. These detected factors can be used to predict the onset of diseases such as diabetic retinography, glaucoma, retina neoplasms, and macular degeneration [22]

Diabetic retinography (DR) is a condition where high blood sugar level causes damaged vessel in the retina. This condition is one of the common causes of vision loss and impairment amongst American living with diabetes and working age adults [23, 24]. In 2018, the FDA approved of IDx-DR, the first AI-enabled system for detecting DR. The research in [25] and [24] leveraged a multilayer CNN to train independent detectors for the anatomy and the characteristics of DR, including microaneurysms, hemorhages, and lipoprotein exudates. The results of these detectors are combined to return the levels of the diseases. The study in [26] implemented an Inception-v4 CNN architecture, to detect the severity grade of DR and provide a heatmap of regions of interest.

Researchers have also integrated AI into the early diagnosis of **age-related macular degeneration (AMD)**, a chronic and irreversible condition and one of the most common cause of central vision loss. One characteristics that signals the onset of AMD is the appearance of hard or soft drusen on optical coherence tomography (OCT) images or color fundoscopy images. ML can automate the drusen detection to diagnosie early signs of AMD, such as in *Burlina et. al* [27] where they employed Deep CNN on color fundus images. The network employs transfer learning on the AlexNet structure, which includes dropout lauers, ReLU activation, and normalization. The weights was trained using stochastic gradient descent and Nesterov momentum, and a learning rate of 0.001 for 50 epochs. This model was compared to a pretrained Overfeat DCNN (New York University) with a linear SVM for classification. The results showed comparable performance to expert human graders, with accuracy ranging from 88.4% and 91.6% and an ROC-AUC score betwen 0.94 and 0.96. In the research done by *Schlegel et al.* [28], the authors adopted an autoencoder architecture and CNN for learning a simplified representation of the raw images (Figure 2 shows the architecture of this network). The encoder portion of this networks generates an optimized abstract representation of the data embedding, which is then used to re-generate a corresponding image with class labels.

Glaucoma is another medical condition resulting in damaged optic nerve. The condition happends when there is a blockage preventing the drainage of the aqueous humor fluid and leading to pressure build-up in the

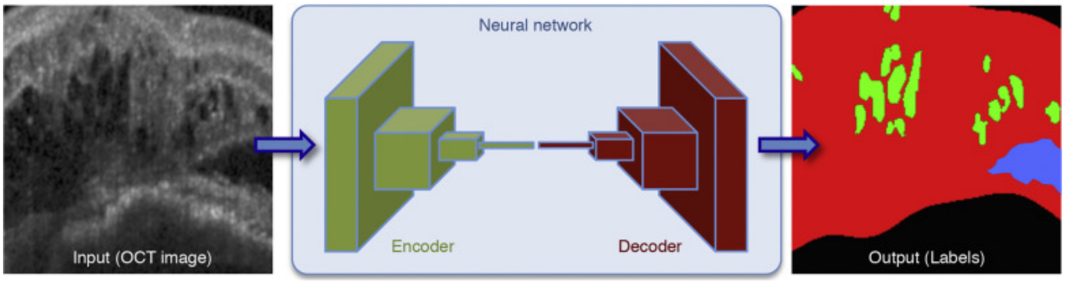

Figure 2: The proposed network with an autoencoder architecture using CNN in the encoder portion to form a simplified representation of the input features

eye. In *Shibata et al.* [29], the authors developed a deep residual network (ResNet) for glaucoma detection from 2D fundus photography. The network is based on the CNN architecture with residual connections to prevent the gradient vanishing and gradient divergence problem, thus allowing for more effective training of deeper networks. In a recent study by *Thompson et al.* [30], the authors implemented a deep learning solution for glaucoma detection using B-scan from OCT images. This reseach employed a residual deep convolutional neural network (ResNet34) and transfer learning to leverage the previously trainined feature extraction layers on the ImageNet dataset.

3.5 Oncology

The integration of AI and ML systems in cancer treatment is an active branch of research. Some potential use cases of AI in this field of research are radiotherapy dosage optimization and image segmentation for tissue abnormality detection. AI algorithms have shown promising improvements compared to manual planning in these fields [31]. In rafiation treament, the prescribed dose is determined by specialist prior to initial treatment. However, the variations in tumour biology poses challenges to this process as the dosage can significantly vary. Further complication may arise depending on the location of the tumour and surrounding organs, thus preventing the desired doseage delivery [32]. In this sense, AI methods can personalize the treatment plan and the optimal prescription achievable depending on the contour of the tumour and organs. In *Nguyen et al.* [?], the authors leverage the modified U-net CNN network (Figure 3) to achieve contour-to-dose mapping. Different

