Research Article

Health Informatics Journal

Health Informatics Journal
1–32

DOI: 10.1177/14604582241307839
journals.sagepub.com/home/jhi
Sage

# Researching public health datasets in the era of deep learning: a systematic literature review

**Rand Obeidat**
Department of Management Information Systems, Bowie State University, Bowie, USA

**Izzat Alsmadi**
Department of Computational, Engineering and Mathematical Sciences, Texas A & M San Antonio, San Antonio, USA

**Qanita Bani Baker**
Department of Computer Science, Jordan University of Science and Technology, Irbid, Jordan

**Aseel Al-Njadat**
The University of Jordan, Amman, Jordan

**Sriram Srinivasan**
Department of Management Information Systems, Bowie State University, Bowie, USA

**Godswill Ashong* and Ifeanyi Osigwe***

## Abstract

**Objective:** Explore deep learning applications in predictive analytics for public health data, identify challenges and trends, and then understand the current landscape. **Materials and Methods:** A systematic literature review was conducted in June 2023 to search articles on public health data in the context of deep learning, published from the inception of medical and computer science databases through June 2023. The review focused on diverse datasets, abstracting applications, challenges, and advancements in deep learning. **Results:** 2004 articles were reviewed, identifying 14 disease categories.

*Collaborators, Department of Management Information Systems, Bowie State University, Bowie, USA.

**Corresponding author:**
Rand Obeidat, Department of Management Information Systems, Bowie State University, 14000 Jericho Park Road Bowie, MD 20715-9465, USA.
Email: robeidat@bowiestate.edu

Observed trends include explainable-AI, patient embedding learning, and integrating different data sources and employing deep learning models in health informatics. Noted challenges were technical reproducibility and handling sensitive data. **Discussion:** There has been a notable surge in deep learning applications on public health data publications since 2015. Consistent deep learning applications and models continue to be applied across public health data. Despite the wide applications, a standard approach still does not exist for addressing the outstanding challenges and issues in this field. **Conclusion:** Guidelines are needed for applying deep learning and models in public health data to improve FAIRness, efficiency, transparency, comparability, and interoperability of research. Interdisciplinary collaboration among data scientists, public health experts, and policymakers is needed to harness the full potential of deep learning.

**Keywords**

public health datasets, deep learning applications, predictive modeling, EHR analysis, trends and challenges

## Introduction

The fusion of deep learning methodologies with extensive and heterogeneous health datasets has emerged as a powerful and trans-formative field, revolutionizing the way of analyzing, interpreting, and extracting insights from vast and complex healthcare data.[1] Public health datasets encompass a diverse range of information, including electronic health records, epidemiological data, images data[2]; where the traditional analytical approaches have limited capabilities to uncover hidden relationships and patterns within these datasets due data volume and complexity.[3]

Deep learning applications have helped medical professionals through several areas such as symptom monitoring, diseases and outbreak predictions, patients profiling, patients' disease history, predictive modeling, and decision support systems. Deep learning could revolutionize public health by providing insights and surveillance tools to prevent and treat diseases.

Health-related public data provides valuable resources for researchers. Wanga et al.[4] mentioned 28 of those health- related repositories such as OCHIN which is "a collaborative, member-based network of federally qualified health centers and similar organizations, which delivers primary health care to vulnerable populations".[5] It is the largest network of CHCs using a single instance of the Epic EHR.[6] It has more than 500 community health centers (CHCs) in its network and serves approximately 2.8 million patients.[7] High-dimensional EHR data provides a rich source of information for machine and deep learning algorithms, allowing them to capture complex patterns and relationships between different variables. Deep learning and high-dimensional EHR data have the potential to improve healthcare outcomes by enabling the development of more accurate and personalized predictions and decisions, leading to improved patient care and better population health management.

This systematic review aims to explore the scope and limits of cutting-edge deep learning techniques that researchers are using for predictive analytics in public health data and to review associated issues and trends, such as challenges, in this area of research. Specifically, in this review we tried to answer the following questions: (1) What are major deep learning and predictive modeling approaches discussed in public health research papers? (2) What are some of the significant research trends and associated issues in this area of research?

## Preliminary review of the literature

We conducted a preliminary SLR to learn about the general popular public health datasets that are available and used frequently in research papers. Based on this initial research, we focused on the three terms combined “public health datasets” in one source, Google Scholar. The initial search from Google Scholar for those three terms together retrieve around 750 results, excluding patents. We focus on two main aspects searching through those results:

- What are the main subjects discussed in those papers?
- What are the main public health datasets mentioned in those papers?

## *Preliminary SLR: Main subjects*

*Visualization.* Visualization is a large theme in surveyed research papers related to public health datasets. Some of the focused visualization subjects include: geospatial,[8–11] interactive data visualization,[12–15] customizable visualization,[16] and visualization libraries.[17] Research indicated that visualization is not as advanced in health applications as compared with other scientific disciplines despite its significant positive influence on better understanding data in those applications.[12]

While most of medical or health visualization focused on images (e.g.[10,18–21]), yet few research papers discussed text visualization, e.g. medical documents.[22]

*Health insurance issues and analysis.* Health information concerns from privacy perspectives span from personal, to health and also financial information about patients. Public and private health providers can have access to all those categories of sensitive information. There are few public insurance datasets such as:

- All-payer claims databases: (e.g. cited by:[23–27]).
- : CPS ASEC.[28–30]
- HealthCare.gov datasets.[31–33]
- Primary Care Reimbursement Service.[34,35]

*Public health datasets and predictive analytic.* Predictive analytic is a major tool for the evaluation and analysis of health datasets. Among all different public datasets and their focuses, machine learning and predictive analytics are on top of the researched subjects.

## *Preliminary SLR: popular public health datasets*

There are several public health datasets that are available online. The nature of variables and details in those datasets can vary from privacy and user information perspectives. Open data is a new public health model, where previously only reports or outputs in certain formats are provided in public, while raw data can be provided upon request.[36] The formats and structure for those datasets are not consistent due to several factors. Public health datasets may have to deal with different state and federal regulations specially in relation to privacy and sensitive information.

We include below a comprehensive list of some of the public health datasets that are frequently used in public health research. Most of the datasets are publicly available, but there are a few that may not be fully public or may have restrictions.

- US Census: General use popular important public dataset in US. It contains a lot of important information about people lives. Census data are used to distribute funds for government programs such as Medicaid, and for planning the locations of schools, roads, and other public facilities. Census data are also used to track population trends over time as well. In the preleiminary SLR, some papers reported using health-related data from US Census (e.g.[37]).

- OCHIN (Oregon Community Health Information Network) dataset (https://ochin.org/data-exchange). OCHIN is used in several health studies and publications and not fully publicly available.[38] OCHIN is a collaborative, member-based organization of federally qualified health centers and similar entities.
- The Uniform Hospital Discharge data set (UHDDS) with Restricted access.[39]
- The Minimum data set for long term care (MDS).
- The Health Plan Employer Data and Information Set (HEDIS) with Restricted access.
- Healthdata. gov which incorporates more than a century of US healthcare data.
- The World Health Organization.
- Data. gov which includes thousands of data sets which, among others, include health, public safety, and scientific research data sets.
- The Human Mortality Database (HMD).
- OpenFDA, launched by the U.S. Food and Drug Administration.
- Medicare. gov
- Canadian Community Health Survey (CCHS).
- National Health Interview Survey (NHIS).
- Mississippi Youth Risk Behavior Survey (YRBS).
- Behavioral Risk Factor Surveillance System (BRFSS).
- National Health and Nutrition Examination Survey (https://www.cdc.gov/nchs/nhanes/index.htm/).
- Medical MNIST (https://medmnist.com/).
- Medical imaging and MRICloud with Restricted access to specific research groups.[40]
- Open EEG (https://github.com/meagmohit/EEG-Datasets/).
- Ontario Incidence Study of Reported Child Abuse and Neglect-2003 (OIS 2003): (https://cwrp.ca).

## Searching strategy to retrieve studies

After completing the preliminary SLR process, we decided to focus on public health datasets in the era of deep learning. In this focused Systematic Literature Review (SLR), the guidelines were followed from References.[41,42] The research process is divided into three phases. In the first planning phase, the stages of defining research questions, developing, and validating review protocols are covered. In the second phase, identification and selection of relevant studies, data extraction, and the information synthesis process are covered; and in the third phase, writing and validating the review are reported.

### *Plan review*

*Review protocol.* The development and validation of the review protocol highlight the searching of related articles with the appropriate keywords and the literature sources.

*Searching Keywords.* To guarantee that the review closely covers Public Health dataset and relevant to deep learning and predictive modeling, we tried to limit our search to the most relevant search term. Thus, we started with the keywords, and then we went through the following steps:

- Extracting the major distinct terms from our research questions;
- Using different spellings of the terms;
- Updating our search terms with keywords from relevant papers.

We used the main alternatives and added “OR operator” and “AND operator” to get the maximum amount of directly relevant works in the literature as shown in (Table 1). This table outlines the alternative search keywords used to ensure comprehensive coverage of relevant literature in our systematic review. Each ID represents a distinct search query combination aimed to capture various studies related to public health datasets and deep learning techniques. For example, ID 1 queries include core terms related to public health data and the fundamental deep learning methods, ID 2 expands on ID 1 by including additional advanced deep learning methods to capture more recent advancements and specialized applications of deep learning in public health contexts, ID 3 queries further broadens the scope by incorporating even more specific deep learning techniques to cover a wide array of deep learning methodologies that may be applied in the analysis of public health data. Appendix A provides the specific search formulas used in each database

*Literature resources.* Primary review studies: PubMed, Web of Science, Scopus, and Springer databases were chosen for selection of relevant articles. These databases have maximum coverage of quality articles in our domain. We restricted our search to research articles published in English and in peer-reviewed journals or conferences available from the inception of each database through 2023.

## Conduct review

In this phase, we conducted the review according to the research questions, keywords, and protocols. This phase mostly emphasizes the inclusion and exclusion of articles, according to (Table 2).

