Open Access

Journal of Artificial Intelligence & Robotics

Article type: Research Article

# State-of-the-art artificial intelligence techniques in healthcare publications, and their correlation with disease and data: A data driven analysis

***Sadegh Keshtkar[1,2]*; Dagmar Krefting[3#]; Anne-Christin Hauschild[3#]; Zully Maritza Ritter[3#]; Narges Lux[1#]; Aasish Kumar Sharma[1,2#]; Pavan Kumar Siligam[1#]; Julian Kunkel[1,2]***

*[1]AG-C, GWDG, Burckhardtweg 4, 37077 Göttingen, Lower Saxony, Germany.*

*[2]Department of Mathematics and Computer Science, Georg-August University, Wilhelmsplatz 1, 37073 Göttingen, Lower Saxony, Germany.*

*[3]Department of Medical Informatics, Georg-August University, Wilhelmsplatz 1, 37073 Göttingen, Lower Saxony, Germany.*

*[#]These authors contributed equally to this work.*

***Corresponding Author: Sadegh Keshtkar***

*AG-C, GWDG, Burckhardtweg 4, 37077 Göttingen, Lower Saxony, Germany.*

*Email: sadegh.keshtkar@gwdg.de*

**Received:** Oct 15, 2024

**Accepted:** Nov 19, 2024

**Published Online:** Nov 26, 2024

**Website:** www.joaiar.org



**Volume 1 [2024] Issue 2**

**Abstract**

Artificial Intelligence (AI) has become a transformative tool in medicine, improving disease prediction, its management and patient healthcare. This study employs a data driven approach to analyze the usage of AI techniques and Machine Learning (ML) models, the diseases they are applied to, and the utilized data sets, and their development in healthcare research from 2018 to 2022. A dataset of over 52,000 abstracts is obtained from PubMed and serves as the basis for the investigation.

To identify and label the abstracts, we employ n-grams mining in conjunction with two lists of known AI techniques and diseases, resulting in three categories: AI technique, disease, and data. The analytical approach involves uniquely counting the occurrences of these labeled instances.

The study reveals that deep learning, Convolutional Neural Networks (CNN), and neural networks are the most prevalent AI techniques in healthcare research. Among diseases, Neoplasms and COVID-19 emerged as the most extensively studied topics. Clinical Information and Magnetic Resonance Imaging (MRI) datasets were found to be the most widely used for ML applications.

Additionally, the paper delves into the increasing significance of Federated Learning (FL) and Reinforcement Learning (RL) in the healthcare domain. FL demonstrates promise as a privacy-preserving ML model, while RL shows potential in optimizing treatment plans and resource allocation, particularly for diseases like the nervous system and Neoplasms.

**Keywords:** Artificial intelligence; Machine learning; Diseases; Data; Healthcare; Data-driven analysis.

**Citation:** Keshtkar S, Krefting D, Hauschild AC, Ritter ZM, Lux N, et al. State-of-the-art artificial intelligence techniques in healthcare publications, and their correlation with disease and data: A data driven analysis. J Artif Intell Robot. 2024; 1(2): 1014.

## Introduction

AI techniques have made remarkable advancements, opening new avenues for transformative applications in healthcare. This paper presents a systematic analytical review aimed at unravelling the landscape of state-of-the-art AI and machine learning approaches in healthcare publications. Our investigation revolves around two fundamental research questions: we explore prevalent artificial intelligence techniques and machine learning models, diseases, and data encountered in healthcare-related research while deciphering their associations. This allows to identify best-practices and a gap analysis.

The first research question delves into the prevalence and popularity of AI techniques, diseases, and data in healthcare. By understanding the top AI techniques used in the literature, the medical conditions studied, and the data types employed, we gain valuable insights into current trends and research priorities in the field.

The second research question focuses on the relationships between the most utilized AI techniques, the most discussed diseases, and the referred data in healthcare-related research. This exploration aims to identify prevailing associations between AI techniques and machine learning techniques, specific medical conditions, and data types, potentially revealing key patterns and opportunities for targeted applications.

To conduct our data driven approach for a systematic analysis, we initiated a comprehensive data collection process using PubMed [1-5], a renowned biomedical literature database. Our search query incorporated Medical Subject Headings (MeSH) terms, enabling meticulous identification of relevant research articles for our investigation.

The gathered literature underwent rigorous data processing to transform PubMed records into labeled abstracts, facilitating subsequent analysis. By employing a strategic analytical framework, we extracted essential information on the top AI techniques, diseases, and data, elucidating their prevalence and prominence in healthcare-related research. This trend analysis enables us to gauge the dynamic landscape of AI and machine learning medical applications in the past five years, from 2018 to 2022.

Additionally, we conducted a correlation analysis to uncover intriguing relationships between the top 10 AI techniques and the most discussed diseases and data sources.

In the section titled "Exploring the Potential Impact of Federated Learning and Reinforcement Learning in Healthcare" we present detailed findings of these two ML models from 2018 to 2022. Our analysis reveals the trends of Federated Learning and Reinforcement Learning in healthcare, along with their correlations with specific diseases and data trends.

Our data driven approach significantly contributes to advancing the comprehension of artificial intelligence's profound impact on healthcare. By illuminating prevalent trends and correlations, including the promising potential of Federated Learning and Reinforcement Learning as leading AI methodologies in the future of health, we aim to catalyze future research agendas and guide the strategic direction of service providers within the healthcare domain. Moreover, our findings offer invaluable insights to inform clinical applications, facilitating the development of more targeted and effective AI interventions in medical practice. Through this endeavor, we aspire to empower service providers, researchers, and practitioners alike to harness the transformative potential of AI to improve patient outcomes and revolutionize healthcare delivery.

In the forthcoming sections, we will provide a comprehensive overview of the most influential AI techniques and machine learning models, the diseases they address, and the data they utilize in healthcare research. This paper consists of seven sections: introduction, literature review, basic terminologies, methodology, findings from a systematic analytical review, conclusion, followed by references and bibliography.

Ultimately, we envision a deeper appreciation of the potential and challenges associated with artificial intelligence and machine learning applications in healthcare, catalyzing advancements for improved patient outcomes and enhanced healthcare practices.

## Literature review

Researchers and healthcare professionals have increasingly turned to machine learning and Artificial Intelligence (AI) techniques to analyze health-related data, enabling enhanced diagnostic and treatment capabilities. As a result, there has been a surge in research exploring the potential of these models in healthcare applications. This literature review serves as an analytical overview, explicitly focusing on machine learning algorithms and corresponding data deployed in healthcare, while addressing the challenges and implications for researchers and practitioners. AI technology has witnessed increasing adoption in the field of healthcare, with numerous studies show casing its potential across various applications.

This section provides an overview of the existing literature, focusing on selected studies and developments that show case the transformative impact of ML in healthcare. In the subsequent sections, we will present a systematic analysis of these works.

### Disease diagnosis and prognosis

Several studies have explored the application of ML models in disease diagnosis and prognosis. For instance, [4] describes a disease prediction system that uses multiple machine learning algorithms to diagnose diseases based on symptoms, age, and gender, with the weighted KNN algorithm giving the best results with 93.5% accuracy [7]. Presents a machine learning algorithm designed to assess the likelihood of survival or mortality for patients who are either confirmed or suspected to be infected. The algorithm utilizes a dataset comprised of medical history, demographic information, and COVID-19-related data to train the model. By analyzing these factors, the algorithm can predict whether a patient is more likely to survive or succumb to the infection, providing valuable insights for clinical decision- making and resource allocation.

Treatment Prediction and Personalized Medicine Artificial intelligence and machine learning promise to transform cancer therapies by accurately predicting the most appropriate therapies to treat individual patients [3,2]. Demonstrates the potential of machine learning in clinical trials by uncovering patient-specific treatment effects and identifying responsive subgroups. The study highlights the capabilities of ML algorithms

in analyzing heterogeneous treatment responses, allowing for more targeted and personalized interventions.

**Image analysis and medical imaging**

ML models have significantly advanced image analysis in medical imaging. For instance, [1] presents a machine learning-based multi model computing approach for medical imaging that can be used for the classification and detection of Alzheimer's disease. The proposed approach uses MRI images that are preprocessed using the CLAHE algorithm to improve image quality [8]. Provides an overview of deep learning techniques for medical image analysis, covering topics such as Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), recurrent neural networks (RNNs), and their application to medical image analysis tasks. According to [9] the field of DL image analysis could help to reduce the barrier for Low-Middle Income Countries (LMIC) to accompany the innovation, driven largely by problem domain expertise and the creative application of this technology to analyze the medical data with less specialized software skills individuals in an effective way.

**Genomics and precision medicine**

The integration of ML with genomics has opened doors to precision medicine [6]. Discusses the contributions of machine learning algorithms in genomic medicine.

**Terminology**

This section elucidates the fundamental concepts and terms employed in our forthcoming methodology description, presented in the subsequent chapter. Our analysis encompasses three principal metadata or categories: Model, Disease, and Data.

**Technique:** As a specific type of category, a technique is delineated by the names of well-established AI and machine learning techniques or instances (e.g., "convolu- tional neural network") and includes their numerous variant names, as provided by the authors (e.g., "Convnet").

**Disease:** Disease is defined for this work as declared by the disease ontology organization [https://disease-ontology.org/]: A disease is a disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism. We used the WHO International Classification of Diseases (ICD) [https://www.who.int/standards/classifications/classification-of-diseases] for disease categorization and related analyses.