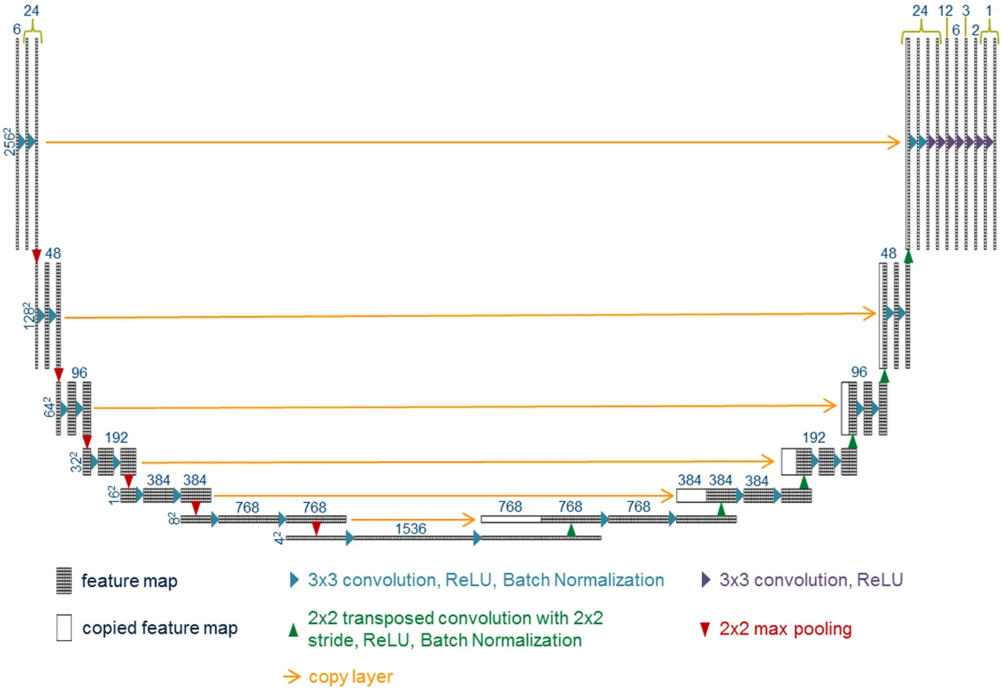

Figure 3: Modified U-Net architecture for personalized prostate cancer dosage prescription based on patient anatomy

4 Challenges

With such an impressive showing of studies wherein machine learning algorithms were leveraged within healthcare for the classification, inspection, and detection of several serious radiological, dermatological, haemotological and many other such diseases, it might come across as rather surprising to learn that real world deployments of machine learning algorithms within clinical practice are rare [33]. There are many reasons for this, and it is these inherent challenges that plague the widespread adoption and deployment of machine learning algorithms into the field of healthcare as a whole. Within this section several such key challenges will be highlighted and examined through both an ethical and legalistic lens, as well as pertaining to the disconnect between the design of machine learning algorithms and the understanding of those algorithms by medical practitioners.

4.1 Adaptation to a Clinical Setting

As outlined by Kelly et al. in [33], many challenges arise pertaining to the translation of machine learning models into clinical practice. While many studies have been performed that showcase seemingly impressive results within a clinical setting, in reality these studies have been retrospective in nature, that is they have used historically labelled data to train and test algorithms. In reality to best understand how these AI systems perform within a true clinical setting, *prospective* studies must be performed. Prospective studies represent the dynamic nature of a clinical setting, wherein patients are monitored as their medical characteristics or circumstances change as a function of time. In reality, the performance of these AI systems is likely to be worse in prospective studies as compared to retrospective studies, as the real-world data collected from a prospective study will vary substantially from the retrospective data used to train the algorithm [33].

Another problem that presents itself relates to the nature of machine learning literature. As a common practice within the machine learning community, many studies that are performed are published first to pre-print servers, and not to peer-reviewed journals [33]. Due to the stringent nature of the medical community, peer-reviewed evidence will be imperative for the trust and adoption of AI systems within the broader context of the healthcare industry. The common, so called ”gold-standard” of medical research is the use of randomized, controlled trials [33, 34], which is not common practice within the machine learning community. As such, high quality reporting of machine learning studies is required for the adoption of these systems into the healthcare industry. For clinicians and medical practitioners to trust and adopt these models into their daily use, there must exist clear and concise reporting on every aspect of these models, such that their usefullness within a clinical setting can be quantified [33].

Clinicians and medical practitioners are foremost qualified in medicine, and are not well versed in the knowledge and terminology of machine learning models; as such, many of the metrics that are often utilized within the machine learning community have no tangible meaning to them. As outlined by Kelly et al., ”accuracy does not necessarily represent clinical efficacy” [33]. Area under the curve of a receiver operating characteristic curve is a widely used metric within the machine learning community, however it not necessarily the best metric to represent clinicial applicability as it is not easily understandable.

Similarly, other such metrics such as specificity and sensitivity recorded at a given model operating point, might have no meaning to a doctor who is not well versed in these commonly used terminologies. Arguably the most important factor is that these measure do not reflect what is most important to patients, namely whether the use of the model results in a beneficial change in the care of the patient [33]. Therefore for ML algorithms to be widely adopted within the healthcare industry, these metrics must become more conducive of the clinical understanding of efficacy. Medical students and practicing clinicians should be properly trained in introductory machine learning courses to provide them with the knowledgebase to understand, digest, and adopt these models into their daily workflow.

There are numerous other challenges related to the adoption of machine learning models within a clinical setting. There exists difficulty in comparing models due to the variable methodologies used on datasets with differing distributions. The ”dataset-shift” problem also exists, where the introduction of AI models will generate a shift in medical practices, which will cause a change in the dataset used to train these models: effectively chasing a moving target. Current AI models are also far from generalisability, as there exist many unknown variables such as differences in equipment, laboratory equipment, electronic health record systems (EHRs), as well as differences in their codebases. Finally, there exist many logistical difficulties in implementing AI systems, as most healthcare data is not currently stored in a form that is easily digestable by these systems. Data is stored in numerous different mediums, such as medical imaging archival systems, pathology systems, EHRs, insurance databases, and other locations. The amalgamation of all of this data into a single, unifed dataset is a very difficult undertaking [33].

When these AI systems make false decisions in automated diagnoses, the results can sometimes be quite harmful to those affected. Due to the nature of EHRs, the data that these machine learning algorithms are trained on is not always the highest quality, or contains numerous inaccuracies. These so called ”data errors” can lead to potentially disastrous results, and these ”decision errors” have tangible, real effects that affect real patients [35]. Similarly, Habehh & Gohel highlight that the probabilty of error in prediction as well as the impact of this error are downsides of using these models within a clinical setting. Due to how these models rely on probablistic distributions, there is skepticism as to the validity of the predictions made by these models, and the implications of ML-based algorithms failing to operate as intended have severe and potentially fatal repercussions. Finally, they highlight how the rise of ML within the healthcare industry could fuel a decline in the personal relationship that many patients have between the healthcare practitioner, resulting in potential rejection from the public [36]. For these aforementioned reasons, these ML models are far from widespread adaptation into clinical settings, from the perspective of healthcare practitioners and clinicians.