**Table 1.** Searching keywords alternatives.

| ID | Keywords |
|---|---|
| 1 | (“Public health data” OR “public health dataset” OR “EHR” OR “electronic health record”) AND (“deep learning” OR “representation learning” OR “neural network” OR “convolutional neural network” OR “con- vNet” OR “CNN” OR “recurrent neural network” OR “RNN” OR “long short-term memory” OR “LSTM” OR “generative adversarial network” OR “GAN” OR “autoencoder” OR “restricted Boltzmann machine” OR “deep belief network” OR “DBN”) |
| 2 | (“Public health data” OR “public health dataset” OR “EHR” OR “electronic health record”) AND (“deep learning” OR “representation learning” OR “neural network” OR “convolutional neural network” OR “con- vNet” OR “CNN” OR “recurrent neural network” OR “RNN” OR “long short-term memory” OR “LSTM” OR “generative adversarial network” OR “GAN” OR “autoencoder” OR “restricted Boltzmann machine” OR “deep belief network” OR “DBN” OR “deep reinforcement learning” OR “DRL”) |
| 3 | (“Public health data” OR “public health dataset” OR “EHR” OR “electronic health record”) AND (“deep learning” OR “representation learning” OR “neural network” OR “convolutional neural network” OR “con- vNet” OR “CNN” OR “recurrent neural network” OR “RNN” OR “long short-term memory” OR “LSTM” OR “generative adversarial network” OR “GAN” OR “autoencoder” OR “AE” OR “restricted Boltzmann machine” OR “deep belief network” OR “DBN” OR “deep reinforcement learning” OR “DRL” OR “gated recurrent units” OR “GRU” OR “pre-trained model” OR “transfer learning” OR “graph neural network” OR “feed-forward neural network”) |

**Table 2.** (a) Inclusion criteria description and (b) exclusion criteria description.

| Category | Criteria |
|---|---|
| (a) Inclusion | • The research was relevant to public health datasets<br>• The research was directly related to the data<br>• The research was related to deep learning applications<br>• The research was conducted using deep learning techniques related to various data types<br>• For duplicate publications of the same study, the newest and most complete one was selected. This is recorded for only one study whose related work appeared two times<br>• The research is a journal article or a conference proceeding |
| (b) Exclusion | • Studies that were irrelevant to public health datasets and deep learning domain were skipped |

Generally, this study focused on peer-reviewed publications that was relevant to deep learning and directly related to the data, applied deep learning techniques used to analyze data available from public health datasets. Studies that were irrelevant to public health datasets and deep learning domain were skipped.

*Study selection.* As mentioned in section 2, we conducted a primary review of research articles using various databases, including PubMed, Web of Science, Scopus, and Springer. Our findings revealed that Scopus yielded the highest number of relevant articles among these databases. Specifically, approximately 45.80% of Springer's unique records, 59.16% of Web of Science's unique records, and 84.29% of PubMed's unique records were also included in Scopus. Consequently, we have chosen to focus our analysis primarily on Scopus results. The whole process of Scopus study selection is illustrated in (Figure 1). A total of 2545 articles appeared in the online search. By applying filtration with title, keyword, inclusion, and exclusion criteria, a total of 2534 papers were short-listed. Inclusion and exclusion criteria are defined in (Table 2). Among them, 530 articles were excluded; 200 articles were found irrelevant to deep learning, and 330 were found irrelevant to Public Health. At the end, 2004 articles are kept in the list after going through full reading. To ensure the quality of our selection, we assessed the remaining articles with a focus on identifying studies that are not only more pertinent and comprehensive but also address all the core research questions. These primary inquiries include whether the studies center around public health data and if they dig into the essential deep learning techniques, such as neural networks, CNNs, RNNs, LSTM, and GANs, particularly in the context of large datasets.

The article selection process involved three researchers: Authors "X", "B", and "C". The 2545 articles identified from the initial search were equally divided among the three researchers for screening. Each researcher independently reviewed their assigned articles based on predefined inclusion and exclusion criteria. To ensure consistency, a subset of 10% of the articles was randomly selected and reviewed by all three researchers, resulting in a Cohen's kappa statistic of 0.78, indicating substantial agreement. Authors "A" and "F" contributed to other essential tasks in the study but were not involved in the article screening process. Cohen's kappa statistic, a measure of inter-rater reliability, was calculated using the formula $\kappa = P_0 - Pe/1\text{-}Pe$, where $P_0$ is the observed proportion of agreement among raters and Pe is the expected proportion of agreement by chance. This confirming the reliability of the selection process. Collaborators "G" and "I" contributed to the data formatting process, ensuring that the data collected was in the appropriate formats for analysis and reporting.

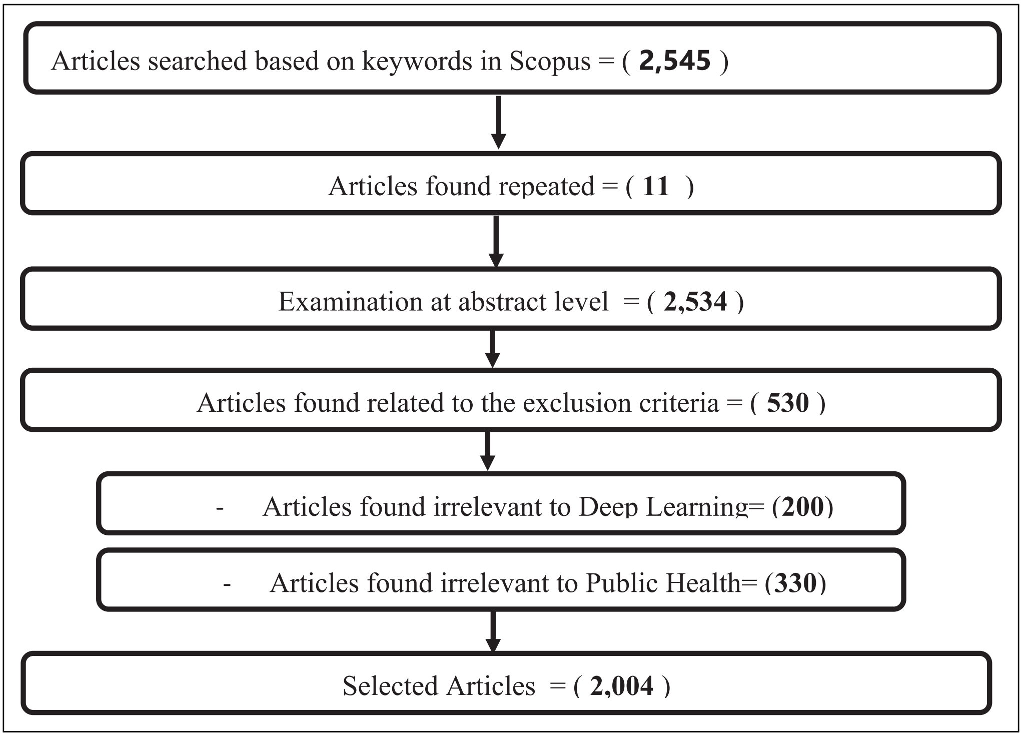



**Figure 1.** Study selection process.

*Report review.* At this stage, the extracted data were aggregated to answer the research questions. For our research questions, we used the narrative synthesis method. Accordingly, we used tables and charts to present our results. The guidelines of References[41] were closely followed in the reporting of results in the Statistical Analysis section below. The provided guidelines emphasize the importance of a thorough methods overview, calling for clear and detailed descriptions of research methods encompassing data collection, processing, and analysis to ensure reproducibility. Additionally, they stress the significance of presenting research results coherently through structured and organized formats, including visual aids like figures, tables, and graphs for improved clarity. Furthermore, the guidelines highlight the need for offering a comprehensive overview of how the research findings are applicable to real-world contexts and contribute to practical applications and services.

## Statistical analysis

In this section, we provide few stat figures and tables. (Figure 2) shows the years of publications for the selected papers. The Figure indicates the recent increase of interests in this field with a significant growth starting the year 2016. The same finding can be noticed observing the top 10 cited publications where they are between the years 2016-2019.

The top different disease categories are highlighted in (Figure 3) that used public health datasets and deep learning techniques.

The top different data formats or sources of data studied are presented in (Figure 4) that used public health datasets and deep learning techniques.

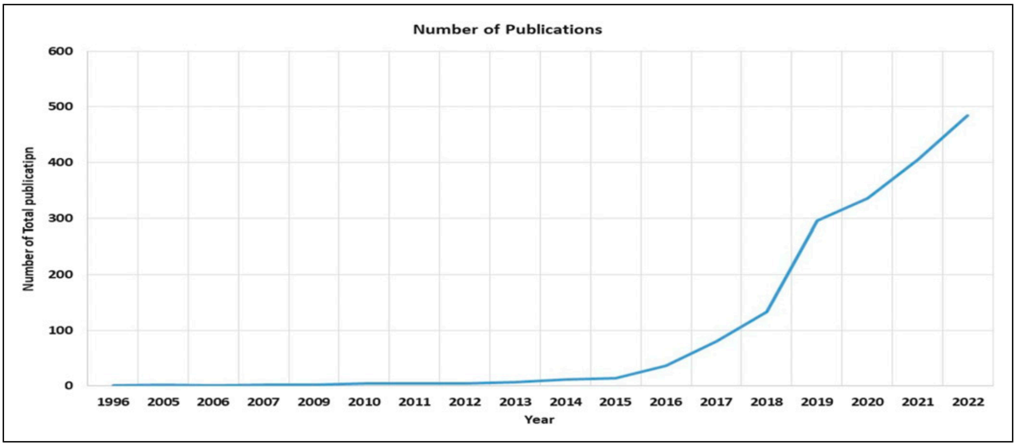



**Figure 2.** The Number of publications per year.

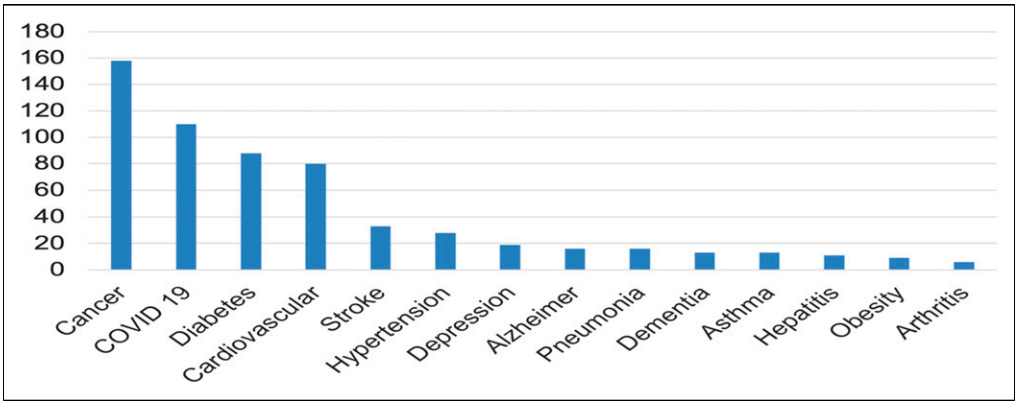



**Figure 3.** The Top Common Disease Categories Studied using Public Health Datasets.

(Tables 3 and 4) shows the distribution of our included publications based on their type. The overall included are about 2000 with roughly 1/4-th conference papers and 3/-4th journal articles.