**Data:** This category encompasses data utilized to feed machine learning models or address medical conditions. Within the data category, a term is attributed to this classification when it corresponds to well-known datasets (e.g., "Image-net"), any information generated or employed to address a disease (e.g., "MRI images", "medical data", "patients' surveys", "clinical information", "blood tests"), any medical technique generating data for subsequent therapeutic purposes (e.g., "chromatography- mass spectrometry"), as well as data explicitly referred to as such in research papers (e.g., "Parkinson's Progression Markers Initiative").

Furthermore, our methodology incorporates several essential concepts, which have been instrumental in its formulation:

**Abstract:** In the context of this study, "abstract" refer to the succinct summaries extracted from PubMed records. These abstracts encapsulate the core content of research papers and serve as the foundation for the current study's systematic analysis.

**Label:** A label is a concatenated term representing the category of a corresponding instance in an abstract and is used for instance identification.

The generation of a label involves the concatenation of a category (model, disease, or data) with a subsection of an abstract that is identified as an instance of that particular category (e.g., a label for "deep learning" would be "model deep learning"). Alternatively, a label may be assigned to a group of AI-models, diseases, or data that pertain to the same instance but possess different nomenclatures (e.g., "disease covid19" as a label for "sars2", "covid 19", or "coronavirus"). However, this latter labeling approach is exclusively employed when we are aware of distinct naming conventions. Further- more, we also explored one additional tiers of labels to further refine our categorization of instances, enhancing the depth of our analysis. For instance, both "deep learning" and "convolutional neural networks" would fall under the overarching label of "Artificial Neural Networks" (i.e., for the data we could not find a reference to apply this higher level of categorization).

**Collection:** In the context of this research, a "collection" pertains to a list of tuples comprising instances' names and their corresponding labels, for instance: Tuple ("parkinson", "disease parkinson").

**Generalization:** The primary goal was to uphold label uniqueness within each abstract. To achieve this, when terms like "sars2", "coronavirus", and "covid 19" were present in an abstract, we unified them under a single label, namely "disease covid19". This ensured that only one label instance was retained in each abstract. By employing this approach, we eliminated label duplication and streamlined the study. Consequently, the number of papers or abstracts considered would be determined by searching for the "disease covid19" label, aligning with our focus on articles rather than duplicating category names within the same abstract.

**Word cloud:** Word Cloud is a data visualization technique widely used for representing textual data in a visually engaging manner. It generates a graphical representation of words within a given dataset, where the size of each word is proportional to its frequency in the dataset. Typically, the most frequent words appear larger and more prominent, creating an easily understandable visual summary of textual content. Word Clouds are often employed to quickly identify key themes, trends, or important terms within a body of text, making complex textual data more accessible and interpretable at a glance.

## Methodology

**Data collection and preprocessing**

In the current study, the following steps were carried out to discover the most popular machine learning models, diseases, and data in papers addressing healthcare using machine learning: (i) definition of research questions, (ii) building a general query to gather a collection of literature on machine learning and health from PubMed, (iii) processing the downloaded abstracts from PubMed, and generating labeled abstracts.

**Definition of the research question**

The purpose of this study, based on big data analytic, was

to answer this question: "What are the most popular AI techniques and machine learning algorithms used in the health domain, and what were the associated diseases and data utilized in conjunction with these algorithms, focusing on the potential implications for privacy concerns?".

The main question is detailed in three research questions.

1. What are the most popular AI techniques, diseases, and data frequently encountered in healthcare-related machine learning papers?.

2. What are the prevailing associations between top-utilized AI techniques, diseases, and data in healthcare related machine learning research?.

3. Considering the future of AI and machine learning in healthcare, especially the growing significance of privacy and decision-making concerns, what potential impact can be attributed to reinforcement learning and federated learning?.

Answering these questions provides an overview of existing practices in this field, which will guide future commercial and academic activities incorporating machine learning in healthcare. It highlights the distribution and interactions of AI techniques, diseases, and data over time, revealing recent trends in various sectors. The research process involved creating a search query, selecting data sources, and generating labeled abstracts. Statistics derived from labeled abstracts answer the paper's questions effectively.

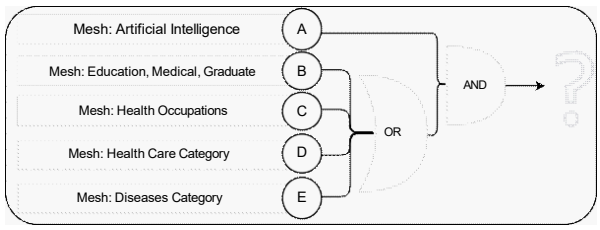


**Figure 1:** The mesh terms used in the main query.

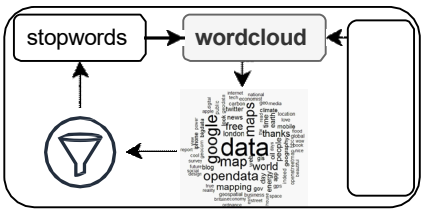


**Figure 2:** The iterative wordcloud filtering method.

**General PubMed query**

The initial stage was gathering a literature collection on machine learning and health. A systematic search strategy was implemented to identify relevant research articles for this data driven approach. The primary database utilized was PubMed, renowned for its comprehensive coverage of biomedical literature. The search query was formulated using a combination of Medical Subject Headings (MeSH) terms. The relevant Mesh terms and the query structure can be seen in (Figure 1). Ultimately, only the papers from 2018 to 2022 are considered.

**Approach selection and consideration of alternatives**

We encountered several challenges during our quest to extract the three categories of meta-data (technique, disease, and data) from the abstracts downloaded from PubMed. We explored various machine learning models, tools, and libraries, but unfortunately, none could automatically extract the desired meta-data from the abstracts.

With the limitation of time and resources, manually labeling over 52,000 abstracts proved infeasible as the first option. As an alternative, we sought to create a set of known names for our meta-data by compiling lists of machine learning models and diseases. To evaluate the results we applied "iterative word cloud filtering", as illustrated in (Figure 2).

This method involved feeding abstracts, a combination of different lengths of grams generated from abstracts, into the word cloud and then identifying words unrelated to our intended category, for example, any top words generated by the word cloud that did not pertain to machine learning models. These identified irrelevant words were then included in the stop words to be excluded in the next iterations. We repeated this process until eventually, in the end, the majority of the top words produced by the word cloud were relevant to our intended category, for example belonging to AI techniques.

Nonetheless, after applying "iterative word cloud filterin" on the abstracts and comparing the results, we realized that our lists were inadequate in covering most of the meta-data in the abstracts. The word cloud filtering brought to light ML model names not included in our initial list.

However, relying solely on iterative word clouds presents several limitations that hinder their effectiveness for our work. First and foremost, simply adding irrelevant top words to stop words and rerunning the process is insufficient.

Moreover, repetitive instances skew the results significantly, generating counts for top words necessitates exhaustive iterations, leading to a cumbersome and labor intensive process. The intricate nature of our item names further complicates matters, as merging different variants to represent the same item requires meticulous manual effort. Additionally, the presence of item names with multiple sections exacerbates tokenization challenges, significantly increasing both the token count and processing time.

Considering these factors, necessitating the adoption of more robust and efficient analytical methodologies. As a result, we relied on the hybrid approach, which allowed us to generate a comprehensive collection of terms and subsequently label the abstracts.

We employed "n-grams mining" as a systematic approach to manually extract the information from abstracts, and supplementing it with the lists containing recognized ML models and diseases. The outcomes obtained from this process underwent cross verification through the iterative word cloud filtering approach. This hybrid approach led to acceptable results.

In the next section we discuss the steps we followed in our hybrid approach, from downloaded PubMed abstracts, until the labeled abstracts.

**Processing the data from PubMed format to labeled abstracts**

After our general query, we obtained results matching 67,855 articles. These results were downloaded in the "PubMed" format as a text file. Subsequently, a series of preparatory steps were applied to ready them for analysis.

The process began with selecting key tags like "abstract," "publication date", and "PMID" from the PubMed text. After refining the data, we applied n-grams mining. The results of this mining process were then combined with two lists of known diseases and ML models. As a result of this merger, a collection of names and labels was developed, which was then used to annotate the abstracts. To improve label uniqueness and analytical generalization after annotation, we used a duplication elimination. Refer to (Figure 3) for visualization, with detailed explanations in the following steps.

**Feature/tag selection:** The downloaded PubMed data were meticulously filtered to include specific tags, notably "AB" (abstract) and "TI" (title) to enhance the search usefulness. This refined approach facilitated narrowing the search to sections containing more relevant material.

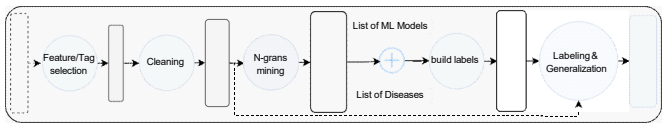


**Figure 3:** Steps applied on PubMed data to generated abstracts with labeled words.

**Cleaning the abstracts:** As part of the text processing, a series of techniques were applied to the abstracts (referred to as raw abstracts). As the concluding stage of the cleaning process, we proceeded to eliminate duplicated abstracts. This effort resulted in a final count of 51,530 cleaned abstracts for publications from the years 2018 to 2022 remain for further processing in the pipeline.