4.2 The Black Box Problem

Another challenge with the use of machine learning algorithms within the healthcare industry is that of the *black box problem*. While this could fall under the section titled Clinical Adaptation, or perhaps the section titled Legal Issues, the author(s) felt it pertinent that this problem recieve its own dedicated section. While the black box problem does factor into why these models are not widely adopted within the healthcare industry from the viewpoint of a clinician, and it also has heavy legal implications, the problem is so inherently intertwined into the inner workings of machine learning that it was given its own section.

The black box problem relates to the interpretability of a machine learning model, and ones own ability to understand and explain the reasoning as to why and how the output of the model was produced. At its core, this problem stems from how these models are formulated. Large swathes of input data are fed in to train the model to detect patterns that are sometimes anomalous in nature that would be otherwise difficult for a human practitioner to detect. These models then take unforeseen test data and either classify or make a decision based on the type of algorithm. But intrisinic to the functioning of a machine learning algorithm is the manner in which this learning takes place. While we can observe the input and output of the model, we have no way of understanding what is happening within the model itself: this is what is known as the hidden layers.

These hidden layers learn new representations of the input, and each hidden layer is a re-representation of the input, as information is extracted at higher and higher levels of abstraction. As human operators we have no idea what occurs within the hidden layer. For a deep learning model with millions or even billions of parameters, one can imagine the complexity of the interactions between these parameters and the abstraction that occurs. Therefore the effectiveness of these AI models in the healthcare industry is limited entirely by their own inability to explain their decision making in a way that is understandable [33, 37]. As one can imagine, this is entirely problematic within the medical industry. A patient is not going to agree to a treatment without first understanding the decision making process that went into determining which treatment was best for them.

As outlined by Kelly et al., the best performing algorithms are unfortunately the least explainable (deep learning), whereas models with poorer performance are often more explainable (linear regression, decision trees) [33]. If an algorithm ends up making an incorrect decision, or incorrectly recommends something to a healthcare practitioner and they act upon that decision, there is no legal basis for explaining or defending the system as we simply do not know what lead to that decision. This makes it difficult for scientists to understand how data correlates to their predictions, and jeopardizes the faith that the public has in the medical system [34]. Therefore the fact that these algorithms cannot be supervised by medical professionals represents a glaring challenge with their implementation and widespread adoption [38].

4.3 Discrimination and Bias

Another challenge that plagues the implementation of ML into the field of healthcare is that of discrimination and bias. As demonstrated by algorithms implemented into non-medical fields, problematic decisions have been made by algorithms that reflect biases that are inherently woven into the data used to train them [39, 34, 33]. These biases stem from systemic problems with the collection of data, such as under-representation of minorities due to racial biases in dataset development, which in turn will lead to subpar prediction results [34]. Similarly, the implementation of AI systems affect groups that are disadvantaged by factors such as race, gender, and socioeconomic background. Such examples of this within the healthcare industry are mortality prediction algorithms with accuracy that varies by ethnicity [40] and dermatological algorithms that classify both benign and malignant moles with accuracy that could rival a board-certified dermatologist, but with severe underperformance on images of these types of moles when tested on people of colour [11, 41].

The unfairness of an algorithm can be classified into three constituents, such as (1) model bias, where the model elects to best represent the majority but not the under-represented, marginalized groups; (2) model variance, due to inadequate data stemming from minority communities; and (3) outcome noise, such as the effect of unobserved variables that interact with the models predictions [40]. Algorithms may become biased if there have been no medical studies done in certain populations, and as such the data is lacking [39]. The result of this is when the algorithm functions as intended, many people benefit. However when the decision making process falters, those that are negatively affected will be from marginalized communities that are under-represented within the very dataset used to train the algorithm, which compounds societal inequities [42].

Price and Cohen outlined bias in action by considering the allocation of scarce medical resources amongst multiple patients. They propose a scenario wherein a particular minority group responds worse off to the proposed medical intervention then other groups, and as such failure to collect information on this minority group would potentially lead to the algorithm giving the minority group more priority than if the data had been included. If the minority group actually responds better to the intervention, the opposite effect could result [43]. It is therefore important that strategies to minimize the discrimination of marginalized communities and bias stemming from disparities in datasets between races be implemented throughout the development of machine learning algorithms for use in the healthcare sector.

A bias that was surprising to the author(s) was one that did not stem from the data itself, but rather the intent of those designing the algorithms, as highlighted by Char and Shah in [39]. They highlight how the medical industry is entering into a new age of privatized development, wherein those responsible for designing ML algorithms for private sector healthcare corporations might be inclined towards malpractice to bolster their algorithm performance. There may exist temptation towards teaching machine learning systems to guide users towards clinical actions that would not necessarily lead to better care, but would lead to better quality metrics and hospital ratings. These systems might be designed in a manner that when they are being scrutinized by health regulators, they skew the data provided for public evaluation. The authors highlight Volkswagen's algorithm for passing emissions test by reducing their nitrogen oxide emissions when tested. Therefore, one must carefully consider the delivate balance between turning a profit and bolstering ratings, and providing improved healthcare for patients. Those responsible for desigining ML for use in healthcare are most likely not healthcare professionals themselves, and therefore need to seriously consider this tension.