(Table 5) shows for journal articles number of papers published in major publishers.

(Table 6) shows the different Deep learning types presented in the literature reviewed. Those are based on citing those types only within the "Abstract" section of the paper. While it captures a broad range of models, it is not exhaustive. Additional models may be identified in specific studies or emerging research.

(Table 7) shows the Top Researchers Names and Number of Publications per Researcher.

(Table 8) shows the top Journals and the Number of Publications shown in each Journal.

## Deep learning and health informatics

We noticed a research trend that started around a decade ago, employing deep learning models in health-related knowledge learning and prediction. Earlier in the statistics section, we showed that

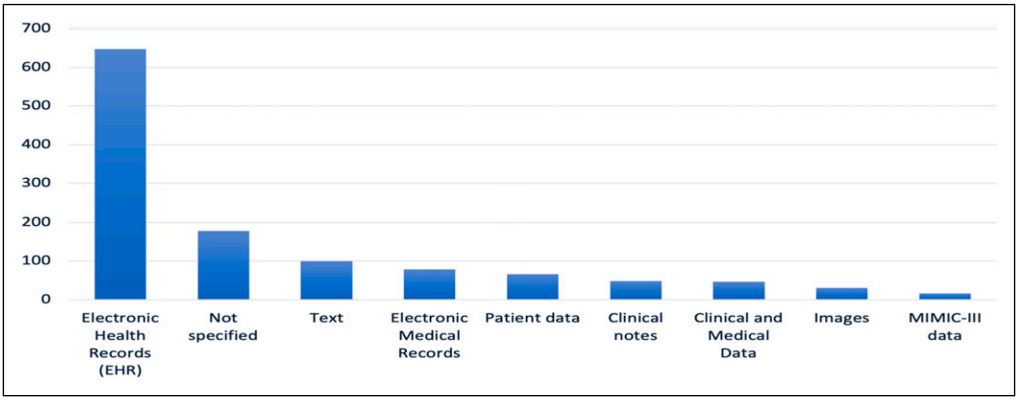



**Figure 4.** The top data formats or sources of data studied.

**Table 3.** Distribution of papers based on publication type. (Table 4) shows top 10 cited papers in our collected list. The list shows that most of the top cited papers focus on using deep learning approaches in health related data.

| Publication type | Number | Status |
|---|---|---|
| Journal articles | 1408 | Included |
| Book or book chapter | 91 | Excluded |
| Conference paper | 596 | Included |
| Conference review | 81 | Excluded |
| Review | 246 | Excluded |
| Others (note, erratum, editorial, short survey, letter) | 112 | Excluded |
| Total included | 2004 | |

**Table 4.** Top 10 cites articles.

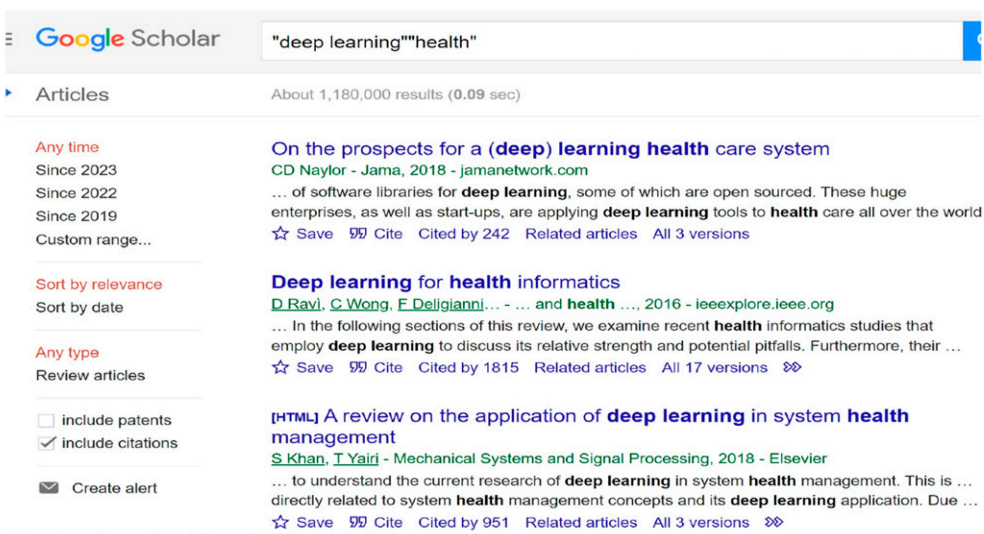
| Article | Year | Subject | Current citations no. (scopus) |
|---|---|---|---|
| 43 | 2018 | Deep learning for healthcare | 1153 |
| 44 | 2018 | Deep learning for healthcare | 1076 |
| 45 | 2018 | Deep learning in biology and medicine | 968 |
| 46 | 2016 | Deep patient | 864 |
| 47 | 2017 | Cardiovascular risk prediction | 626 |
| 48 | 2017 | Deep EHR | 557 |
| 49 | 2016 | Attention mechanisms and health care data | 536 |
| 50 | 2017 | RNN and heart failures detection | 458 |
| 51 | 2019 | Acute kidney injury prediction | 454 |
| 52 | 2017 | Graph-based attention and healthcare learning | 359 |

most of the top 10 cited papers in this review utilized deep learning models. This is not uncommon or surprising as deep learning has the similar hype in many other fields as well. (Figure 5) shows more than million candidate articles in Google Scholar that include the two keywords "Deep learning" and "Health". More than half of those are after 2015.

## Deep learning Models for health public data

Deep learning models (DLMs) or deep neural networks can be used to solve several urgent and important real-world health-related problems on large-scale datasets. The need of effective data analytics in health informatics for better decision making is an ongoing challenging issue.[53] Our analysis focuses on deep learning models that have been extensively validated and are widely adopted in current health data applications. Although newer architectures like transformers and large language models are gaining traction, their application in health informatics is still emerging and is therefore beyond the scope of this study. Our previous sections shows some of the popular DLMs in health informatics (e.g. LSTM, CNN, GAN, AE, RNN, GRU, GNN and DBN). We cover a summary of each one of those models in health informatics in the next sections.

*Long short-term memory networks: LSTM.* LSTM is a variant of RNN designed to alleviate problems related to long- term dependencies.[54] LSTMs are well-suited for processing sequences of data with long-range dependencies. They can capture information from earlier time steps and remember it for a more extended period, making them effective for tasks like natural language processing (NLP) and time series analysis.[55] LSTM can refer to the information of the previous unit when processing the information of this unit, but it does not refer to the information of the subsequent unit.[56] LSTM in health care research have been used to predicting life expectancy,[57] healthcare usage based on medical records,[58] and machine health monitoring.[59] LSTM networks' ability to handle long-range



Deep Learning and Health search on Google Scholar

**Figure 5.** Deep Learning and Health search on Google Scholar.

dependencies in health data makes them effective for modeling lifespan estimations. However, LSTMs can be computationally intensive and prone to overfitting if not properly used. Despite these challenges, their capacity to capture temporal dynamics and manage missing data proves advantageous in health applications like forecasting patient lifespans from historical records.

*Convolution neural networks: CNN.* CNN is an advanced version of DNNs which arranges its neurons in three basic dimensions. The main advantage of CNN is that it does both feature extraction and classification, automatically.[60–63] CNN is used in many image detection and classification in many domains and applications including health related image classification applications (e.g. medical images classification,[64,65] lung cancer,[66] Ultrasound images,[67] breast tumors,[68] skin cancer,[69] *tuberculosis*,[70] and many others.

*Generative adversarial networks: GAN.* GAN is an unsupervised learning method consisting of two neural networks that are competing with each other in trying to learn the pattern of its input by itself and generate new samples with similar characteristics to the real data. One key advantage of GAN is that the model is able to learn the underlying data distribution. Additionally, the existence of a discriminator enforces the generator in generating accurate results. GANs can also reduce the training data needed for training, which is suitable for medical imaging due to the shortage of data.[71] In health data GAN is used in many applications such as: ECG denoising,[72] chest X-ray images classification,[73] tomato leaf disease identification,[74] neuro imaging and clinical neuroscience.[75]

*Auto-encoders: AEs.* Autoencoders are unsupervised neural networks used for representing the structural data by data compression. They have been used for the purpose of reducing dimensionality and also detecting network anomalies in large datasets. Autoencoders benefit from the many technical advances made in the last years in the development of DNNS. One of the main advantages of autoencoders is their ability to learn quickly on unlabeled data.[76]

In health informatics, AEs are used to learn Cardiac-ICU and Neuro-ICU.[77] They are also used in medical imaging,[76] compressing patients features,[78] RNA target prediction,[79] detection of plants diseases,[80] and tumor identification and detection.[81–84]

*Recurrent neural networks: RNNs.* Recurrent neural networks are neural networks with directed cycles that are particularly suitable to process time series data.[85,86] They are designed to process sequences, and can remember or forget information from earlier steps when processing later steps in a sequence.[87] RNNs allow previous outputs to be used as inputs while hiding the state.[88] RNNs are used in health informatics in different examples or applications such as, future diagnosis of heart failure,[89,90] predicting health by suicide,[87] predicting hospital readmission of diabetic patients,[91] predicting HIV status,[92] and end-to-end classification of cell-cycle stages.[93]

*Gated Recurrent Units: GRU.* Gated Recurrent Units are special variants of LSTMs that merge the forget and input gates into a single update gate resulting in a simpler more efficient model than standard LSTM models.[94] GRUs do not incorporate memory cells, they have reset gates instead. These reset gates allow the hidden state to leave the unimportant information and thus, focusing on the quality of content.[95] In health informatics, GRUs are used in life estimation,[96] patient case similarity evaluation,[97] flu forecasting,[98] clinical event prediction,[99] stroke volume estimation,[100] prediction of Parkinson's disease,[101] patients trajectory prediction,[102] predicting blood pressure,[103] and human emotions recognition.[104]

*Graph Neural Networks: GNN.* Graph neural networks (GNNs) are models for representation of learning graphs, they are built upon the multi- layer perceptrons (MLP) architecture with additional message passing layers to allow features to flow a cross nodes.[105] They are increasingly becoming popular in various fields such as computer vision, computational biology and chemistry, where data are naturally explained by graphs.[106] Graph Neural Networks are used in health informatic research and applications such as: classification of mammograms,[107] estimating the state of epidemics spreading,[108–110] identification of providers with similar risk profiles in healthcare claims,[111] virus-human protein-protein interaction prediction,[112] modelling the bioactivities of ligands targeting orphan G protein-coupled receptors,[113] drug-target affinity prediction,[114,115] and prediction of circRNA-miRNA association.[116]

*Deep belief networks: DBN.* Deep belief nets are probabilistic generative models that are composed of multiple layers of stochastic, latent variables.[117] The main advantages of deep belief networks are their efficient learning algorithm, their ability to extract high dimensional features and to represent them in low dimensions, as well as their associative memory ability.[118] Deep belief nets are used in health informatics research and applications such as: seizure detection,[119] heart failure prediction,[120] detecting Neuro-degenerative Disease from MRIs,[121] emotion recognition,[122] and the classification of malaria-infected and uninfected images.[123]

## Deep learning applications in health public data

Applying Deep learning (DL) techniques in health public data has revolutionized how vast amounts of healthcare information are analyzed and interpreted in many health applications.[124] These techniques contribute to several applications that help in assessing population health trends, optimizing resource allocation, and enhancing public health strategies.[125] In this section, we outline the most crucial applications where the Deep Learning (DL) approaches are utilized and elucidate how Deep Learning (DL) approaches contribute to the enhancement of these applications and we discuss how the DL techniques hold immense promise for reshaping healthcare in public health datasets.