To address our research questions, it is imperative to determine the frequency count of each unique technique, disease, and data name exclusively mentioned within the abstracts. In our pursuit of compiling an inclusive array of these names, we endeavored to extract them directly from the cleaned abstracts to the best extent possible. Therefore, we employed an n-grams mining approach, aiming to capture as wide a range as we could while recognizing certain limitations.

**N-grams mining:** The N-grams mining process entailed generating n-grams from abstracts, which were then fed into a word cloud. We selected the most frequently occurring n-grams from the word cloud's output. N-grams of varying lengths, starting from 10 and going down to 2. Starting by higher lengths, we delved into more complex combinations of words, enabling us to identify higher-order relationships between terms. As we decreased the n-gram length, we aimed to capture simple word pairs and their associations.

For instance, the term "Alzheimer's disease" could initially be considered as a disease. However, when we encountered it within a sentence like "Alzheimer's disease neuroimaging initiative" we could confidently determine that it referred to a data source rather than a disease.

**Building collections:** The information extracted through the n-grams mining process, when combined with the data from the two acquired lists of recognized ML models and diseases, forms the fundamental basis for the collection creation process. Within this process, every term, along with its corresponding label, is meticulously stored in what we refer to as the collection.

**Labeling and generalization:** In this critical step, we utilized the previously created collection of terms and their respective labels to systematically replace each term within the abstracts with the corresponding label. For instance, the term "sars2" in an abstract was replaced by its designated label, "disease covid19". This labeling procedure allows for a more organized and structured data representation, enabling efficient analysis and further investigation.

Furthermore, to ensure the accuracy of our analyses, we only considered a unique occurrence of labels or their higher categories from each abstract. By doing so, we assured that only papers with distinct labels or higher categories made valuable contributions to our research.

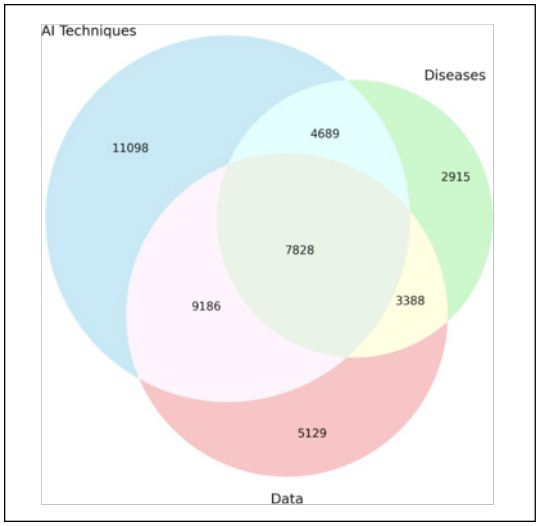


**Figure 4:** Venn diagram of 51530 abstracts labeled by technique (AI technique), disease, and data.

After the process, we successfully obtained a set of labeled abstracts along with their higher categories. This valuable resource allowed us to determine the number of unique occurrences for each them. These labelled abstracts provide the essential meta-data required to address the questions posed in the paper. Through this comprehensive approach, we have curated a well-structured dataset, empowering us to conduct in-depth analyses and draw meaningful insights related to the prevalence and utilization of various ML models, diseases, and data in healthcare-related machine learning papers.

**Resulting data set**

Following the implementation of our hybrid approach, we are left with two distinct categories of abstracts: those in which we identified instances (i.e., 44,458 abstracts contain any instances) and those where our efforts were unsuccessful in uncovering such instances (i.e., 7,072 abstracts contain no instance). The Venn diagram of all extracted instances can be seen in (Figure 4).

For abstracts without labeled instances, we conducted a thorough examination. Most of these uncovered abstracts either contained instances that did not contribute to our primary investigation of the top 10 instances, or those available not extracted instances already present in our collection with different names (e.g., "conv-net" as "convolutional neural network"). The number of such instances was not substantial enough to significantly impact the top ten rankings in each category.

Despite these not covered abstracts, we relied on our curated collection as the primary metric for analyzing the abstracts. This approach aligns with our research objectives, allows us to gain valuable insights into the prevalence and distribution of

disease, model, and data instances in healthcare-related machine learning research. In our article, we present a collection containing 286 data instances, 284 disease instances, and 658 model instances (with varying names) relevant to healthcare-related machine learning mentioned in the abstracts. As you can see in (Figure 4) out of 51,530 abstracts, 19,739 abstracts include disease labels, 32,113 abstracts include AI technique labels, 26,214 abstracts include data labels, and 7072 abstracts without any label. The details of overlapping of instances extracted from the abstracts can be seen in (Figure 4).

In the following, using this data, we discuss the popularity and correlations of top models, diseases, and data in healthcare research. Additionally, we explore the adoption of advanced methods, Federated Learning (FL) and Reinforcement Learning (RL), for addressing privacy and decision-making concerns in healthcare.

**Exploring AI impact: Techniques, diseases, and data in healthcare**

In this section, we present our findings addressing the research questions posed in the methodology section. Our analysis focuses on three critical aspects: Machine Learning (ML) models, diseases, and data within the context of healthcare.

Our approach unfolds in the following stages:

**Word cloud visualization:** To provide an immediate overview of the research landscape, we use word cloud images to show case the relative occurrence of terms related to AI techniques, diseases, and data. This visual method allows us to quickly understand prevailing themes and trends in the field.

**Temporal analysis with 2D tables and heatmaps:** We dive deeper into our investigation by analyzing the distribution of abstracts across different years. Utilizing 2D tables and heatmaps, we visualize the percentage of abstracts focusing on various AI techniques, diseases, and data types for each year. This enables us to identify potential shifts in trends over time.

**Correlation exploration:** We explore correlations between different factors, specifically investigating connections between diseases and AI techniques. This analysis aims to uncover valuable insights that could drive advancements in healthcare research.

**Federated learning and reinforcement learning:** Lastly, we explore the Potential impact of Federated Learning (FL) and Reinforcement Learning (RL) in Healthcare, specially as a solution for privacy and decision-making concern in this domain.

Our approach aims to provide a holistic understanding of how AI techniques, diseases, and data intersect in the healthcare domain. By offering these insights, we contribute to shaping future research directions and enhancing the broader understanding of the field.

In recent years, the field of healthcare has witnessed an unprecedented transformation due to the integration of state-of-the-art machine learning techniques. These techniques have revolutionized medical research and practice by enabling the analysis of complex datasets, prediction of disease outcomes, personalized treatment plans and more. This paper presents a systematic analytical review of the increasing adoption of machine learning in healthcare publications, shedding light on the exponential growth observed over the past few years.

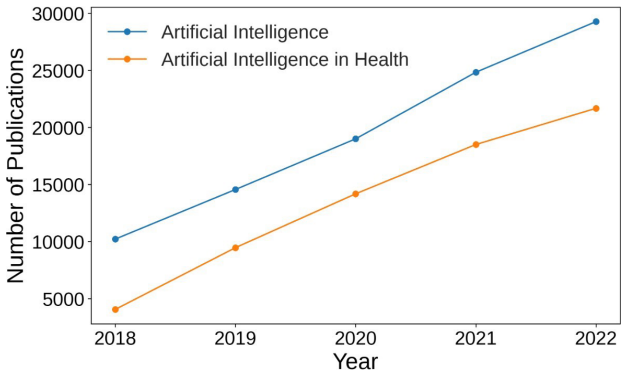


**Figure 5:** The number of publications for AI vs. AI in health (extracted from PubMed).

As it is shown in (Figure 5), starting from 2018 to 2022, the number of AI publications contributed to healthcare are relatively high. The number of health-related machine learning papers has exhibited a remarkable year-on-year increase, show casing the rapidly evolving interest and advancements in this interdisciplinary field.

**Hierarchical categorization of diseases, models, and datasets:** In our analysis, we employed a two-tiered approach to categorize diseases, models, and datasets, aiming to provide a structured understanding of their relationships.

**Disease categorization:** We utilized the World Health Organization International Classification of Diseases (ICD) [https://www.who.int/standards/classifications/classification-of-diseases] that plays a pivotal role on a global scale, offering comprehensive insights into the prevalence, origins, and ramifications of human ailments and mortality worldwide. Clinical terminologies categorized within the ICD form the cornerstone of health documentation and disease statistics across primary, secondary, and tertiary healthcare settings, as well as on mortality records. In contrast, for model categorization, we initially consolidated various model variants into overarching categories, guided by both existing literature and our own findings from the abstracts and chapters within our study.

**Technique categorization:** Our technique categorization encompasses a diverse array of AI and machine learning techniques, aiming to encapsulate the breadth of methodologies employed in health-related research.

The top-level (level 2) AI technique categories include:

*Artificial Neural Network* includes convolutional neural networks, deep learning, MLPs, LSTMs, and related variants. Bayesian incorporating bayesian statistical methods and probabilistic modeling approaches. *Clustering* encompassing various clustering algorithms. *Dimensionality Reduction* covering techniques such as PCA and t-SNE.