4.4 Data Privacy and Ethics

Arguably the largest concern with the use of machine learning in the healthcare field is that of data privacy and ethics. As we all know, ML algorithms require copious amounts of data to sufficiently train them to a deployable state. In the context of healthcare, there are often issues with acquiring such a large repository of data. This stems from the fact that patient records are often regarded as confidential, and as such many healthcare organizations are not yet ready to offer up this personable information [34]. The deployment of ML systems into the healthcare field has vast implications towards the privacy and confidentiality of data. As early as 1982 some researchers within the medical field were describing the once Hippocratic cornerstone of medicine, confidentiality, as a decrepit and dated concept [44], an opinion that might be more relevant now than ever.

Char and Shah highlight how the implementation of ML as a decision making support tool into the healthcare field will make confidentiality an increasingly difficult concept to maintain, as patients whose data are not recorded cannot benefit from machine learning analyses [39], and the training of ML models will require as much data as possible to ensure clinical accuracy. There also lies the issue in conglomerating health records from all over into one central location, as health records are important and vulnerable data that might be targeted by hackers. There must exist sufficient security measures to prevent data breaches and protect this vulnerable, confidential information [34, 35]. To protect against these kinds of breaches, under the Health Insurance Portability and Accountability Act (HIPAA) patient information is modified by removing a set of 18 specified identifiers, such as names and emails [43]. Murdoch states in [38] however that several emerging computational strategies can be used to re-identify individuals in health data breaches, even if the information has been anonymized and scrubbed of identifiers, with success rates as high as 86% [38, 45, 46, 47, 48].

But what if a patient does not want their data to be used to train a ML model? Patient consent therefore becomes a glaring issue, with problems related to informed consent [49, 37], consent to data collection [49, 35, 38], privacy and consent [38], and the extent of consent [50, 43] being topics of debate. Informed consent in the context of data management refers to a persons right to know where and how their data is being used, as well as what kinds of data is being used. In terms of data collection, a patient should have consent as to whether or not their information can be used in the training of an ML model. In most recent years, Google has gotten into trouble with their predatory data collection tactics regarding healthcare data being harvested from patients unknowingly.

In 2017 Google's DeepMind partnered with the British NHS Foundation to use machine learning as an assistive tool in the management of acute kidney injury. What users of this app did not know, however, was that their data was inappropriately being annexed from the United Kingdom to the United states, and this private data was collected without the consent or agency of the user [49, 38]. As Murdoch states in [38], corporations may not always be encouraged to maintain privacy protection if they can monetize the data or otherwise gain from it, provided that the legal penalties are not high enough to offset this behaviour. In a 2018 survey of four thousand American adults, it was found that only 11% were willing to share their health data with tech companies, compared to 72% with physicians. Similarly, only 31% were either "somewhat confident" or "confident" in the abilities of tech companies to safeguard their data [38].

Furthermore, there are concerns about the extent of data collection, and where the line is drawn. Racine et al. highlight that the amount and intrusiveness of health data collection may grow over time as more and more healthcare facilities implement these technologies, highlighting how the development of AI will be used to justify extensive digitization of health status, healthcare practices, treatment use, and even the data collected from wearable devices [50]. It is therefore important to consider consent as a dynamic concept, where a patient can track the use of their data and reaffirm their consent if they still wish to. Giving a patient the ability to monitor the evolution of their data as a function of time will allow the patient to have more control over their data, and help to build a trusting relationship between the patient and the healthcare practitioner or ML engineer responsible for the algorithm using their data. Finally, another issue pertaining to consent is whether or not patients should have a role in deciding how their data should be used [43], specifically should they be able to decide what gets used from their dataset, and if so, how should it be used?

5 Case Study: Bias in ML

In a paper titled *Why Is My Classifier Discriminatory?*, Chen et al. examined the concept of fairness in the context of data, and sought to reduce discriminatory classification by not simply constraining the models used, but rather through augmenting the data collection methodologies used [40]. Their paper dealt with automatic decision support systems, like those discussed within this survey. They state that it is often hoped that there will be reduced bias and improved accuracy by training models on observational data, but this is not always the case. Other factors such as data quality and the choice of model however may encode unintentional discrimination, which has a systemically disparate impact [40].

The authors evaluted fairness with respect to protected groups of individuals defined by attributes such as gender and ethnicity, and this paper sought to address the impact of data collection and processign on discrimination and fairness, as this concept was less discussed within the literature compared to constraining the model itself. They describe discrimination in terms of differences in prediciton across all protected groups. In terms of automatic decision support systems, it can often be difficult to balance the tradeoff between predictive accuracy and fairness, as the predictions influence high-states decisions when used in areas such as the healthcare industry [40].

A model was defined as fair if its errors were distributed similarly accross all protected groups, as measured by the cost function γ. Discrimination was defined as referring to specific kinds of differences in the predictive power of models when applied to these different protected groups. Without diving too deep into the mathematical formulation of their predictive methodology, they based their predictions on a set of covariates:

$$X \in \chi \subseteq \mathrm{I\!R}^{\mathrm{k}}$$

and a *protected attribute*:

$$A \in \mathcal{A}$$

Where X represents the medical history of a patient in critical care, A is the self-reported ethnicity, and Y is mortality. Predictions learned from a training set *d* are denoted as:

$$\hat{Y}_d := h(X, A)$$

Fairness was quantified using a popular cost-based definition known as the equalized odds criterion, which states that a binary classifier $\hat{Y}$ is fair if its false negative rates (FNR) and false positive rates (FPR) are equal across all groups. FPR and FNR were defined for a member of a protected group $a \in \mathcal{A}$ as:

$$\mathrm{FPR}_a(\hat{Y}) := \mathbb{E}_X[\hat{Y} \mid Y = 0, A = a], \quad \mathrm{FNR}_a(\hat{Y}) := \mathbb{E}_X[1 - \hat{Y} \mid Y = 1, A = a]$$