*Disease prediction.* Deep learning applications in the health public provide invaluable tools for accurate disease prediction, offering unprecedented insights into early detection, improving prevention strategies, and disease progressive assessment. With the development of DL techniques, the accuracy of medical disease prediction is continuously improved.[126] Several works reviewed the deep learning algorithms for disease prediction as.[127] In[127], Yu et al. showed the use of Structured data algorithms and Unstructured data algorithms to predict several types of disease. The DL techniques are extensively used for Chronic Disease prediction and for Epidemic Outbreak Prediction. The application of deep learning in infectious disease prediction has proven the efficiency in the surveillance of outbreaks.

*Drug discovery.* Drug discovery research generally aims to identify molecules with therapeutic effects against specific diseases The primary objective of drug discovery research is typically to identify molecules with therapeutic effects targeted at specific diseases.[128] DL techniques have emerged as a powerful tool being used in all stages of drug discovery and development. Several research utilized DL techniques in Deep learning in drug discovery.[129] For instance, In[130], Bai et al. proposed a tool called MolAICal that employs a two-component approach to design 3D drugs: the first component utilizes deep learning (DL) and a genetic algorithm trained on FDA-approved drugs, and the second component integrates molecular docking with a DL. Popova et al.[131] designed a deep reinforcement learning-based approach called ReLeaSE utilized for drug design. ReLeaSE integrates two deep neural networks (DNN), known as generative and predictive. The generative model was used to produce new compounds, and the predictive model was used to predict the properties of the compound.[131] The SLR in[129] presented the integration between the recent DL technologies and several drug discovery applications.

*Image analysis.* Medical image analysis plays a crucial role in healthcare that enhances early and accurate diagnosis and treatment planning and monitoring. DL techniques have revolutionized medical image analysis, offering unparalleled capabilities in interpreting imaging data and improving diagnostic decision-making. DL techniques are used for several medical image types and utilized in several disease diagnostic processes. Suzuki et al.[132] overviewed the area of deep learning in medical imaging and included the deep learning effects in the medical imaging analysis field, the focus of this overview is two models: a massive-training artificial neural network (MTANN) and a convolutional neural network (CNN) their applications to medical imaging. Several systematic reviews have already discussed the current approaches of DL in Medical image analysis as.[133,134] Other research focuses on using medical images and DL to study specific diseases, for example, the research in[135] determined the types of medical images utilized in the classification of COVID-19 and explored the employed methodologies in the COVID-19 classification tasks.

*Health monitoring.* The utilization of public datasets with deep learning techniques in health monitoring provides many applications with powerful tools for advancing personalized healthcare solutions.[136] The public health datasets, are comprised of diverse sources and kinds of data such as electronic health records and population health statistics that provide a comprehensive foundation of information for the health monitoring systems.[137,138] Deep learning techniques with their capabilities in handling large and complex datasets, bring unparalleled analytical prowess to extract meaningful insights from the available public health data.[139,140] By utilizing deep learning techniques, researchers can develop predictive models for health monitoring offering a nuanced understanding of health dynamics.[141] combining the accessibility of public data with the analytical power of deep learning holds significant potential to enhance the precision of health monitoring which contributes to early disease detection, supports the development of more targeted health monitoring systems, and provides effective healthcare interventions on a population scale. Sujith et al. in[136] presented a review of the smart health monitoring frameworks that utilized deep learning and machine learning techniques and algorithms for enhancing the handling of the healthcare data generated in Smart health monitoring.

*Medical research.* The availability of many public datasets and the strength of DL techniques fosters a collective approach to problem- solving in medical research.[142] This collaborative paradigm accelerates the discovery, leading to breakthroughs in disease understanding, drug development, and personalized medicine. The utilization of deep learning in public datasets marks a paradigm shift in medical research, where open access to data enhances medical innovation,[143] and drives advancements that have the potential to revolutionize healthcare practices and outcomes.[144]

## *Common disease categories and techniques studied*

In this paper, different disease categories studied are discussed. The most common disease categories studied are presented in (Table 9). The top common disease categories studied include

Hepatitis, Cardiovascular, Cancer, Stroke, Pneumonia, Diabetes, Alzheimer, Dementia, Obesity, Asthma, Hypertension, Arthritis, COVID-19, and Depression. The most common techniques used among these disease categories include Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bi-directional Long Short-Term Memory (BiLSTM), Deep Neural Network (DNN), Perceptron Neural Network Autoencoder Networks, BERT (Bidirectional Encoder Representations from Transformers) Deep Auto-encoder Artificial Neural Network, Longitudinal Deep Learning Model, MLP (Multilayer Perceptron), and Transfer Learning.

## Research trends and challenges: EXAMPLES

One of the promising research trends in the scope of this paper is obviously to use deep learning algorithms to improve the accuracy of data analysis tasks. Another trend that we discussed already is the usage of different deep learning models and settings in health information data analysis.

We cite examples of research trends in health informatics datasets utilizing machine and deep learning methods.

### *The integration of different data sources or types*

In (Figure 4), data formats or sources of data are presented. Many recent research publications integrate data from different sources and types. So far, those are the different types of data used :

- Electronic Health Records Structured data
- Images
- Text (unstructured)
- Biological sequences
- ECG, EEG, MEG, etc signals
- Patient, clinical and medical data

In the last couple of decades the amount of data used in data analytic research increased rapidly. Additionally, there was a significant increase in the usage of unstructured data or text that is coming from several input sources such as: web-based contents, and documents, patients and medical records, and so on. Such spectrum of data is not only different in terms of those types mentioned earlier, but also in many other aspects. For example, how you deal with combining data from social networks that may have a lot of misinformation with data from hospitals, clinics and medical doctors?.

In addition, the presence of a significant quantity of data in various unspecified formats or originating from diverse sources in reviewed studies necessitates adherence to the principles of FAIRness in open and responsible research. This concept emphasizes the importance of data being Findable, Accessible, Interoperable, and Reusable, ensuring that research data management is conducted efficiently and effectively.

### *Deep learning research and reproducibility issues*

The issue of reproducibility is an ongoing problem in deep learning research in general. Many research papers present models, settings and discuss results and accuracy, however, with no solid public data about these details such results can be hardly verifies.

Technical reproducibility refers to the ability to reproduce a paper's results precisely as presented in the paper under the same condition.[145,146] Reproducibility is also critical for deep learning research, whose goal is to develop algorithms to reliably solve complex tasks at scale.[147] Reproducibility is an essential metric when assessing for possible clinical deployments of evaluated algorithms. of any algorithm.

## *Explainable-AI*

Explainable-AI is a rapidly emerging research area that focuses on annotating model's decisions as well as decision- making characteristics.[148,149] It is used to describe an AI model, its expected impact and potential biases and helps characterize model accuracy, fairness and transparency.[150] Explanations serve as a bridge between humans and AI systems.[151] The EU's guidelines on AI robustness and explainability[152] emphasize three key elements for the proper utilization of AI: transparency, reliability, and safeguarding of individual data. There are many popular frameworks of explainable AI including SHAP, LIME, MUSE, PDP, CE, etc. In health data Explainable-AI is important to interpret results and utilize their findings. Explainable AI is critical in health informatics helps avoid practical consequences.[153,154] It can also improve fairness and remove bias in decision making.[155] "Positional SHAP" (PoSHAP),[156] is proposed to interpret models trained from biological sequences by exploiting SHAP to generate positional model interpretations.

## *Patient embedding learning*

Patient embedding learning has several potential applications in healthcare, including:

- Predictive modeling: Patient embeddings can be used as input features in predictive models, allowing for more accurate predictions of patient outcomes, such as hospital readmission or disease progression.
- Improved patient clustering: Patient embeddings can be used to cluster patients into subpopulations with similar medical histories and demographics, improving population health management and reducing health disparities.
- Personalized medicine: Patient embeddings can be used to predict the likelihood of a patient responding to a specific treatment, allowing for the development of more personalized treatment plans.

## *Imbalanced data*

Handling of imbalanced health-related public data poses a considerable challenge for machine learning and deep learning approaches.[157] This imbalance data can impact model training, where the algorithm may prioritize the majority class. To overcome this challenge, employing machine learning and deep learning techniques in health public dataset should explore several imbalance handling strategies,[158] such as oversampling the minority class,[159] or utilizing techniques like Synthetic Minority Over-sampling Technique (SMOTE).[160,161]

Deep learning techniques offer a promising approach for handling imbalanced health datasets.[162] As examples, Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) excel in handling and capturing spatial and temporal dependencies in medical data.[163,164] Moreover, transfer learning Techniques where pre-trained deep learning models are fine-tuned for specific health challenges, can be particularly advantageous in scenarios with limited imbalanced labeled data.[165]

### Data ownership

Health data ownership and privacy concern are debated issues in the context of healthcare systems.[166] The current use of Big Data in healthcare faces challenges due to a lack of rules that manage data ownership and responsibilities.[167] These challenges arise when considering the collaborative and constructed nature of clinical data.[168] In[169] et al suggested discussing the use of Blockchain as shown in[170] to offer a new way to empower patients to control their data while ensuring secure and confidential data sharing through encrypted transactions.[168] However, despite Blockchain techniques' potential benefits in enhancing security and privacy, the widespread adoption of blockchain for healthcare data transfer and storage is still pending.[171]

## Discussions, limitations of the study and recommendations for future work

In conclusion, this systematic literature review delves into the dynamic landscape of public health research within the era of deep learning. The synthesis of existing studies has provided a comprehensive overview of the applications, challenges, and advancements in utilizing deep learning methodologies for the analysis of diverse datasets in public health contexts. The findings underscore the potential of deep learning to revolutionize the way we approach and address public health issues, offering innovative solutions for data-driven decision-making.