*Ensemble* including ensemble learning methods such as random forests, boosting, and bagging. *Federated Learning* reflecting collaborative learning approaches across decentralized data sources. *Instance Based* incorporating methods like K-Nearest Neighbors (KNN) and other instance-based learning algorithms. *Kernel Based* encompassing kernel methods such as Support Vector Machines (SVM) and kernel PCA. *Natural Language Processing* focused on Natural Language Processing (NLP) techniques and applications in healthcare. Regression covering regression-based predictive modeling approaches. Reinforcement Learning reflecting *reinforcement learning* al-

gorithms and applications in healthcare decision-making. Rule Based including apriori, sequential pattern discovery and other *rule-based* systems utilized in healthcare analytics. *Time Series* encompassing arima, moving average, and other time series algorithms tailored for temporal data analysis and forecasting. *Tree Based* incorporating decision trees, regression tree, and other tree based machine learning algorithms.

**Dataset categorization:** While our dataset categorization faced certain challenges due to the heterogeneity of data sources and content, we endeavored to provide a cohesive grouping based on shared characteristics. Despite the absence of a standardized reference for dataset categorization, we aimed to offer meaningful insights by grouping datasets according to their primary data modality or theme. For instance, datasets were grouped based on common attributes such as gene data, X-ray imaging, Magnetic Resonance Imaging (MRI), and other relevant features present within the data. Although some datasets contained diverse information, such as the Cancer Genome Atlas, spanning gene data, CT scans, MRI, and more, we aimed to provide a coherent categorization that would offer valuable insights into the types of data utilized in machine learning-driven healthcare publications.

In summary, for our classification of models and diseases, we established two distinct categories. The first category involves grouping various variants of a single instance, while the second category, representing a higher level of classification, amalgamates instances based on International Classification of Diseases (ICD) codes for diseases, as well as insights gleaned from various papers, online resources, and sections within our paper for the machine learning models. However, for our data analysis, we focused solely on the first category, which entails combining instances based on shared information. Subsequent analyses will be predicated on these two levels of categorization.

**Top analysis using wordcloud**

The wordcloud picture in this section generated using "iterative wordcloud filtering" applied on labeled abstracts. This approach leveraged the results of our labeling process rather than the two levels of categories. However, subsequent sections of our analysis will be based on these two category levels.

In this way, for example, to generate picture for top AI techniques, we tried to filter any word which is not start with "technique". Therefore, this tailored filtering process significantly accelerated the generation of the wordcloud compared to using raw abstracts applied in our evaluation process.

The wordcloud plots offer a visually engaging representation of the top AI tech- niques, diseases, and data in healthcare. The AI techniques wordcloud, (Figure 6).

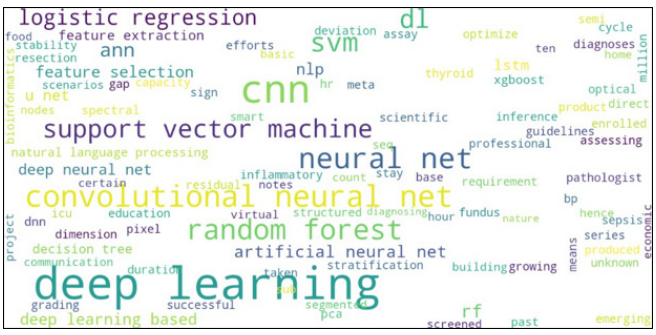


**Figure 6:** Word cloud visualization: Most frequent AI techniques.

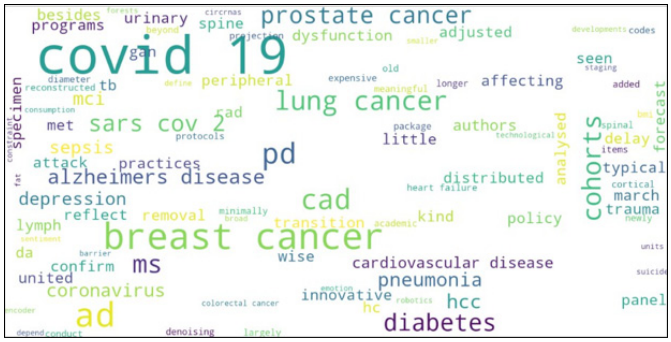


**Figure 7:** Word cloud visualization: Most frequent diseases.

Showcases the most frequently utilized AI and machine learning approaches in healthcare. Models like deep Learning, neural network, CNN, random forest, and SVM emerge as dominant players, signifying their widespread use and effectiveness in medical applications.

The diseases wordcloud (Figure 7), provides insights into the most extensively studied medical conditions in conjunction with ML models. Covid-19 stands out as the most prevalent disease, indicating the substantial efforts invested in research and diagnosis of the Coronavirus using ML techniques. Other diseases like breast cancer, diabetes, Alzheimer's Disease (AD) prostate cancer, and lung cancer also feature prominently, reflecting their significance in healthcare and the potential impact of AI-driven advancements in their management.

The data wordcloud, (Figure 8), illustrates the most data that have been instrumental in training and validating AI and ML techniques in healthcare. Identifying terms as datasets solely based on their use in abstracts is challenging. However, we've made an effort to compile terms that, according to our internet search results, could potentially qualify as data. The difficulty arises from the complexity and diverse nature of datasets and databases mentioned in the abstracts. For example, some databases are so general that are capable of storing a wide array of data types, which adds another layer of difficulty to our task of selecting and categorizing terms based solely on abstracts. As depicted in the wordcloud [8], terms such as biomarkers, experimental results, or data generated by some participants are among top used data. However, the more technical data such as Electroencephalogram (EEG), Electronic Health Records (EHR), Computed tomography, Magnetic Resonance Imaging (MRI), clinical data, and Genomics data occupy prominent positions, emphasizing their critical role in shaping the performance and accuracy of ML models in medical applications.

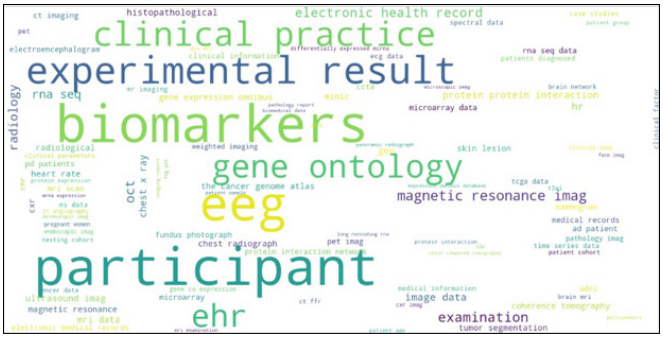


**Figure 8:** Wordcloud visualization: Most frequent data.

By examining these wordcloud pictures, we gain fast valuable insights. The recurring themes of certain models, diseases, and data highlight the key focus areas and pave the way for future research and advancements in data-driven disease analysis using ML in healthcare.

As evident from the word cloud visuals, relying solely on the instances outlined in the abstracts might not accurately depict the actual population of AI techniques, diseases, and data in the abstracts. Thus, in the subsequent sections, we endeavor to employ a more refined categorization approach. We posit that presenting these elements at a higher level of categorization will provide a clearer insight into the genuine correlations among the models, diseases, and data.

### Temporal analysis

Explore healthcare research through three key dimensions-top AI techniques, diseases, and data-within the years 2018 to 2022, unveiling the dynamic landscape of medical investigation. This analysis delves into each dimension, capturing the essence of utilization, growth, and attention that define the forefront of healthcare research.

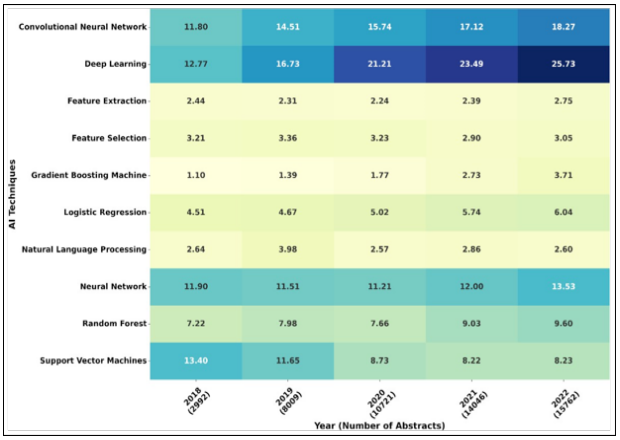


**Figure 9:** Trend: Top AI techniques (level I: grouped variant naming of the instances).

### Top AI techniques

In this section, we delve into the intricate details of top 10 ML models/AI techniques employed in the health domain spanning the years 2018 to 2022. Recognizing the complexity and diversity of techniques utilized, we have structured our analysis into two levels. The first level entails grouping variants of the same instance under a unified name, while the second level categorizes models based on the specific tasks they perform, such as regression models, tree based techniques, ensemble techniques, and more. This hierarchical approach allows for a nuanced understanding of the trends and patterns observed in the utilization of AI techniques and ML models within the dynamic landscape of healthcare.

Analyzing the heatmap plots reveals significant utilization patterns of AI and machine learning techniques in healthcare research.

As depicted in (Figure 9), Deep Learning (DL) and Convolutional Neural Networks (CNN) remain the dominant AI techniques, showing continuous growth. Their effectiveness in handling complex medical data contributes to their sustained popularity.

AI Techniques like Gradient Boosting, K-Nearest Neighbors (KNN), and Natural Language Processing (NLP) exhibit minimal and steady utilization. This suggests either limited applicability or saturation in exploration. Random Forest, Neural Network, and Logistic Regression experience moderate growth. However, they lag behind CNN and DL, indicating potential challenges in adapting these techniques to healthcare complexities.