The cost function γ was defined as:

$$\gamma_a \in \{\mathrm{FPR}_a, \mathrm{FNR}_a, \mathrm{ZO}_a\}$$

Where ZO_a is the *zero-one loss*, which is a measure of how many mistakes a hypothesis function h makes on the training set, and is given by:

$$\text{ZO}_a(Y) := \mathbb{E}_X[\mathbf{1}(Y \neq \hat{Y}) \mid A = a]$$

They utilized the MIMIC-III dataset, which is a collection of all clinical notes from 25,879 adult patients from the Beth Israel Deaconess Center, and used this dataset to predict hospital mortality of patients in critical care. Fairness was studied with respect to five self-reported ethnic groups, namely Asian (2.2%), Black (8.8%), Hispanic (3.4%), White (70.8%), and Other (14.8%). From the dataset, notes collected in the first 48 hours of an intensive care unit (ICU) stay were used, and patients that stayed for longer than 48 hours were considered. Features were identified by utilizing the tf-idf statistics of the 10,000 most used words from the notes, which assigned a weight to each word based on its frequency relative to the entire document. Their model was trained using 50% of the datam with hyper-parameters selected on 25%, and tested on 25%. The model that they used was a logistic regression model with L1-regularization, and they found that it achieved an AUC of 0.81.

Their findings were that Other and Hispanic groups had the highest and lowest generalized zero-one loss, respectively. The largest ethnic group (White) did not have the best accuracy, whereas the smaller ethnic groups tended towards the extremes. Despite having similar base-rates, Black and Hispanic error rates differed drastically. By increasing the training data size by subsampling and splitting the data, holding out at least 20% for testing, they found that differences in loss decreased with additional training data. Following this, they used clustering to identify clusters where the difference in prediction errors between protected groups was large. This revealed subpopulations for which the differences in zero-one loss were high, which allowed for a more targeted method to reducing discimination. Their work highlights that manipulation of data can be sufficient to improve fairness while maintaining predictive accuracy [40].

6 Case Study: ML for Antimicrobial Resistance Predictions

Antimicrobial resistance (AMR) occurs when microbes dangerous to humans and other living creatures evolve and develop resistances to the drugs that are used to combat these microbes. According to the World Health Organization (WHO), AMR and drug-resistant infections are a major threat to global health, with AMR-related deaths projected to reach roughly 10 million by 2050 [51]. As such, AMR is an active area of research as the global health community works to predict, identify, and prevent AMR cases around the world. Innovations in the fields of artificial intelligence (AI) and machine learning (ML) have allowed for their application in the fight against AMR as prediction and decision support systems. This section is a case study on how AI and ML is successfully being applied in the AMR medical field and the results it is yielding.

6.1 Application of ML in AMR Prediction

An important step in tackling AMR and drug-resistant infections is identifying resistant pathogens before they have the ability to cause harm. As such, ML is being applied in the classification and identification of potential drug-resistant pathogens through the analysis of their gene contents. The gene contents of available pathogen genomes can be used to predict resistances of those pathogens to various different antibiotics [52]. Classification is a common problem solved by AI and ML, hence its usefulness in this area. Training of these predictive models that will be tasked with classifying potentially drug-resistant genomes is accomplished through supervised learning, where genomes that are already labeled as drug-resistant or non-drug-resistant are used to train and test the accuracy of the model. In the context of AMR, there is already an abundance of identified drug-resistant and non-drug-resistant pathogen genomes, as such these pre-labeled genomes are used for training [52].

As with all applications of ML, it is critical that the trained model generalizes well to data it has not seen before. In the context of AMR predictions, this can be a challenge due to several factors surrounding the genomic data sets. First of all, the location, time, and habitat from which samples are taken matters greatly as samples with similar factors can produce large sets of similar genomes, causing a confounding effect in the ML model [52]. In addition to this, it is common for genetic data sets to posses an uneven distribution of AMR phenotypes [52], leading to a model incorrectly classifying the majority of examples as the AMR phenotype of majority. Finally, a major difficulty affecting ML model generalization is the potential for poor or contaminated genome assemblies [52]. Contaminated data may introduce false positives into the classification process, incorrectly training the model, and resulting in a model that generalizes poorly with non-contaminated data. Given the above reasons, researchers in the field of AMR believe that no single, universal AMR ML model can generalize perfectly and predict drug-resistance in any given pathogen. As such, ML in AMR prediction is being applied in much more specific cases where prediction models are looking to classify genomic data from specific environments or sample cultures as to reduce the confounding effects discussed above [52]. I believe that this idea extends past AMR prediction and into the general methodology seen in ML training. Modeling a subset of a given problem is not only easier and far more reasonable, but it can also yield far higher accuracy and potentially provide more utility than a universal solution.

6.2 ML Algorithms Employed in AMR Prediction

As was discussed above, AMR prediction is a classification problem in the context of ML. There exists a plethora of classification algorithms commonly used in ML, and many of these have seen application in AMR prediction studies. Gerontini et al [53] developed a system to early detect special cases of antibiotic resistance occurring in hospitals through the the use of the Naive Bayes classifier, Support Vector Machines (SVMs), and the C4.5 classification algorithm. This study, which looked to not only solve an AMR related issue but also compare ML classifiers, found that SVMs achieved the best results when compared to the other algorithms according to each test's F-score [53], however this is only in the context of the problem being solved. Sousa et al. [54] employed a decision tree (DT) algorithm to predict a patient's likelihood of infection with a β-lactamase, an enzyme which renders certain antibiotics inefective, producing organisms. DTs are a very popular classification algorithm employed in the field of AMR prediction, with Goodman et al. [55] also using DTs to predict the probability of carbapenem-resistant organisms colonization. Y. Kherabi et al. [56] performed a review of 36 AMR studies where ML models were being applied and found that 42% of those studies used DTs and 36% used a derivative of DTs, Random Forests (RF).