Throughout the review, common themes such as predictive modeling, disease surveillance, and risk assessment have emerged as focal points of deep learning applications in public health. However, the review also highlights challenges such as data privacy concerns, interpretability of models, and the need for standardized methodologies. These challenges underscore the importance of a thoughtful and ethical integration of deep learning into public health research.

As the field continues to evolve, future research should focus on addressing the identified challenges, exploring novel applications, and fostering interdisciplinary collaboration between data scientists, public health experts, and policymakers. By doing so, we can harness the full potential of deep learning to advance our understanding of public health phenomena, ultimately contributing to more effective strategies for disease prevention, health promotion, and healthcare delivery. This review serves as a valuable resource for researchers, practitioners, and policymakers navigating the intersection of public health and deep learning, providing a roadmap for future investigations in this rapidly evolving field.

There are a few limitations of this review to consider. We were unable to review all of the initial results due to resource constraints. We conducted a primary review of research articles using various databases, including PubMed, Web of Science, Scopus, and Springer. Our findings revealed that Scopus yielded the highest number of relevant articles among these databases. Therefore, our screening process could be considered a convenience sample which optimizes for recent research and highly visible research based on citation frequency. A future review could take the time to review all initial search results rather than adopting our visibility approach.

## Conclusions

The employment of Deep Learning Models in public health settings continually proves to play a transforming role that can advance multiple areas particularly those of disease prediction, surveillance and risk assessment. Despite the challenges highlighted in this study, Deep learning methods hold promising future in improving healthcare and disease prevention. This study serves as a helpful resource for future research to address these challenges and work as a guideline for further study in this field.

## Author contribution

Rand Obeidat, Izzat Alsmadi, and Qanita Bani Baker: These three researchers were responsible for the initial screening of the 2,545 articles identified from the initial search. Each researcher independently reviewed their assigned articles based on predefined inclusion and exclusion criteria. Aseel Al-Njadat and Sriram Srinivasan: These authors contributed to other essential tasks as analyzing, reporting and writing of the paper in the study but were not involved in the article screening process. Godswill Ashong and Ifeanyi Osigwe: These authors contributed to the data formatting process, ensuring that the data collected was in the appropriate formats for analysis and reporting.

## Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by grant funding from the National Science Foundation (NSF) for the collaborative project titled "CAP: Expanding Use-Inspired AI Research and Instruction Across Campus," Award #2401658.

## ORCID iDs

Rand Obeidat https://orcid.org/0000-0001-8271-3829
Qanita Bani Baker https://orcid.org/0000-0003-4722-6632
Aseel Al-Njadat https://orcid.org/0009-0009-6935-4759

## References

1. Ravì D, Wong C, Deligianni F, et al. Deep learning for health informatics. *IEEE journal of biomedical and health informatics* 2016; 21(1): 4–21.
2. Shortreed SM, Cook AJ, Coley RY, et al. Challenges and opportunities for using big health care data to advance medical science and public health. *Am J Epidemiol* 2019; 188(5): 851–861.
3. Ngiam KY and Khor W. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* 2019; 20(5): e262–e273.
4. Wang Y, Kung L, Wang WYC, et al. An integrated big data analytics-enabled transformation model: application to health care. *Inf Manag* 2018; 55(1): 64–79.
5. DeVoe JE, Likumahuwa S, Eiff MP, et al. Lessons learned and challenges ahead: report from the OCHIN Safety Net West practice-based research network (PBRN). *J Am Board Fam Med* 2012; 25(5): 560–564.
6. Hatch B, Tillotson C, Huguet N, et al. Implementation and adoption of a health insurance support tool in the electronic health record: a mixed methods analysis within a randomized trial. *BMC Health Serv Res* 2020; 20(1): 1–12.
7. Clark KD, Woodson TT, Holden RJ, et al. Translating research into agile development (TRIAD): development of electronic health record tools for primary care settings. *Methods Inf Med* 2019; 58(01): 001–008.
8. Sopan A, Noh AS-I, Karol S, et al. Community Health Map: a geospatial and multivariate data visualization tool for public health datasets. *Govern Inf Q* 2012; 29(2): 223–234.
9. Hanchette CL. *Geographic information systems*. Springer, 2014.
10. Yasobant S, Vora KS, Hughes C, et al. Geovisualization: a newer GIS technology for implementation research in health. *J Geogr Inf Syst* 2015; 7(01): 20.

11. Pereira G, Stankus L and Manzo R. (n.d.). HealthDashboard: A urban public health geospatial visualization platform. Retreived from https://www.academia.edu/79602941/HealthDashboard_A_Urban_Public_Health_Geospatial_Visualization_Platform.
12. Shneiderman B, Plaisant C and Hesse BW. Improving healthcare with interactive visualization. *Computer* 2013; 46(5): 58–66.
13. Yu Z, Pepe K, Rust G, et al. Patient-provider geographic map: an interactive visualization tool of patients' selection of health care providers. In: *2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC)*. IEEE, 2017, pp. 1–8.
14. Narayan KA and Nayak MSDP. Need for interactive data visualization in public health practice: examples from India. *Int J Prev Med* 2021; 4: 12.
15. Gill IS. *Interactive visualizations for two large public health datasets*. PhD Thesis. University of British Columbia, 2023.
16. Soman S, Kumar S and Kumar S. Healthanalytic: a concept application for customizable visualization and analysis of health informatics datasets. In: *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2016, pp. 1033–1038.
17. Mundargi ZK, Patel K, Patel A, et al. An automated data visualization website using Python and plotly. In: *2023 International Conference for Advancement in Technology (ICONAT)*. IEEE, 2023, pp. 1–4.
18. Yi Q, Hoskins RE, Hillringhouse EA, et al. Integrating open-source technologies to build low-cost information systems for improved access to public health data. *Int J Health Geogr* 2008 2008; 7: 1–13.
19. Del Corral J, Blumenthal MB, Mantilla G, et al. Climate information for public health: the role of the IRI climate data library in an integrated knowledge system. *Geospatial Health* 2012 2012; 6(3): S15–S24.
20. Gotz D, Caban JJ and Chen AT. Visiual analytics for healthcare. *Healthcare Data Analytics*. 2015; 403–431.
21. Chhabra M and Kumar R. A smart healthcare system based on classifier DenseNet 121 model to detect multiple diseases. *Mobile radio communications and 5G networks: proceedings of second MRCN 2021*. Springer, 2022, pp. 297–312.
22. Kenei JK, Opiyo ETO, Moso JC, et al. Clinical documents summarization using text visualization technique. *Dedan Kimathi University of Technology Official Repository*. 2018. http://41.89.227.156:8080/xmlui/handle/123456789/767.
23. Love D, Custer W and Miller P. *All-payer claims databases: state initiatives to improve health care transparency*. Commonwealth Fund New York (NY); 2010.
24. Miller PB, Love D, Sullivan E, et al. All-payer claims databases: an overview for policymakers. *Prepared for State Coverage Initiatives by the National Association of Health Data Organizations and the Regional All-Payer Healthcare Information Council*. 2010. http://www.statecoverage.org/files/SCI_All_Payer_Claims_ReportREV.pdf.
25. Peters A, Sachs J, Porter J, et al. The value of all-payer claims databases to states. *N C Med J* 2014; 75(3): 211–213.
26. Perraillon MC, Liang R, Sabik LM, et al. The role of all-payer claims databases to expand central cancer registries: experience from Colorado. *Health Serv Res* 2022; 57(3): 703–711.
27. Han A, Lee K-H and Park J. The impact of price transparency and competition on hospital costs: a research on all-payer claims databases. *BMC Health Serv Res* 2022; 22(1): 1321.
28. Hill SC, Solomon KT, Maclean JC, et al. Effects of improvements in the CPS on the estimated prevalence of medical financial burdens. *Health Serv Res* 2019; 54(4): 920–929.
29. Ruhter J, Conmy AB, Chu R, et al. Tracking health insurance coverage in 2020-2021. In: *Office of the Assistant Secretary for Planning and Evaluation, Office of Health Policy*. Washington, DC: US Department of Health and Human Services, 2021.
30. Berchick ER and Jackson H. Data processing improvements for estimates of health insurance coverage in the current population survey annual social and economic supplement. *Med Care Res Rev* 2022; 79(2): 308–316.

31. Branham DK and DeLeire T. Zero-premium health insurance plans became more prevalent in federal marketplaces in 2018. *Health Aff* 2019; 38(5): 820–825.
32. Wengle E and Blumberg LJ. Marketplace coverage enrollment by metal tier, 2016–18: trends in states using the healthcare. Gov enrollment platform. *Washington, DC: Urban Institute.* 2020.
33. Branham DK, Conmy AB, DeLeire T, et al. *Access to marketplace plans with low premiums on the federal platform*. Washington, DC: US Department of Health and Human Services, Assistant Secretary for Planning and Evaluation, 2021.
34. Hurley E, Sinnott SJ, Browne J, et al. *P23 Using national pharmacy claims data to estimate the prevalence and patterns of medication use for COPD in Ireland*. BMJ Publishing Group Ltd, 2019.
35. Doogue R, McCann D, Fitzgerald N, et al. Blood pressure control in patients with a previous stroke/ transient ischaemic attack in primary care in Ireland: a cross sectional study. *BMC Fam Pract* 2020; 21(1): 1–9.
36. Martin EG and Begany GM. Opening government health data to the public: benefits, challenges, and lessons learned from early innovators. *J Am Med Inf Assoc* 2017; 24(2): 345–351.
37. Terry RL and Fond M. Experimental US Census Bureau race and Hispanic origin survey questions: reactions from Spanish speakers. *Hisp J Behav Sci* 2013; 35(4): 524–541.
38. DeVoe JE and Sears A. The OCHIN community information network: bringing together community health centers, information technology, and data to support a patient-centered medical village. *J Am Board Fam Med* 2013; 26(3): 271–278.
39. Usdo H, Services H and others. Health information policy council; 1984 revision of the Uniform hospital discharge data set. *Fed Regist* 1985; 50: 31038–31040.
40. Mori S, Wu D, Ceritoglu C, et al. MRICloud: delivering high-throughput MRI neuroinformatics as cloud-based software as a service. *Comput Sci Eng* 2016; 18(5): 21–35.
41. Wang D, Miwa T and Morikawa T. Big trajectory data mining: a survey of methods, applications, and services. *Sensors* 2020; 20(16): 4571.
42. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *Ann Intern Med* 2009; 151(4): W–65.
43. Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: review, opportunities and challenges. *Briefings Bioinf* 2018; 19(6): 1236–1246.
44. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine* 2018; 1(1): 18.
45. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018; 15(141): 20170387.
46. Miotto R, Li L, Kidd BA, et al. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci Rep* 2016; 6(1): 1–10.
47. Weng SF, Reps J, Kai J, et al. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* 2017; 12(4): e0174944.
48. Shickel B, Tighe PJ, Bihorac A, et al. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics* 2017; 22(5): 1589–1604.
49. Choi E, Bahadori MT, Sun J, et al. Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. *Adv Neural Inf Process Syst* 2016; 1: 29.
50. Choi E, Schuetz A, Stewart WF, et al. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inf Assoc* 2017; 24(2): 361–370.
51. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 2019; 572(7767): 116–119.