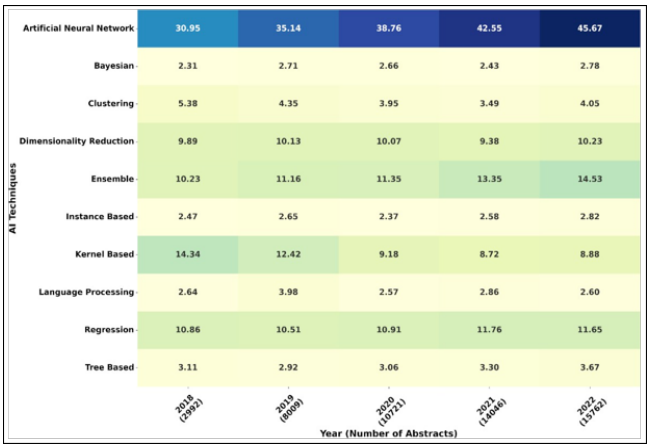


**Figure 10:** Trend: Top AI techniques (level II: task/nature based categorization).

Support Vector Machine (SVM) witnessed a decline in popularity between 2018 and 2022. This decline could potentially be attributed to the emergence of DL and CNN. Nonetheless, SVM retains its stature as one of the five most frequently employed AI techniques in healthcare applications.

The insights point to a need for refining CNN and DL for healthcare contexts. Techniques with stagnant attention may require re-evaluation, while the growth of others suggests further customization. Overall, these trends guide future research to optimize AI and machine learning's impact on healthcare challenges. In a higher level point of view, as it is depicted in (Figure 10), over the past five years (2018-2022), the family of Artificial Neural Networks, has experienced a remarkable surge in popularity, establishing itself as the leading paradigm in the field of health.

On the other side, in scenarios where data is largely unstructured, such as with text, images, or extensive sequences, techniques or models demonstrating slight growth often involve ensemble methods, dimensionality reduction, and regression. These approaches are pivotal in organizing and enriching data while retaining its crucial features. They accomplish this by either amalgamating multiple models to refine prediction accuracy or simplifying intricate datasets for tasks like forecasting or estimation. Given the prevalence of tasks requiring predictions about future conditions, such as the health status of patients, these techniques become invaluable for informed decision-making and proactive intervention.

The decrease in the utilization of kernel-based models reflects the rise of alternative solutions, particularly the family of artificial neural networks, which has reduced the necessity for kernel-based approaches. Despite this trend, kernel-based models persist as one of the top five utilized models in the health domain.

The gap in the utilization of Bayesian models, Clustering, and, more notably, Natural Language Processing techniques, underscores a promising area for research within the health domain. With the advent of Large Language Models (LLMs), such as GPT-3, there is an evident applicability in processing the di-

verse data types prevalent in healthcare and generating outputs akin to those produced by domain experts.

Investigating Bayesian models offers opportunities for probabilistic reasoning and uncertainty quantification, crucial in medical decision-making. Furthermore, delving into clustering techniques can unveil hidden patterns within healthcare datasets, aiding in patient stratification and personalized treatment plans. These avenues for research hold the potential to enhance diagnostic accuracy, treatment effectiveness, and overall patient outcomes.

### Top diseases

In this section, by examining the periods from 2018 to 2022, we delve into the realm of medical research and explore the top 10 most extensively studied diseases addressed using ML techniques, based two levels of categories of diseases we have built using ICD database of diseases.

Considering the first category, as depicted in (Figure 11), the prominence of Covid-19 in healthcare research from 2020 to 2022 is unsurprising, considering the global impact of the pandemic during this period. Interestingly, Malignant Neoplasm and Heart Disease also garnered significant attention, representing the second-tier focus among the top diseases. Furthermore, the attention distribution of other diseases remained relatively consistent throughout 2018-2022, following a same distribution pattern.

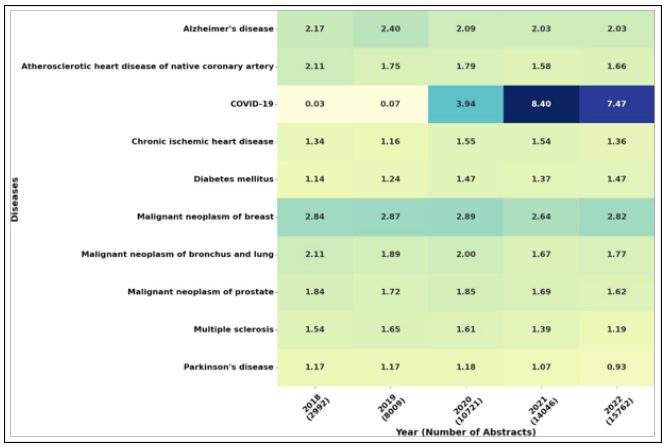


**Figure 11:** Trend: Top diseases (level I: based on ICD).

Upon delving into the second categorization based on ICD, intriguing patterns emerge. Despite the consistent rise in cases of Coivid-19 (referred to here as Codes for special purposes), an interesting shift occurs, placing Neoplasms at the forefront during the same timeframe. This highlights a significant focus on Neoplasms within current research and societal concerns. Following closely behind are diseases of the circulatory and Nervous systems, indicating a notable trend in attention and research focus during the years 2018-2022.

### Top data

In this part, we delve into the realm of data-driven research and explore the top 10 most utilized data in the health-related machine learning publications. As previously noted, categorization approach of extracted datasets from abstracts is primarily by identifying similarities in their content. By analyzing trends in data-centric ML research over the past five years, we aim to provide valuable insights into the prevailing data priorities and their transformative potential in advancing healthcare research.

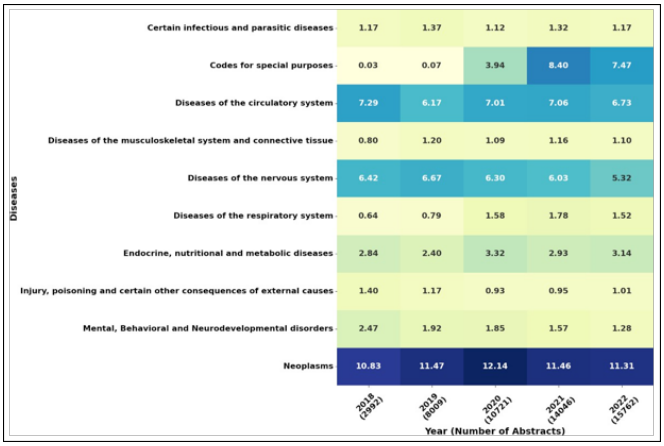


**Figure 12:** Trend: Top diseases (level II: based on ICD)

As depicted in (Figure 13), we present a heatmap diagram showcasing the prominence of various datasets in healthcare publications.

The diagram emphasizes the critical role of clinical information, which encompasses datasets extracted from abstracts originating from clinics and hospitals. These datasets, distinct from those listed in our data inventory, represent a vital source of information. Among the top five most utilized datasets are Experimental Health Records, Gene Ontology, Magnetic Resonance Imaging (MRI) data, and RNA Sequencing. Notably, there is a notable surge in the utilization of Computed Tomography (CT) and X-ray Imaging, attributed to the increasing adoption of artificial neural networks and their capabilities.

Conversely, slight decreases are observed in Participants, Electronic Health Records, and RNA and gene related data such as RNA Sequencing and Gene Ontology. The significant rise in Clinical Information between 2020 and 2022 underscores researchers' focus on leveraging data from hospitals and clinics, particularly during the COVID-19 pandemic.

Apart from Clinical Information, which comprises a fusion of diverse datasets from clinics, other data categories exhibit relatively comparable utilization within the healthcare domain. This indicates researchers' efforts to leverage all available data sources in healthcare research.

Moreover, this analysis highlights the importance of integrating various datasets to facilitate improved disease diagnosis, treatment planning, and personalized patient care. It underscores the potential of machine learning in driving evidence-based medical decisions.

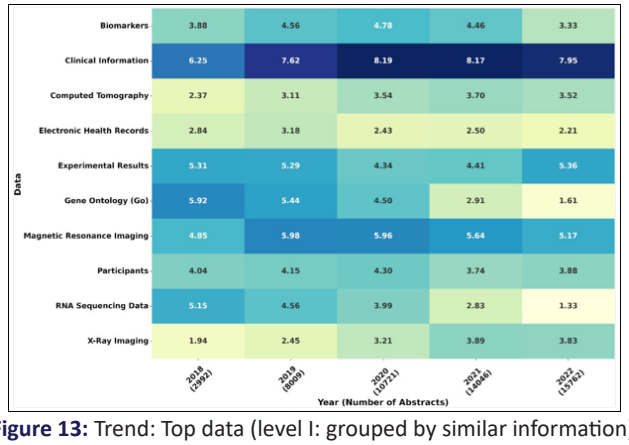


**Figure 13:** Trend: Top data (level I: grouped by similar information)

### Correlation analysis

In this subsection, we embark on a correlation analysis focused on the period from 2018 to 2022. We present correlations at both levels of categorization; however, for cases where data correlation is involved, we utilized the first level of categorization for the data.