An important factor that impacts the ML classification algorithm chosen is its transparency. As with all ML applications in the field of medicine, ML algorithms applied in the field of AMR must be able to explain their reasoning behind a prediction. Physicians and researchers will not blindly trust the judgment of a model, especially when human health is involved, as such the model's results must be interpretable. Kim et al. [52] defines three aspects of interpretability relevant to AMR predictions: ability to evaluate individual input features, traceability, and the ability to assess the interactions of features. The reason why DTs are so popular within this field is due to the fact that they satisfy the above three aspects of interpretability. DTs can rank the important of training features by identifying features that reduce variance [52], and the hierarchical structure of DTs naturally allows for a model's decision path to be traced through each node of the tree. As a result, DTs provide an intuitive explanation to the classification process, making them an excellent algorithm of choice for AMR predictions.

6.3 Implications of ML in AMR Prediction

The threat presented by AMR and drug-resistant infections is a cause for global concern. The advent of AI's application in this fight is a sign of hope; however, this area of research is still in its infancy. The application of AI and ML in clinical decision making causes much ethical concern. When it comes to human life, how can a model be trusted? Although this concern may not be fully understood, it is certainly being respected. In the context of AMR research, ML is being used as a tool to assist in the surveillance and diagnosis of potentially drug-resistant microbes. This often excludes these models from making decisions directly involving patients, and is more of a tool to be used by researchers that will eventually benefit patients down the pipeline. I believe that this is an effective application of ML in the field of medicine as it does not place the responsibility of human life on a ML model, but instead provides an additional tool for experts in the field. In my opinion, in its current form, AI and ML should be used as a tool by the expert, it should not replace the expert.

7 Conclusion

References

[1] C. Kulikowski, “Beginnings of artificial intelligence in medicine (AIM): Computational artifice assisting scientific inquiry and clinical art - with reflections on present AIM challenges,” *Yearbook of medical informatics*, vol. 28, no. 1, pp. 249–256, 2019.

[2] V. Kaul, S. Enslin, and S. A. Gross, “History of artificial intelligence in medicine,” *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 807–812, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0016510720344667

[3] D. A. Wolfram, “An appraisal of INTERNIST-I,” *Artificial Intelligence in Medicine*, vol. 7, no. 2, pp. 93–116, Apr. 1995. [Online]. Available: https://www.sciencedirect.com/science/article/pii/093336579400028Q

[4] I. Kononenko, “Machine learning for medical diagnosis: history, state of the art and perspective,” *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, Aug. 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S093336570100077X

[5] FDA, “Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices | FDA,” 2024. [Online]. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices

[6] T. Wu, L. R. Sultan, J. Tian, T. W. Cary, and C. M. Sehgal, “Machine learning for diagnostic ultrasound of triple-negative breast cancer,” *Breast Cancer Research and Treatment*, vol. 173, no. 2, pp. 365–373, Jan. 2019. [Online]. Available: https://doi.org/10.1007/s10549-018-4984-7

[7] B. van Ginneken, A. A. A. Setio, C. Jacobs, and F. Ciompi, “Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans,” in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, Apr. 2015, pp. 286–289. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7163869

[8] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning,” Dec. 2017. [Online]. Available: http://arxiv.org/abs/1711.05225

[9] J. Arevalo, F. A. González, R. Ramos-Pollán, J. L. Oliveira, and M. A. Guevara Lopez, “Convolutional neural networks for mammography mass lesion classification,” in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2015, pp. 797–800. [Online]. Available: https://ieeexplore.ieee.org/document/7318482/?arnumber=7318482

[10] F. Ercal, A. Chawla, W. V. Stoecker, H. C. Lee, and R. H. Moss, “Neural network diagnosis of malignant melanoma from color images,” *IEEE transactions on bio-medical engineering*, vol. 41, no. 9, pp. 837–845, Sep. 1994.

[11] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017. [Online]. Available: https://www.nature.com/articles/nature21056

[12] W. Walter, C. Pohlkamp, M. Meggendorfer, N. Nadarajah, W. Kern, C. Haferlach, and T. Haferlach, “Artificial intelligence in hematological diagnostics: Game changer or gadget?” *Blood Reviews*, vol. 58, p. 101019, Mar. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0268960X22000935

[13] Y.-Y. Wu, T.-C. Huang, R.-H. Ye, W.-H. Fang, S.-W. Lai, P.-Y. Chang, W.-N. Liu, T.-Y. Kuo, C.-H. Lee, W.-C. Tsai, and C. Lin, “A Hematologist-Level Deep Learning Algorithm (BMSNet) for Assessing the Morphologies of Single Nuclear Balls in Bone Marrow Smears: Algorithm Development,” *JMIR Medical Informatics*, vol. 8, no. 4, p. e15963, Apr. 2020. [Online]. Available: https://medinform.jmir.org/2020/4/e15963

[14] Y. M. Alomari, S. N. H. Sheikh Abdullah, R. Zaharatul Azma, and K. Omar, “Automatic Detection and Quantification of WBCs and RBCs Using Iterative Structured Circle Detection Algorithm,” *Computational and Mathematical Methods in Medicine*, vol. 2014, no. 1, p. 979302, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2014/979302

[15] S. Alférez, A. Merino, L. E. Mujica, M. Ruiz, L. Bigorra, and J. Rodellar, "Automatic classification of atypical lymphoid B cells using digital blood image processing," *International Journal of Laboratory Hematology*, vol. 36, no. 4, pp. 472–480, 2014. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ijlh.12175

[16] S. Alférez, A. Merino, L. Bigorra, L. Mujica, M. Ruiz, and J. Rodellar, "Automatic recognition of atypical lymphoid cells from peripheral blood by digital image analysis," *American Journal of Clinical Pathology*, vol. 143, no. 2, pp. 168–176, 2015.