52. Choi E, Bahadori MT, Song L, et al. GRAM: graph-based attention model for healthcare representation learning. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 2017, pp. 787–795.
53. Navamani TM. Efficient deep learning approaches for health informatics. In: *Deep learning and parallel computing environment for bioengineering systems*. Elsevier, 2019, pp. 123–137.
54. Hochreiter S and Schmidhuber J. Long short-term memory. *Neural Comput* 1997; 9(8): 1735–1780.
55. Srivatsavaya P. LSTM "implementation, advantages and diadvantages. *University of Twente Student Theses*. 2023. https://medium.com/@prudhviraju.srivatsavaya/lstm-implementation-advantages-and-diadvantages-914a96fa0acb.
56. Zhou L, Zhao C, Liu N, et al. Improved LSTM-based deep learning model for COVID-19 prediction using optimized approach. *Eng Appl Artif Intell* 2023; 122: 106157.
57. Beeksma M, Verberne S, van den Bosch A, et al. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. *BMC Med Inf Decis Making* 2019; 19(1): 1–15.
58. Pham T, Tran T, Phung D, et al. Predicting healthcare trajectories from medical records: a deep learning approach. *J Biomed Inf* 2017 2017; 69: 218–229.
59. Zhao R, Wang J, Yan R, et al. Machine health monitoring with LSTM networks. In: *2016 10th international conference on sensing technology (ICST)*. IEEE, 2016, pp. 1–6.
60. Bayoudh K, Hamdaoui F and Mtibaa A. Transfer learning based hybrid 2D-3D CNN for traffic sign recognition and semantic road detection applied in advanced driver assistance systems. *Appl Intell* 2021; 51: 124–142.
61. Fu Y, Zhang Y, Gao Y, et al. Machining vibration states monitoring based on image representation using convolutional neural networks. *Eng Appl Artif Intell* 2017; 65: 240–251.
62. Hazra A, Choudhary P, Inunganbi S, et al. Bangla-Meitei Mayek scripts handwritten character recognition using convolutional neural network. *Appl Intell* 2021; 51: 2291–2311.
63. Liang P, Deng C, Wu J, et al. Compound fault diagnosis of gearboxes via multi-label convolutional neural network and wavelet transform. *Comput Ind* 2019; 113: 103132.
64. Sharma AK, Nandal A, Dhaka A, et al. Medical image classification techniques and analysis using deep learning networks: a review. *Health informatics: A Computational Perspective in Healthcare* 2021; 1: 233–258.
65. Kumar R, Kumbharkar P, Vanam S, et al. Medical images classification using deep learning: a survey. *Multimed Tool Appl* 2023; 1: 1–46.
66. Saleh AY, Chin CK, Penshie V, et al. Lung cancer medical images classification using hybrid CNN-SVM. *International Journal of Advances in Intelligent Informatics* 2021; 7(2): 151–162.
67. Kareem FQ and Abdulazeez AM. Ultrasound medical images classification based on deep learning algorithms: a review. *Fusion: Practice and Applications* 2021; 3(1): 29–42.
68. Zaychenko Y, Hamidov G and Varga I. Medical images of breast tumors diagnostics with application of hybrid CNN–FNN network. *Syst Res Inf Technol* 2018; 4: 37–47.
69. Zaman K and Maghdid SS. Medical images classification for skin cancer using convolutional neural network algorithms. *Advances in Mechanics* 2021; 9(3): 526–541.
70. Liu C, Cao Y, Alcantara M, et al. TX-CNN: detecting tuberculosis in chest X-ray images using convolutional neural network. In: *2017 IEEE international conference on image processing (ICIP)*. IEEE, 2017, pp. 2314–2318.
71. Jain S, Indora S and Atal DK. Rider manta ray foraging optimization-based generative adversarial network and CNN feature for detecting glaucoma. *Biomed Signal Process Control* 2022; 73: 103425.
72. Singh P and Pradhan G. A new ECG denoising framework using generative adversarial network. *IEEE ACM Trans Comput Biol Bioinf* 2020; 18(2): 759–764.

73. Mahmud Pranto MA, Al Asad N, Adnan Palash MI, et al. COVID-19 chest X-ray classification with augmented GAN. In: *Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021*. Springer, 2022, pp. 125–139.
74. Wu Q, Chen Y and Meng J. DCGAN-based data augmentation for tomato leaf disease identification. *IEEE Access* 2020; 8: 98716–98728.
75. Wang R, Bashyam V, Yang Z, et al. Applications of generative adversarial networks in neuroimaging and clinical neuroscience. *Neuroimage* 2023; 2023: 119898.
76. Zaychenko Y, Hamidov G and Chapaliuk B. *The application of CNN and hybrid networks in medical images processing and cancer classification*. Cambridge Scholars Publishing, 2023.
77. Sadati N, Nezhad MZ, Chinnam RB, et al. Representation learning with autoencoders for electronic health records: a comparative study. *ArXiv:1801.02961 [Cs, Stat]*. 2018. http://arxiv.org/abs/1801.02961
78. Cabrera-Bean M, Dos Santos VJP, Roso-Llorach A, et al. Autoencoders for health improvement by compressing the set of patient features. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 5917–5920.
79. Lee B, Baek J, Park S, et al. deepTarget: end-to-end learning framework for microRNA target prediction using deep recurrent neural networks. In: *Proceedings of the 7th ACM international conference on bioinformatics, computational biology, and health informatics*, 2016, pp. 434–442.
80. Pardede HF, Suryawati E, Sustika R, et al. Unsupervised convolutional autoencoder-based feature learning for automatic detection of plant diseases. In: *2018 international conference on computer, control, informatics and its applications (IC3INA)*. IEEE, 2018, pp. 158–162.
81. Naga Srinivasu P, Krishna TB, Ahmed S, et al. Variational autoencoders-BasedSelf-learning model for tumor identification and impact analysis from 2-D MRI images. *Journal of Healthcare Engineering* 2023; 2023: 1.
82. Raut G, Raut A, Bhagade J, et al. Deep learning approach for brain tumor detection and segmentation. In: *2020 International Conference on Convergence to Digital World-Quo Vadis (ICCDW)*. IEEE, 2020, pp. 1–5.
83. Lingwal S, Rauthan JS and Negi B. Deep learning for health care in disease identification: a review. *Proceedings of integrated intelligence enable networks and computing: IIENC 2020*, 2021, pp. 641–650.
84. Gupta S and Gupta MK. *A comparative analysis of deep learning approaches for predicting breast cancer survivability*. Archives of Computational Methods in Engineering, 2021, pp. 1–17.
85. Strauman AS, Bianchi FM, Mikalsen KØ, et al. Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks. In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2018, pp. 307–310.
86. Hsu M-H, Juang W-C, Cai Z-X, et al. Evaluation of recurrent neural network model training for health care suggestions. In: *2021 International Conference on Security and Information Technologies with AI, Internet Computing and Big-data Applications*. Springer, 2022, pp. 161–168.
87. Sanderson M, Bulloch AGM, Wang J, et al. Predicting death by suicide using administrative health care system data: can recurrent neural network, one-dimensional convolutional neural network, and gradient boosted trees models improve prediction performance? *J Affect Disord* 2020; 264: 107–114.
88. Yuan J, Wang H, Lin C, et al. A novel GRU-RNN network model for dynamic path planning of mobile robot. *IEEE Access* 2019; 7: 15140–15151.
89. Eddahech A, Briat O, Bertrand N, et al. Behavior and state-of-health monitoring of Li-ion batteries using impedance spectroscopy and recurrent neural networks. *Int J Electr Power Energy Syst* 2012; 42(1): 487–494.

90. Chen R, Stewart WF, Sun J, et al. Recurrent neural networks for early detection of heart failure from longitudinal electronic health record data: implications for temporal modeling with respect to time before diagnosis, data density, data quantity, and data type. *Circulation: Cardiovascular Quality and Outcomes* 2019; 12(10): e005114.
91. Chopra C, Sinha S, Jaroli S, et al. Recurrent neural networks with non-sequential data to predict hospital readmission of diabetic patients. In: *Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics*, 2017, pp. 18–23.
92. Chingombe I, Dzinamarira T, Cuadros D, et al. Predicting HIV status among men who have sex with men in Bulawayo & Harare, Zimbabwe using bio-behavioural data, recurrent neural networks, and machine learning techniques. *Tropical Medicine and Infectious Disease* 2022; 7(9): 231.
93. Jose A, Roy R, Eschweiler D, et al. End-to-end classification of cell-cycle stages with center-cell focus tracker using recurrent neural networks. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
94. Nanduri A and Sherry L. Anomaly detection in aircraft data using Recurrent Neural Networks (RNN). In: *2016 Integrated Communications Navigation and Surveillance (ICNS)*. IEEE, 2016, pp. 1–2.
95. Chandani K. An experimental comparison of deep LSTM and GRUs for event classification in sports. *Master-Thesis*. 2019. https://github.com/KajalChandani/Master-Thesis-.
96. Zhao D, Feng L and Wang X. Bearing remaining useful life estimation based on encoder and gated recurrent units. *IAENG Int J Comput Sci* 2022; 49(1): 7.
97. Naganure N, Ashwin NU and Kamath SS. Leveraging deep learning approaches for patient case similarity evaluation. In: *Intelligent Data Engineering and Analytics: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020)*. Springer, 2021, vol 2, pp. 613–622.
98. Murayama T, Wakamiya S and Aramaki E. Single model for influenza forecasting of multiple countries by multi-task learning. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021, pp. 335–350.
99. Qiao Z, Zhao S, Xiao C, et al. Pairwise-ranking based collaborative recurrent neural networks for clinical event prediction. In: *Proceedings of the twenty-seventh international joint conference on artificial intelligence*, 2018.
100. Urban M, Klum M, Pielmus A-G, et al. GRU neural network improved bioimpedance based stroke volume estimation during ergometry stress test. *Sensors* 2022; 22(20): 7883.
101. Wingate J, Kollia I, Bidaut L, et al. Unified deep learning approach for prediction of Parkinson's disease. *IET Image Process* 2020; 14(10): 1980–1989.
102. Rodrigues-Jr JF, Gutierrez MA, Spadon G, et al. LIG-Doctor: efficient patient trajectory prediction using bidirectional minimal gated-recurrent networks. *Inf Sci* 2021; 545: 813–827.
103. Girkar UM, Uchimido R, Lehman L-H, et al. Predicting blood pressure response to fluid bolus therapy using attention-based neural networks for clinical interpretability. *arXiv Mach Learn*. 2018; pp. 1-6. https://arXiv:1812.00699v1.
104. Khan P, Ranjan P and Kumar S. AT2GRU: a human emotion recognition model with mitigated device heterogeneity. In: *IEEE Transactions on Affective Computing*, 2021.
105. Yang C, Wu Q, Wang J, et al. Graph neural networks are inherently good generalizers: insights by bridging gnns and mlps 2022.
106. Eliasof M, Haber E and Treister E. Pde-gcn: novel architectures for graph neural networks motivated by partial differential equations. *Adv Neural Inf Process Syst* 2021; 34: 3836–3849.
107. Duong LT, Chu CQ, Nguyen PT, et al. Edge detection and graph neural networks to classify mammograms: a case study with a dataset from Vietnamese patients. *Appl Soft Comput* 2023; 134: 109974.