In the pursuit of investigating the correlations between top AI techniques, diseases, and data in healthcare research, we selected abstracts that encompassed both relevant items. For instance, in analyzing the correlation between AI techniques and diseases, we specifically chose abstracts containing labels for both AI techniques and diseases. The rationale behind this selection was to place significant emphasis on elucidating the intricate interplay and relationships between these key categories.

By analyzing this refined subset of abstracts, we seek to shed light on the prevailing trends, knowledge gaps, and potential avenues for further exploration, ultimately contributing to the enhancement of AI applications in healthcare and ultimately, the improvement of patient outcomes and medical progress.

### AI techniques and diseases

Understanding the correlation between the top 10 AI techniques and the top 10 diseases is of significant importance in healthcare research. This subsection delves into the interplay between AI techniques and diseases to shed light on potential applications, research trends, and their transformative impact on healthcare.

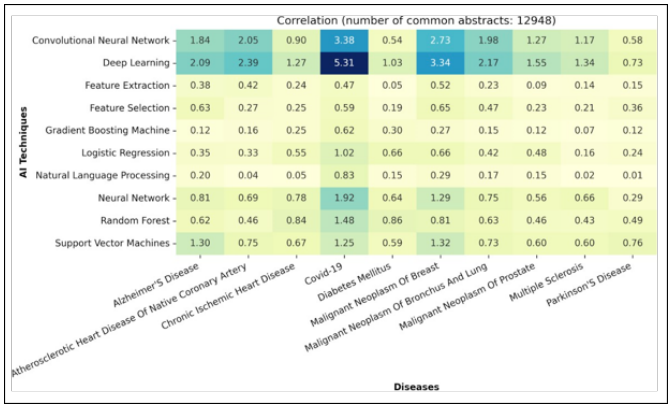


**Figure 14:** Correlation: Top 10 AI techniques and diseases (Level 1: grouped by variant naming of a the same instance).

(Figure 14) showcases the correlation between the top 10 AI techniques and the top 10 diseases based on level 1 categorization. Upon further analysis of the correlation between the top AI techniques and the diseases, distinct preferences and associations emerge.

Significantly, Covid-19 exhibits a strong correlation with Deep Learning, Convolutional Neural Networks (CNNs), and to a lesser extent, Neural Networks, underscoring their collaborative impact. Furthermore, the correlation of Covid-19 with other AI techniques remains notably high, surpassing other diseases except for Malignant Neoplasm of Breast and Alzheimer's Disease. Its association with various AI techniques underscores the thorough exploration of approaches for this unique phenomenon.

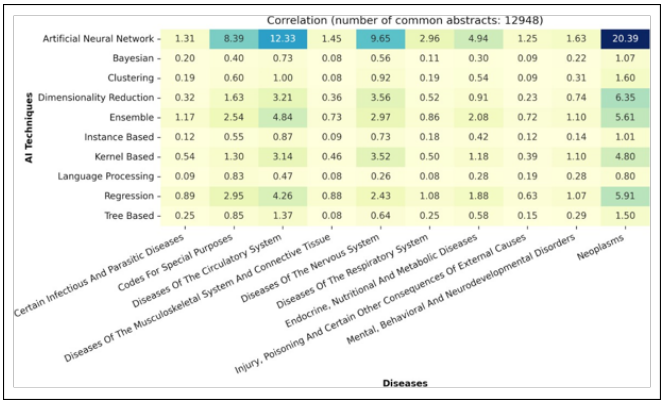


**Figure 15:** Correlation heatmap: Top 10 AI techniques and diseases (2018-2022).

Indeed, this holds true for all other diseases, as they exhibit a stronger correlation with the family of deep learning models. This further underscores the significance and potency of deep learning models within the realm of healthcare.

On the other hand, the second level of correlations is between diseases and Support Vector Machine and Random Forest. It can be attributed to the robust capabilities of these machine learning techniques in handling the intricacies inherent in disease data. SVM and Random Forest algorithms excel in capturing the complexities in disease related information by effectively handling non-linearity, feature selection, and noise reduction.

It's worth noting the remaining AI techniques, like Feature Extraction, Feature Selection, Gradient Boosting Machine (GBM), Logistic Regression, and Natural Language Processing (NLP), display minimal correlation with diseases. They don't exhibit clear preferences for specific diseases across the board, suggesting a more neutral correlation pattern. Though these techniques may show minimal direct correlation with diseases, investing in them is vital for advancing disease-related research. They offer invaluable contributions across the data science pipeline, from preprocessing tasks like Feature Extraction to providing interpretability and computational efficiency through models like Logistic Regression and GBM. Additionally, NLP techniques enrich our understanding of disease-related phenomena. Encouraging researchers to explore these models further can uncover their potential in addressing health-related problems.

The correlation insights offer valuable guidance for researchers and healthcare practitioners to make informed decisions about AI technique selection and its application in addressing critical healthcare challenges.

Based on the second level of categorization, level 2, the correlation between techniques and diseases is depicted in (Figure 15).

At this level, the strongest correlation is observed between Neoplasms and the Artificial Neural Network family. While Neoplasms also show notable correlations with other models, their relationship with Artificial Neural Networks stands out distinctly. The complexity and heterogeneity of Neoplasms data make them well-suited for analysis by deep learning techniques, particularly the family of artificial neural network, due to their ability to capture intricate patterns and relationships within complex datasets.

Interestingly, Diseases of the Circulatory System and Diseases of the Nervous System emerge as the second highest correlated diseases with Artificial Neural Network models. This correlation can be attributed to the nature of the data associated with these diseases, primarily consisting of images and unstructured data, and the inherent capabilities of deep learning models in processing such data types.

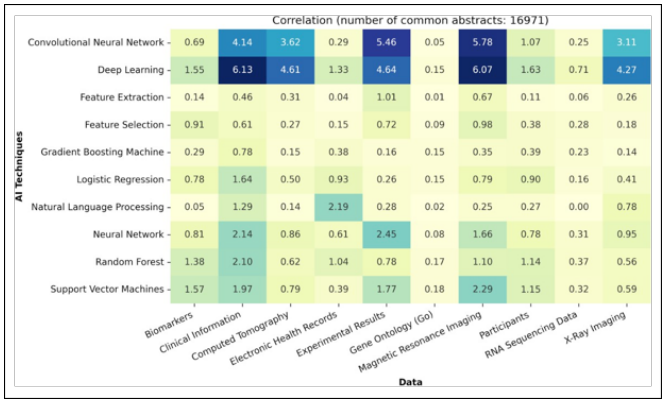


**Figure 16:** Correlation heatmap: Top 10 AI techniques and data (2018-2022).

As noted in the previous correlation plot 14 (i.e., the level 1 categorization), Covid-19, which is named Codes For Special Purpose in level 2 categorization plot 15, has a high correlation with Artificial Neural Network family.

Focusing on high correlations, we can find the correlation between Ensemble techniques and Regression with Disease of the Circularity System. Another notable correlation exists between Diseases of the Nervous System and Dimensionality Reduction, Kernel Based, and Ensemble methods, indicating both the intricacy of the disease data and researchers' endeavors to explore alternatives beyond the deep learning family.

Among the lowest correlations, NLP techniques exhibit minimal exploration of their potential in addressing medical issues, presenting a promising area for future research. Additionally, methods like Clustering and Bayesian techniques are under- rated in the health domain despite their considerable potential for unraveling the complexities inherent in disease data.

Examining the correlation heatmap plots for AI techniques and diseases reveals gaps in machine learning approaches that have not adequately addressed certain diseases. This highlights a valuable area for future research in AI and machine learning to enhance coverage and effectiveness in tackling these specific health issues.

**AI techniques with data**

The correlation analysis between AI techniques and data sets is important for AI service providers in health domain to spend their resources properly with respect to the correlation of some data and AI techniques. Here also we have depicted two levels of categorization for correlations between models and data (i.e., for the data we used in both cases the level 1 categorization).

The correlation analysis in (Figure 16), based on level 1 categorization of models and data, unveils distinct patterns and preferences in AI techniques usage for specific datasets.

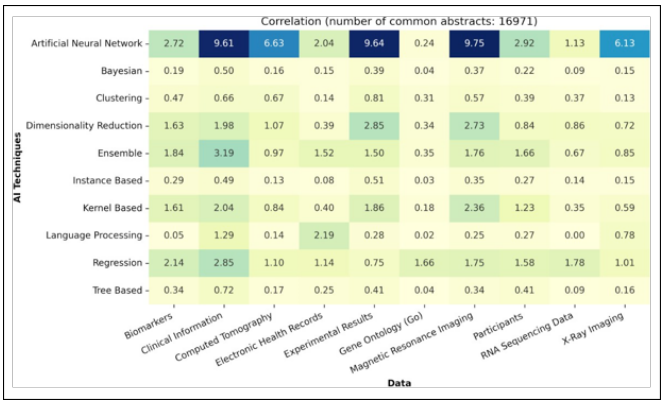


**Figure 17:** Correlation heatmap: Top 10 AI techniques and data (2018-2022).

Essentially, we observe that a majority of data types exhibit noticeable correlations with deep learning models. This trend can be attributed to either the exceptional performance of deep learning models in complex healthcare scenarios or the extensive application of these models in research papers.

For instance, the heatmap prominently showcases a strong correlation between Clinical Information, Magnetic Resonance Imaging, Experimental results, and Computed Tomography with Deep Learning and Convolutional Neural Network models. This correlation is primarily attributed to the complexity of structured/unstructured data (e.g., mixture of images, text, and tabular data) or image data (e.g., MRI images), where deep learning models excel most of the time.