[17] G. Gunčar, M. Kukar, M. Notar, M. Brvar, P. Černelč, M. Notar, and M. Notar, "An application of machine learning to haematological diagnosis," *Scientific Reports*, vol. 8, no. 1, p. 411, Jan. 2018. [Online]. Available: https://www.nature.com/articles/s41598-017-18564-8

[18] C. Matek, S. Krappe, C. Münzenmayer, T. Haferlach, and C. Marr, "Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set," *Blood*, vol. 138, no. 20, pp. 1917–1927, Nov. 2021. [Online]. Available: https://ashpublications.org/blood/article/138/20/1917/477932/Highly-accurate-differentiation-of-bone-marrow

[19] C. Matek, S. Schwarz, K. Spiekermann, and C. Marr, "Human-level recognition of blast cells in acute myeloid leukemia with convolutional neural networks," Feb. 2019. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/564039

[20] C. Cheuque, M. Querales, R. León, R. Salas, and R. Torres, "An Efficient Multi-Level Convolutional Neural Network Approach for White Blood Cells Classification," *Diagnostics*, vol. 12, no. 2, p. 248, Feb. 2022. [Online]. Available: https://www.mdpi.com/2075-4418/12/2/248

[21] M. Zhao, N. Mallesh, A. Höllein, R. Schabath, C. Haferlach, T. Haferlach, F. Elsner, H. Lüling, P. Krawitz, and W. Kern, "Hematologist-Level Classification of Mature B-Cell Neoplasm Using Deep Learning on Multiparameter Flow Cytometry Data," *Cytometry Part A*, vol. 97, no. 10, pp. 1073–1080, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cyto.a.24159

[22] P. Kumar, S. Chauhan, and L. K. Awasthi, "Artificial Intelligence in Healthcare: Review, Ethics, Trust Challenges & Future Research Directions," *Engineering Applications of Artificial Intelligence*, vol. 120, p. 105894, Apr. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197623000787

[23] O. o. t. Commissioner, "FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems," Mar. 2020. [Online]. Available: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye

[24] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *npj Digital Medicine*, vol. 1, no. 1, pp. 1–8, Aug. 2018. [Online]. Available: https://www.nature.com/articles/s41746-018-0040-6

[25] M. D. Abràmoff, Y. Lou, A. Erginay, W. Clarida, R. Amelon, J. C. Folk, and M. Niemeijer, "Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 13, pp. 5200–5206, Oct. 2016. [Online]. Available: https://doi.org/10.1167/iovs.16-19964

[26] R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, S. Xu, S. Barb, A. Joseph, M. Shumski, J. Smith, A. B. Sood, G. S. Corrado, L. Peng, and D. R. Webster, "Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy," *Ophthalmology*, vol. 126, no. 4, pp. 552–564, Apr. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0161642018315756

[27] P. M. Burlina, N. Joshi, M. Pekala, K. D. Pacheco, D. E. Freund, and N. M. Bressler, "Automated Grading of Age-Related Macular Degeneration From Color Fundus Images Using Deep Convolutional Neural Networks," *JAMA Ophthalmology*, vol. 135, no. 11, p. 1170, Nov. 2017. [Online]. Available: http://archopht.jamanetwork.com/article.aspx?doi=10.1001/jamaophthalmol.2017.3782

[28] T. Schlegl, S. M. Waldstein, H. Bogunovic, F. Endstraßer, A. Sadeghipour, A.-M. Philip, D. Podkowinski, B. S. Gerendas, G. Langs, and U. Schmidt-Erfurth, "Fully Automated Detection and Quantification of Macular Fluid in OCT Using Deep Learning," *Ophthalmology*, vol. 125, no. 4, pp. 549–558, Apr. 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0161642017314240

[29] N. Shibata, M. Tanito, K. Mitsuhashi, Y. Fujino, M. Matsuura, H. Murata, and R. Asaoka, "Development of a deep residual learning algorithm to screen for glaucoma from fundus photography," *Scientific Reports*, vol. 8, no. 1, p. 14665, Oct. 2018. [Online]. Available: https://www.nature.com/articles/s41598-018-33013-w

[30] A. C. Thompson, A. A. Jammal, S. I. Berchuck, E. B. Mariottoni, and F. A. Medeiros, "Assessment of a Segmentation-Free Deep Learning Algorithm for Diagnosing Glaucoma From Optical Coherence Tomography Scans," *JAMA Ophthalmology*, vol. 138, no. 4, p. 333, Feb. 2020. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7042899/

[31] R. F. Thompson, G. Valdes, C. D. Fuller, C. M. Carpenter, O. Morin, S. Aneja, W. D. Lindsay, H. J. W. L. Aerts, B. Agrimson, C. Deville, S. A. Rosenthal, J. B. Yu, and C. R. Thomas, "Artificial intelligence in radiation oncology: A specialty-wide disruptive transformation?" *Radiotherapy and Oncology*, vol. 129, no. 3, pp. 421–426, Dec. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167814018302895

[32] E. Huynh, A. Hosny, C. Guthier, D. S. Bitterman, S. F. Petit, D. A. Haas-Kogan, B. Kann, H. J. W. L. Aerts, and R. H. Mak, "Artificial intelligence in radiation oncology," *Nature Reviews Clinical Oncology*, vol. 17, no. 12, pp. 771–781, Dec. 2020. [Online]. Available: https://www.nature.com/articles/s41571-020-0417-8

[33] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1, p. 195, Oct. 2019.

[34] B. Khan, H. Fatima, A. Qureshi, S. Kumar, A. Hanan, J. Hussain, and S. Abdullah, "Drawbacks of Artificial Intelligence and Their Potential Solutions in the Healthcare Sector," *Biomedical Materials & Devices*, vol. 1, no. 2, pp. 731–738, Sep. 2023.