Among various prominent correlations, notable connections emerge between Neural Network, Random Forest, and Support Vector Machine models and a spectrum of data sources including Biomarkers, Clinical Information, Experimental Results, and Magnetic Resonance Imaging. These correlations underscore the efficacy of these techniques in deciphering intricate patterns and relationships within data, as well as their resilience and adaptability, particularly in navigating the complexities of high-dimensional datasets such as biomarkers and clinical information.

In contrast, Gene Ontology and RNA Sequencing Data consistently display a minimum preference across all range of listed top AI techniques. This observation presents an intriguing avenue for future investigation by researchers.

While NLP techniques exhibits minimal correlation with the listed data sources overall, its noteworthy correlation with Electronic Health Records and Clinical Information underscores its emerging potential in extracting valuable insights from intricate datasets. This recent attention highlights NLP's significant promise in this domain, making it a compelling area for future investigation among healthcare researchers.

The models primarily regarded as data preprocessing steps, such as Feature Extraction and Feature Selection, consistently exhibit a low preference across all data types. This suggests that either these preprocessing steps are not prominently mentioned in the abstracts or are underutilized in the research works. However, it's imperative to recognize the significance of these methods as valuable steps in the application of machine learning to various types of data.

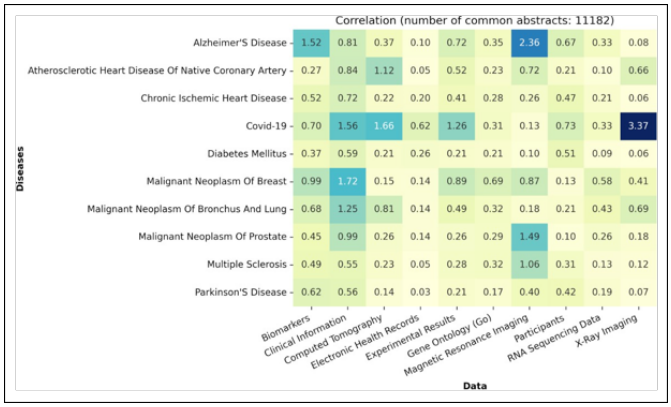


**Figure 18:** Correlation: Top 10 diseases and data (2018-2022).

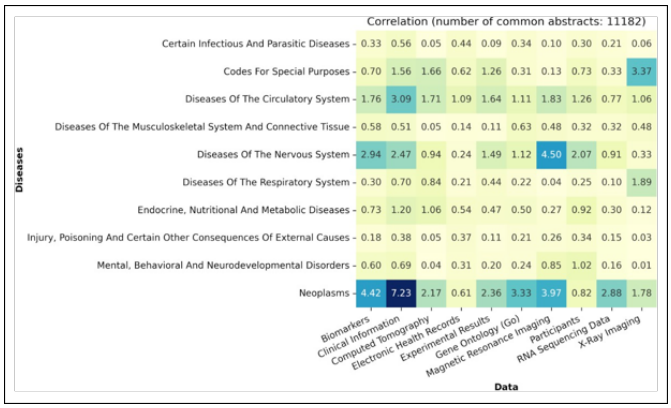


**Figure 19:** Correlation: Top 10 diseases and data (2018-2022).

**Diseases with data trends**

Understanding the correlation between the top 10 diseases and data trends is crucial in healthcare research. This section explores the relationship between prevalent diseases and the data landscape, uncovering valuable insights.

The correlation analysis holds immense significance for multiple stakeholders. It aids researchers and healthcare professionals in recognizing diseases affected by current data trends, enabling data-driven strategies for disease detection, treatment, and prevention. Businesses can leverage this knowledge to tailor products and services, fostering innovation and growth. Similar to other correlations, here we also consider two levels of categorization.

The heatmap diagram (Figure 18), visually represents the correlation between the top 10 diseases and data trends from level 1 categorization. It shows how specific data trends significantly influence certain diseases, while others may have different contributing factors to their development and spread.

This exploration unveils significant correlations. X-ray Imaging demonstrates a robust association with Covid-19, while Magnetic Resonance Imaging exhibits a slightly lower yet substantial correlation with Alzheimer's Disease. These findings align with expectations, considering the pivotal roles of these imaging modalities in diagnosing their respective diseases.

As expected, Covid-19 also demonstrates high correlations with Clinical Information, Computed Tomography, and Experimental Results, mirroring the extensive exploration of Covid-19 treatments during its early spread. Moreover, Alzheimer's Disease exhibits a notably high correlation with Biomarkers. This correlation can be attributed to the intricate biological mechanisms underlying Alzheimer's pathology, where biomarkers play a crucial role in detecting and monitoring disease progression. Electronic Health Records (EHRs) and RNA Sequencing Data consistently exhibit minimal correlations with most diseases, whereas Clinical Information demonstrates the highest correlation with all diseases except for Covid-19 and Alzheimer's Disease.

Notably, the absence of correlation between Alzheimer's Disease and EHRs suggests a significant gap deserving further investigation.

In the plot illustrating the correlation between diseases and data at the level 2 categorization, as depicted in (Figure 19), the most prominent correlation, notably distant from others, is observed between Clinical Information and Neoplasms. This robust association can be justified by the comprehensive nature of clinical data, which encompasses a wide array of patient information, diagnostic records, and treatment histories, all highly relevant to the study and management of Neoplasms. Basically, Neoplasms exhibit the highest correlation compared to other diseases across all listed data types. Conversely, Diseases of Neuron System demonstrate a notably greater correlation with MRI images, while Covid-19, namely, Codes for Special Purposes in this level of categorization, shows a heightened correlation with X-ray Imaging data, particularly when juxtaposed with Neoplasms.

After Neoplasms, Diseases of the Nervous System display a comparably high correlation with the majority of the listed data.

The remaining diseases exhibit a consistently low yet steady correlation with all listed data types, prompting researchers in the health domain to consider them as potential targets for further investigation.

These findings contribute to advancements in healthcare research and patient care by understanding the connections between diseases and data trends.

**The potential impact of FL and RL in healthcare**

In recent years, the healthcare sector has seen an unprecedented flood of data, necessitating the urgent need for powerful machine learning models to exploit its potential fully.

Federated Learning (FL) and Reinforcement Learning (RL) have emerged as attractive contenders for improving healthcare analytic. Federated learning enables independent healthcare institutions to train machine learning models on their own data without explicitly exchanging raw data. Furthermore, reinforcement learning, with its focus on decision-making and sequential tasks, offers promising avenues to address both privacy concerns and decision-making challenges in the health sector.

This chapter digs into their growing significance, and their correlation with the top data and diseases, as evidenced by data-driven trends from 2018 through 2022.

**Trend analysis**

This section delves into an analysis of the trend data from 2018 to 2022, showcasing the remarkable growth of Federated Learning (FL) and Reinforcement Learning (RL) in healthcare applications. (Figure 20) depicts a line graph demonstrating the year- by-year increase in using FL and RL in the healthcare field. The trend data from 2018 to 2022 shows that FL and RL have grown in popularity in healthcare. RL grew significantly, from 50 in 2018 to 212 in 2022 (a more than 4-fold rise), whereas

FL grew more rapidly, from 7 in 2018 to 89 in 2022 (a 12-fold increase). Given the yearly growth patterns, FL is clearly on course to overtake RL in popularity, perhaps becoming one of the top models used in healthcare. Both models are on an upward trend, with each year witnessing a significant increase in popularity rates. Additionally, (Figure 20) illustrates that RL's popularity has grown considerably more rapidly than FL. An essential point to note is that FL is still in its early stages of usage, which may account for its slower growth rate.

From 2018 to 2022, the number of instances of RL and FL may have been relatively low in comparison to other ML techniques. The spectacular annual growth of these two approaches, on the other hand, shows that they are primed to become among the most popular techniques in healthcare in the near future. As a consequence, academics and computing organizations working in healthcare-related fields would be advised to devote more resources in further study into these two methods. Their potential influence on healthcare applications is evident, and dedicated research in this field might produce significant advantages for the sector as a whole.

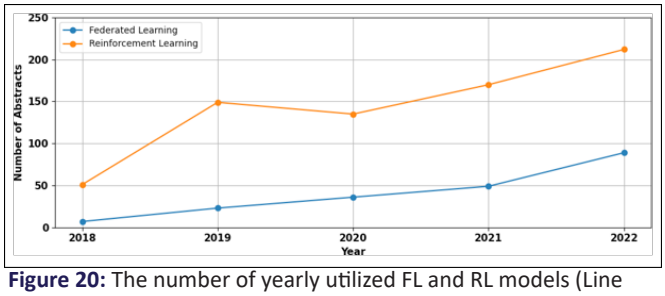


**Figure 20:** The number of yearly utilized FL and RL models (Line plot), and their yearly growth with respect to 2019 (Bar plot).

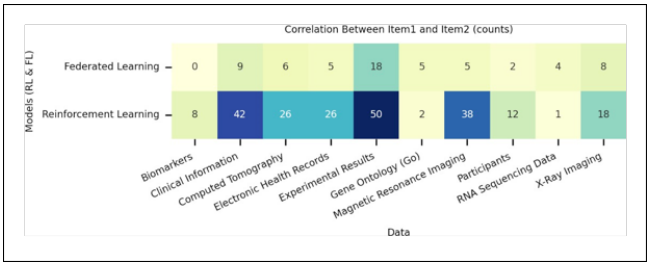


**Figure 21:** The number of top data and diseases addressed by FL and RL (2018-2022).

### Correlation analysis

In the realm of health-related machine learning papers, the relationship between Reinforcement Learning (RL) and Federated Learning (FL) with top data and diseases reveals interesting insights. From the graphs (i.e., based on level 2 categorization) pre- sented in (Figures 21 & 22) several key points emerge: First, for RL, the top data used includes Experimental Results, Clinical Information, and Magnetic Resonance Imaging. In contrast, FL does not exhibit distinguishable data preferences, as all data sources are considered at the same rate (i.e., except for Experimental Results). This discrepancy could be attributed to two potential reasons. Firstly, FL might be relatively new to researchers in this field, leading to limited data usage. Secondly, since privacy is a primary concern in FL, the emphasis may not be on the specific type of data but rather on ensuring secure and privacy-preserving collaborations.

Regarding the relationship between FL and RL with top diseases, as depicted in (Figure 22), the top diseases addressed by reinforcement learning are Neoplasms, Diseases of the Nervous System, and Diseases of the Circularity System, which shows similar preference as other top models depicted in (Figure 15). However, for federated learning, the top addressed disease is Covid-19 (i.e., in the level 2 categorization specified as Codes For Special Purposes), which shows the researchers' try to explore all.

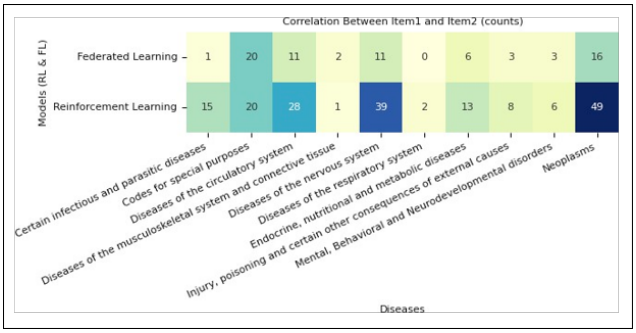


**Figure 22:** The number of top data and diseases addressed by FL and RL (2018-2022).

Possible solutions and approaches in dealing with the coronavirus pandemic. Nonetheless, the overall analysis reveals that FL does not exhibit distinguishable preferences for other diseases. Instead, its utilization seems to be more widespread and generalized across various healthcare applications. Whereas RL's top diseases are tumors, cancer and lesions. The prominence of RL in decision-making problems related to these diseases signifies its potential for optimizing treatment plans and resource allocation in critical healthcare scenarios. Understanding the popularity and utilization patterns of RL and FL in relation to top data and diseases can help organizations in the health domain identify areas where these models can be most impactful. For example, organizations can strategically invest in RL applications to address decision-making challenges, while recognizing that FL's potential is not yet fully tapped due to potential privacy concerns, even when FL seems to be the most appropriate tool to address these concerns.

## Conclusion

### Challenges and future directions

In this data driven approach, we explored the state-of-the-art AI techniques in health-care publications and their correlation with disease and data. Despite facing challenges in extracting the desired meta-data (i.e., AI technique, disease, and data instances) using available tools and platforms, we navigated our way through the vast PubMed journal library to select relevant abstracts for our study.

We encountered difficulties in only relying on lists of known models and disease, applying n-grams mining alone, or solely relying on iterative word cloud filtering. As a result, we adopted a hybrid approach by combining two lists of known models and diseases with n-grams mining, and then evaluated the results using iterative word cloud filtering. While we couldn't extract all occurrences of our meta-data in the abstracts, we were able to assess that the remaining abstracts and metadata would not significantly impact our top metadata, including top AI techniques, top diseases, and top data.

To gain deeper insights, we generated word clouds of the top AI techniques, diseases, and data using labeled abstracts from our carefully curated collection based on our methodology. Furthermore, we created heat maps to visualize the trends of these

top AI techniques, diseases, and data over the years 2018-2022. Additionally, we explored the correlations between the top AI techniques and diseases, as well as the top AI techniques and data, providing valuable information for scientists and service providers, such as High-Performance Computing (HPC) centers, to make informed decisions and allocate resources wisely.

Throughout our investigation, we encountered intriguing trends that indicate the potential impact of specific AI and machine learning models in the future of healthcare. Particularly, we found that Federated Learning (FL) and Reinforcement Learning (RL) are emerging as promising contenders among the top ML models used in health-related research. These models hold significant promise in addressing critical aspects such as privacy and decision-making concerns in healthcare applications.

It is essential to acknowledge that the quality and in formativeness of abstracts play a significant role in the validity of our findings. Since the analysis solely relies on the abstracts and their associated tags, it is important to highlight that some abstracts may lack sufficient information or be too brief to provide a comprehensive representation of the actual paper's content. This limitation could impact the accuracy and depth of the extracted metadata. Moreover, discrepancies between the abstracts provided by PubMed and the original papers' abstracts may introduce variations in the data used for analysis, which could potentially affect the overall conclusions drawn from the study.

Additionally, while comparing AI techniques in their abstracts could contribute valuable insights to the analysis, it is essential to acknowledge that our main intention was to consider both the AI techniques utilized in the research and those were compared to.

The impact of the COVID-19 pandemic and its related statistics may consider as a sort of proof of concept for our approach. By incorporating COVID-19-related meta-data into the analysis, we observed that it emerged prominently in the results exactly during COVID-19 pandemic period, and the percentage of meta-data related to COVID-19 demonstrated the reliability of the approach to some extent. This finding provides supporting evidence that our systematic analytical review effectively captures and identifies relevant meta-data, even in the context of rapidly evolving and significant events like the COVID-19 pandemic.

Looking forward, we acknowledge the importance of labeled articles to train NLP techniques and models (e.g., LLMs) for more accurate and efficient meta-data extraction. This allows us to further enhance the precision and reliability of our analysis. Additionally, we see potential for further investigation into the correlations we have uncovered and the top meta-data we have extracted. Such research endeavors hold promise in advancing the field of healthcare machine learning and supporting informed decision-making in medical research and service provision.

In conclusion, the analysis conducted in this paper is subject to the limitations posed by the quality and representativeness of the abstract. However, the inclusion of COVID-19 statistics and meta-data as a proof of concept provides confidence in the reliability of your approach and its ability to capture pertinent information. Overall, our research intends to shed light on the present landscape of AI and machine learning approaches in healthcare publications, as well as their relationships with disease and data, and to demonstrate the potential of FL and RL as top AI techniques in the future of health. Despite the challenges faced, we have made significant strides in understanding the trends and correlations, presenting valuable knowledge for the scientific community and service providers. We anticipate that our findings will serve as a stepping stone for future research, and pave the way for more sophisticated NLP Techniques and data driven insights in the healthcare domain and beyond, ultimately contributing to improved patient outcomes and enhanced healthcare practices.

**Declaration of generative AI and AI-assisted technologies in the writing process:** During the preparation of this work, the author(s) used Chat GPT to enhance language and readability. After using Chat GPT, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

**Acknowledgment:** We gratefully acknowledge the funding provided jointly by KISSKI and the German Federal Ministry of Education and Research (BMBF) under grant number 01IS22093. Their combined support has been instrumental in enabling this research, including the resources and infrastructure necessary for data analysis and study implementation.

## References

1. Fatemah H Alghamedy, Muhammad Shafiq, Lijuan Liu, Affan Yasin, Rehan Ali Khan, et al. Machine learning-based multimodel computing for medical imaging for classification and detection of alzheimer disease. Computational Intelligence and Neuroscience. 2022; 2022.

2. Paola Berchialla, Corrado Lanera, Veronica Sciannameo, Dario Gregori, Ileana Baldi. Prediction of treatment outcome in clinical trials under a personalized medicine perspective. Scientific Reports. 2022; 12(1): 4115.

3. Henry Gerdes, Pedro Casado, Arran Dokal, Maruan Hijazi, Nosheen Akhtar, et al. Drug ranking using machine learning systematically predicts the efficacy of anti- cancer drugs. Nature communications. 2021; 12(1): 1850.

4. Rinkal Keniya, Aman Khakharia, Vruddhi Shah, Vrushabh Gada, Ruchi Man- jalkar, et al. Disease prediction from various symptoms using machine learning. Available at SSRN. 2020; 3661426.

5. NCBI. Pubmed. Accessed. 1996.

6. Sameer Quazi. Artificial intelligence and machine learning in precision and genomic medicine. Medical Oncology. 2022; 39(8): 120.

7. Mario A Quiroz-Ju´arez, Armando Torres-G´omez, Irma Hoyo-Ulloa, Roberto de J Le´on-Montiel, Alfred B U'Ren. Identification of high-risk covid-19 patients using machine learning. Plos one. 2021; 16(9): e0257234.

8. S Suganyadevi, V Seethalakshmi, K Balasamy. A review on deep learning in medical image analysis. International Journal of Multimedia Information Retrieval. 2022; 11(1): 19-38.

9. Douglas Williams, Heiko Hornung, Adi Nadimpalli, Ashton Peery. Deep Learning and its Application for Healthcare Delivery in Low and Middle Income Countries. Frontiers in Artificial Intelligence. 2021; 4.