[35] O. Ali, W. Abdelbaki, A. Shrestha, E. Elbasi, M. A. A. Alryalat, and Y. K. Dwivedi, "A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities," *Journal of Innovation & Knowledge*, vol. 8, no. 1, p. 100333, Jan. 2023.

[36] H. Habehh and S. Gohel, "Machine Learning in Healthcare," *Current Genomics*, vol. 22, no. 4, pp. 291–300, Dec. 2021.

[37] J. Guan, "Artificial Intelligence in Healthcare and Medicine: Promises, Ethical Challenges and Governance," *Chinese Medical Sciences Journal*, vol. 34, no. 2, pp. 76–83, Jun. 2019.

[38] B. Murdoch, "Privacy and artificial intelligence: challenges for protecting health information in a new era," *BMC Medical Ethics*, vol. 22, no. 1, p. 122, Sep. 2021.

[39] D. S. Char, N. H. Shah, and D. Magnus, "Implementing Machine Learning in Health Care — Addressing Ethical Challenges," *New England Journal of Medicine*, vol. 378, no. 11, pp. 981–983, Mar. 2018.

[40] I. Chen, F. D. Johansson, and D. Sontag, "Why Is My Classifier Discriminatory?" in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018.

[41] H. A. Haenssle, C. Fink, A. Rosenberger, and L. Uhlmann, "Reply to the letter to the editor 'Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists' by H. A. Haenssle et al." *Annals of Oncology*, vol. 30, no. 5, pp. 854–857, May 2019.

[42] T. Panch, P. Szolovits, and R. Atun, "Artificial intelligence, machine learning and health systems," *Journal of Global Health*, vol. 8, no. 2, p. 020303, Dec. 2018.

[43] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nature Medicine*, vol. 25, no. 1, pp. 37–43, Jan. 2019. [Online]. Available: https://www.nature.com/articles/s41591-018-0272-7

[44] M. Siegler, "Confidentiality in Medicine — A Decrepit Concept," *New England Journal of Medicine*, vol. 307, no. 24, pp. 1518–1521, Dec. 1982.

[45] E. Check Hayden, “Privacy loophole found in genetic databases,” *Nature*, Jan. 2013.

[46] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, “Identifying personal genomes by surname inference,” *Science (New York, N.Y.)*, vol. 339, no. 6117, pp. 321–324, Jan. 2013.

[47] Y. Erlich, T. Shor, I. Pe’er, and S. Carmi, “Identity inference of genomic data using long-range familial searches,” *Science*, vol. 362, no. 6415, pp. 690–694, Nov. 2018.

[48] S. Ji, Q. Gu, H. Weng, Q. Liu, Q. He, R. Beyah, and T. Wang, “De-Health: All Your Online Health Information Are Belong to Us,” Jun. 2019.

[49] S. Gerke, T. Minssen, and G. Cohen, “Ethical and legal challenges of artificial intelligence-driven healthcare,” in *Artificial Intelligence in Healthcare*. Elsevier, 2020, pp. 295–336.

[50] E. Racine, W. Boehlen, and M. Sample, “Healthcare uses of artificial intelligence: Challenges and opportunities for growth,” *Healthcare Management Forum*, vol. 32, no. 5, pp. 272–275, Sep. 2019.

[51] H. H. Balkhy, “World health organization: Update on amr,” 2021. [Online]. Available: https://www.hhs.gov/sites/default/files/who-update-amr.pdf

[52] J. I. Kim, F. Maguire, K. K. Tsang, T. Gouliouris, S. J. Peacock, T. A. McAllister, A. G. McArthur, and R. G. Beiko, “Machine Learning for Antimicrobial Resistance Prediction: Current Practice, Limitations, and Clinical Perspective,” *Clinical Microbiology Reviews*, vol. 35, no. 3, pp. e00 179–21, May 2022. [Online]. Available: https://journals.asm.org/doi/10.1128/cmr.00179-21

[53] M. Gerontini, M. Vazirgiannis, A. C. Vatopoulos, and M. Polemis, “Predictions in antibiotics resistance and nosocomial infections monitoring,” in *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*, Jun. 2011, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/5999112

[54] A. Sousa, M. T. Pérez-Rodríguez, M. Suarez, N. Val, L. Martínez-Lamas, A. Nodar, and M. Crespo, “Validation of a clinical decision tree to predict if a patient has a bacteraemia due to a b-lactamase producing organism,” *Infectious Diseases*, vol. 51, no. 1, pp. 32–37, Jan. 2019. [Online]. Available: https://doi.org/10.1080/23744235.2018.1508883

[55] K. E. Goodman, P. J. Simner, E. Y. Klein, A. Q. Kazmi, A. Gadala, M. F. Toerper, S. Levin, P. D. Tamma, C. Rock, S. E. Cosgrove, L. L. Maragakis, A. M. Milstone, f. t. C. P. E. Program, and t. C. M.-H. Program, “Predicting probability of perirectal colonization with carbapenem-resistant Enterobacteriaceae (CRE) and other carbapenem-resistant organisms (CROs) at hospital unit admission,” *Infection Control & Hospital Epidemiology*, vol. 40, no. 5, pp. 541–550, May 2019. [Online]. Available: https://www.cambridge.org/core/journals/infection-control-and-hospital-epidemiology/article/predicting-probability-of-perirectal-colonization-with-carbapenemresistant-enterobacteriaceae-cre-and-other-carbapenemresistant-organisms-cros-at-hospital-unit-admission/90C71FEB6EE4B84DFA6149A8967E18B5

[56] Y. Kherabi, M. Thy, D. Bouzid, D. B. Antcliffe, T. M. Rawson, and N. Peiffer-Smadja, “Machine learning to predict antimicrobial resistance: future applications in clinical practice?” *Infectious Diseases Now*, vol. 54, no. 3, p. 104864, Apr. 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666991924000198