108. Gao J, Sharma R, Qian C, et al. STAN: spatio-temporal attention network for pandemic prediction using real-world evidence. *J Am Med Inf Assoc* 2021; 28(4): 733–743.
109. Kapoor A, Ben X, Liu L, et al. Examining covid-19 forecasting using spatio-temporal graph neural networks. *Google Research*. 2020. https://arXiv:2007.03113.
110. Tomy A, Razzanelli M, Di Lauro F, et al. Estimating the state of epidemics spreading with graph neural networks. *Nonlinear Dynam* 2022; 109(1): 249–263.
111. Aguinaga S, Lasaga D, Danthuri S, et al. Identification of providers with similar risk profiles in healthcare claims using graphs. In: *2023 11th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, 2023, pp. 1–6.
112. Koca MB, Nourani E, Abbasoğlu F, et al. Graph convolutional network based virus-human protein-protein interaction prediction for novel viruses. *Comput Biol Chem* 2022; 101: 107755.
113. Huang S, Zheng S and Chen R. Multi-source transfer learning with Graph Neural Network for excellent modelling the bioactivities of ligands targeting orphan G protein-coupled receptors. *Math Biosci Eng* 2023; 20(2): 2588–2608.
114. Nguyen TM, Nguyen T, Le TM, et al. Gefa: early fusion approach in drug-target affinity prediction. *IEEE ACM Trans Comput Biol Bioinf* 2021; 19(2): 718–728.
115. Najam-ul-Lail IM, Naseer MA, Tabussam I, et al. Application of explainable artificial intelligence in drug discovery and drug design. In: *Explainable Artificial Intelligence for Biomedical Applications*, 2023, p. 213.
116. Qian Y, Zheng J, Jiang Y, et al. Prediction of circRNA-miRNA association using singular value decomposition and graph neural networks. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.
117. Hinton GE. Deep belief networks. *Scholarpedia* 2009; 4(5): 5947.
118. Fink O and Weidmann U. Using deep belief networks for predicting railway operations failures. In: *Advances in Risk and Reliability Technology Symposium. Citeseer*, 2013, p. 274.
119. Turner JT, Page A, Mohsenin T, et al. Deep belief networks used on high resolution multichannel electroencephalography data for seizure detection. *AAAI Spring Symp Series*. 2014; pp 75-81.
120. Madhavan M and Gopalakrishnan G. Long non-coding RNAs in heart failure: a deep belief network based cluster analysis. *Curr Bioinf* 2021; 16(7): 983–991.
121. Noor MBT, Zenia NZ, Kaiser MS, et al. Detecting neurodegenerative disease from MRI: a brief review on a deep learning perspective. In: *Brain Informatics: 12th International Conference, BI 2019, Haikou, China, December 13–15, 2019, Proceedings 12*. Springer, 2019, pp. 115–125.
122. Kim Y, Lee H and Provost EM. Deep learning for robust feature generation in audiovisual emotion recognition. In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 3687–3691.
123. Maqsood A, Farid MS, Khan MH, et al. Deep malaria parasite detection in thin blood smear microscopic images. *Appl Sci* 2021; 11(5): 2284.
124. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med* 2019; 25(1): 24–29.
125. Kaul D, Raju H and Tripathy BK. Deep learning in healthcare. In: *Deep Learning in Data Analytics: Recent Techniques, Practices and Applications*, 2022, pp. 97–115.
126. Xie S, Yu Z and Lv Z. Multi-disease prediction based on deep learning: a survey. *CMES-Computer Modeling in Engineering & Sciences* 2021; 128(2).
127. Yu Z, Wang K, Wan Z, et al. Popular deep learning algorithms for disease prediction: a review. *Cluster Comput* 2023; 26(2): 1231–1251.
128. Berdigaliyev N and Aljofan M. An overview of drug discovery and development. *Future Med Chem* 2020; 12(10): 939–947.

129. Askr H, Elgeldawi E, Aboul Ella H, et al. Deep learning in drug discovery: an integrative review and future challenges. *Artif Intell Rev* 2023; 56(7): 5975–6037.
130. Bai Q, Tan S, Xu T, et al. MolAICal: a soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. *Briefings Bioinf* 2021; 22(3): bbaa161.
131. Popova M, Isayev O and Tropsha A. Deep reinforcement learning for de novo drug design. *Sci Adv* 2018; 4(7): eaap7885.
132. Suzuki K. Overview of deep learning in medical imaging. *Radiol Phys Technol* 2017; 10(3): 257–273.
133. Aggarwal R, Sounderajah V, Martin G, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine* 2021; 4(1): 65.
134. Zhou SK, Greenspan H, Davatzikos C, et al. A review of deep learning in medical imaging: imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc IEEE* 2021; 109(5): 820–838.
135. Cenggoro TW and Pardamean B. others. A systematic literature review of machine learning application in COVID-19 medical image classification. *Proc Comput Sci* 2023; 216: 749–756.
136. Sujith A, Sajja GS, Mahalakshmi V, et al. Systematic review of smart health monitoring using deep learning and Artificial intelligence. *Neuroscience Informatics* 2022; 2(3): 100028.
137. Allam Z and Jones DS. On the coronavirus (COVID-19) outbreak and the smart city network: universal data sharing standards coupled with artificial intelligence (AI) to benefit urban health monitoring and management. In: *Healthcare*. MDPI, 2020, p. 46.
138. Ali F, El-Sappagh S, Islam SMR, et al. An intelligent healthcare monitoring framework using wearable sensors and social networking data. *Future Generat Comput Syst* 2021; 114: 23–43.
139. Che Azemin MZ, Hassan R, Mohd Tamrin MI, et al. COVID-19 deep learning prediction model using publicly available radiologist-adjudicated chest X-ray images as training data: preliminary findings. *Int J Biomed Imag* 2020; 2: 1.
140. Shorten C, Khoshgoftaar TM and Furht B. Deep learning applications for COVID-19. *Journal of big Data* 2021; 8(1): 1–54.
141. Wang Y, Nazir S and Shafiq M. An overview on analyzing deep learning and transfer learning approaches for health monitoring. *Comput Math Methods Med* 2021; 2021: 1–10.
142. Purushotham S, Meng C, Che Z, et al. Benchmarking deep learning models on large healthcare datasets. *J Biomed Inf* 2018; 83: 112–134.
143. Chui KT, Alhalabi W, Pang SSH, et al. Disease diagnosis in smart healthcare: innovation, technologies and applications. *Sustainability* 2017; 9(12): 2309.
144. Walport M and Brest P. Sharing research data to improve public health. *Lancet* 2011; 377(9765): 537–539.
145. McDermott MBA, Wang S, Marinsek N, et al. Reproducibility in machine learning for health research: still a ways to go. *Sci Transl Med* 2021; 13(586): eabb1655.
146. Goodman SN, Fanelli D and Ioannidis JPA. What does research reproducibility mean? *Sci Transl Med* 2016; 8(341): 341ps12.
147. McDermott M, Wang S, Marinsek N, et al. Reproducibility in machine learning for health. *Representative Reproducibility Mach. Learn. Workshop*. 2019.
148. Jahan S, Abu Taher K, Kaiser MS, et al. Explainable AI-based Alzheimer's prediction and management using multimodal data. *PLoS One* 2023; 18(11): e0294253.
149. Giuste F, Shi W, Zhu Y, et al. Explainable artificial intelligence methods in combating pandemics: a systematic review. In: *IEEE Reviews in Biomedical Engineering*, 2022.
150. Nair PC, Gupta D, Devi BI, et al. Building an explainable diagnostic classification model for brain tumor using discharge summaries. *Proc Comput Sci* 2023; 218: 2058–2070.

151. Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Comput Surv* 2018; 51(5): 1–42.
152. Hamon R, Junklewitz H and Sanchez I. *Robustness and explainability of artificial intelligence*. Publications Office of the European Union, 2020, p. 207.
153. Zednik C. Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & technology* 2021; 34(2): 265–288.
154. Karim MR, Islam T, Shajalal M, et al. Explainable ai for bioinformatics: methods, tools and applications. *Briefings Bioinf* 2023; 24(5): bbad236.
155. Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning. *ACM Comput Surv* 2021; 54(6): 1–35.
156. Dickinson Q and Meyer JG. Positional SHAP (PoSHAP) for Interpretation of machine learning models trained from biological sequences. *PLoS Comput Biol* 2022; 18(1): e1009736.
157. Rao RR and Makkithaya K. Learning from a class imbalanced public health dataset: a cost-based comparison of classifier performance. *Int J Electr Comput Eng* 2017; 7(4): 2215.
158. Hasib KM, Iqbal MS, Shah FM, et al. A survey of methods for managing the classification and solution of data imbalance problem. *arXiv:2012.11870*. 2020.
159. Sharma S, Gosain A and Jain S. A review of the oversampling techniques in class imbalance problem. In: *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021*. Springer, 2022, vol 1, pp. 459–472.
160. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res* 2002; 16: 321–357.
161. Turlapati VPK and Prusty MR. Outlier-SMOTE: a refined oversampling technique for improved detection of COVID-19. *Intelligence-based medicine* 2020; 3: 100023.
162. Gao L, Zhang L, Liu C, et al. Handling imbalanced medical image data: a deep-learning-based one-class classification approach. *Artif Intell Med* 2020; 108: 101935.
163. Sarvamangala DR and Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. *Evolutionary intelligence* 2022; 15(1): 1–22.
164. Weerakody PB, Wong KW, Wang G, et al. A review of irregular time series data handling with gated recurrent neural networks. *Neurocomputing* 2021; 441: 161–178.
165. Alzubaidi L, Al-Amidie M, Al-Asadi A, et al. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers* 2021; 13(7): 1590.
166. Kostkova P, Brewer H, De LS, et al. Who owns the data? Open data for healthcare. *Front Public Health* 2016; 4: 7.
167. Agrawal R and Prabakaran S. Big data in digital healthcare: lessons learnt and recommendations for general practice. *Heredity* 2020; 124(4): 525–534.
168. Hicks J. The future of data ownership: an uncommon research agenda. *Socio Rev* 2023; 71(3): 544–560.
169. Jaiman V and Urovi V. A consent model for blockchain-based health data sharing platforms. *IEEE Access* 2020; 8: 143734–143745.
170. Gordon WJ and Catalini C. Blockchain technology for healthcare: facilitating the transition to patient-driven interoperability. *Comput Struct Biotechnol J* 2018; 16: 224–230.
171. Saeed H, Malik H, Bashir U, et al. Blockchain technology in healthcare: a systematic review. *PLoS One* 2022; 17(4): e0266462.
172. Micheal A. Breast cancer risk prediction combining a convolutional neural network-based mammographic evaluation with clinical factors. *Breast Cancer Res Treat* 2023; 200(2): 237–245. DOI: 10.1007/s10549-023-06966-4.
173. Julia M. Use of a convolutional neural network-based mammographic evaluation to predict breast cancer recurrence among women with hormone receptor-positive operable breast cancer. *Breast Cancer Res Treat* 2022; 194(1): 35–47. DOI: 10.1007/s10549-022-06614-3.

174. Kevin A. Class imbalance in out-of-distribution datasets: improving the robustness of the TextCNN for the classification of rare cancer types. *J Biomed Inf* 2022; 125: 2. DOI: 10.1016/j.jbi.2021.103957.
175. Hasley M. Dynamic changes of convolutional neural network-based mammographic breast cancer risk score among women undergoing chemoprevention treatment. *Clin Breast Cancer* 2021; 21(4): e312–e318. DOI: 10.1016/j.clbc.2020.11.007.
176. Hasley W. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. *Artif Intell Med* 2020; 2: 110. DOI: 10.1016/j.artmed.2020.101977.
177. Hasley SH. A quantitative quality control method of big data in cancer patients using artificial neural network. In: *CCIS 2014 - proceedings of 2014 IEEE 3rd international conference on cloud computing and intelligence systems*, 2014. DOI: 10.1109/CCIS.2014.7175787.
178. Yang YY and Fasching PA. Tresp. predictive modeling of therapy decisions in metastatic breast cancer with recurrent neural network encoder and multinomial hierarchical regression decoder. In: *Proceedings - 2017 IEEE international conference on healthcare informatics, ICHI 2017*, 2017. DOI: 10.1109/ICHI.2017.51.
179. D Qjx G, Yoon HJ and Fearn PA. Deep learning for automated extraction of primary sites from cancer pathology reports. In: *IEEE Journal of Biomedical and Health Informatics* 2018. DOI: 10.1109/JBHI.2017.2700722.
180. Tap P, Bashiri FS and Larose E. Diagnostic classification of lung CT images using deep 3D multi-scale convolutional neural network. In: *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, 2018. DOI: 10.1109/ICHI.2018.00078.
181. Baik SM, Hong KS and Park DJ. Deep learning approach for early prediction of COVID-19 mortality using chest X-ray and electronic health records. *BMC Bioinf* 2023; 24(1): 1–13.
182. Varde AS, Karthikeyan D and Wang W. Facilitating COVID recognition from X-rays with computer vision models and transfer learning. *Multimed Tool Appl* 2023; 2: 1–32.
183. Bashir SR, Raza S, Kocaman V, et al. Clinical application of detecting COVID-19 risks: a natural language processing approach. *Viruses* 2022; 14(12): 2761.
184. Butler L, Karabayir I, Tootooni MS, et al. Image and structured data analysis for prognostication of health outcomes in patients presenting to the ED during the COVID-19 pandemic. *Int J Med Inf* 2022; 158: 104662.
185. Seung R. Deep learning approach for early prediction of COVID-19 mortality using chest X-ray and electronic health records. In: *BMC Bioinformatics*, 2023.
186. Lalila R. Recurrent neural network models (CovRNN) for predicting outcomes of patients with COVID-19 on admission to hospital: model development and validation using electronic health record data. In: *The Lancet Digital Health*, 2022.
187. Sophie P. COVID-19 patient outcome prediction using selected features from emergency department data and feed-forward neural networks. In: *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*. LNICST, 2021 2021.
188. Sharmila S. Deep learning medical solutions: a short review on COVID-19 detection using chest X-ray image with CNN multiple architectures. In: *Proceedings - 2022 5th International Conference on Computational Intelligence and Communication Technologies, CCICT 2022*, 2022.
189. Ronald W. Using similar patients to predict complication in patients with diabetes, hypertension, and lipid disorder: a domain knowledge-infused convolutional neural network approach. *J Am Med Inf Assoc: JAMIA* 2023; 30(2): 273–281. DOI: 10.1093/jamia/ocac212.
190. Shrestha M, Alsadoon OH, Alsadoon A, et al. A novel solution of deep learning for enhanced support vector machine for predicting the onset of type 2 diabetes. *Multimed Tool Appl* 2023; 82(4): 6221–6241.

191. Jin Y, Li F, Vimalananda VG, et al. Automatic detection of hypoglycemic events from the electronic health record notes of diabetes patients: empirical study. *JMIR medical informatics* 2019; 7(4): e14340.
192. Pereira V, Matos S and Oliveira JL. Automated ICD-9-CM medical coding of diabetic patient's clinical reports. In: *Proceedings of the First International Conference on Data Science, E-learning and Information Systems*, 2018, pp. 1–6.
193. Woodman Richard J and Mangoni Arduino A. Patient multi-relational graph structure learning for diabetes clinical assistant diagnosis. *Aging Clinical and Experimental Research*. 2023. https://doi.org/10.1007/s40520-023-02552-2
194. Orw NC, Hsu W and Lee ML. Using similar patients to predict complication in patients with diabetes, hypertension, and lipid disorder: a domain knowledge-infused convolutional neural network approach. *J Am Med Inf Assoc: JAMIA* 2023; 2: 3. DOI: 10.1093/jamia/ocac212.
195. A SM, Alsadoon OH, Alsadoon A, et al. A novel solution of deep learning for enhanced support vector machine for predicting the onset of type 2 diabetes. *Multimed Tool Appl* 2023; 1: 1. DOI: 10.1007/s11042-022-13582-9.
196. A GI, Saccani S, Coser A, et al. Deep-learning-based prognostic modeling for incident heart failure in patients with diabetes using electronic health records: a retrospective cohort study. *PLoS One* 2023; 1: 3. DOI: 10.1371/journal.pone.0281878.
197. Manikandan R, Kuwelkar S and Sivakumar R. An hybrid technique for optimized clustering of EHR using binary particle swarm and constrained optimization for better performance in prediction of cardiovascular diseases. *Measurement: Sensors* 2023; 25: 100577.
198. An Y, Tang K and Wang J. Time-aware multi-type data fusion representation learning framework for risk prediction of cardiovascular diseases. *IEEE ACM Trans Comput Biol Bioinf* 2021; 19(6): 3725–3734.
199. Rustam F, Ishaq A, Munir K, et al. Incorporating CNN features for optimizing performance of ensemble classifier for cardiovascular disease prediction. *Diagnostics* 2022; 12(6): 1474.
200. Zhao J, Feng Q, Wu P, et al. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Sci Rep* 2019; 9(1): 717.
201. Su J, Hu J, Jiang J, et al. Extraction of risk factors for cardiovascular diseases from Chinese electronic medical records. *Comput Methods Progr Biomed* 2019; 172: 1–10.
202. P LJ and Hashemi AS. Explainable graph neural networks for atherosclerotic cardiovascular disease. In: *Studies in health technology and informatics*, 2023. DOI: 10.3233/SHTI230214.
203. B GA, Longato E, Fadini GP, et al. Explainable deep-learning model reveals past cardiovascular disease in patients with diabetes using free-form visit reports. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2023. DOI: 10.1007/978-3-031-25599-1_32.
204. Ge Y, Wang Q, Wang L, et al. Predicting post-stroke pneumonia using deep neural network approaches. *Int J Med Inf* 2019; 132: 103986.
205. Han L, Askari M, Altman RB, et al. Atrial fibrillation burden signature and near-term prediction of stroke: a machine learning analysis. *Circulation: Cardiovascular Quality and Outcomes* 2019; 12(10): e005595.
206. Hung C-Y, Lin C-H, Lan T-H, et al. Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database. *PLoS One* 2019; 14(3): e0213007.
207. Hung C-Y, Lin C-H and Lee C-C. Improving young stroke prediction by learning with active data augmenter in a large-scale electronic medical claims database. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 5362–5365.

208. Dev S, Wang H, Nwosu CS, et al. A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics* 2022; 2: 100032.
209. Oei RW, Hsu W, Lee ML, et al. Using similar patients to predict complication in patients with diabetes, hypertension, and lipid disorder: a domain knowledge-infused convolutional neural network approach. *J Am Med Inf Assoc* 2023; 30(2): 273–281.
210. Datta S, Morassi Sasso A, Kiwit N, et al. Predicting hypertension onset from longitudinal electronic health records with deep learning. *JAMIA open* 2022; 5(4): ooac097.
211. Ren Y, Fei H, Liang X, et al. A hybrid neural network model for predicting kidney disease in hypertension patients based on electronic health records. *BMC Med Inf Decis Making* 2019; 19: 131–138.
212. Velazquez M and Lee Y. Multimodal ensemble model for Alzheimer's disease conversion prediction from Early Mild Cognitive Impairment subjects. *Comput Biol Med* 2022; 151: 106201.
213. Akter S, Das D, Haque RU, et al. AD-CovNet: an exploratory analysis using a hybrid deep learning model to handle data imbalance, predict fatality, and risk factors in Alzheimer's patients with COVID-19. *Comput Biol Med* 2022; 146: 105657.
214. Shen Z, Schutte D, Yi Y, et al. Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision. *BMC Med Inf Decis Making* 2022; 22(1): 1–11.
215. Prabhu SS, Berkebile JA, Rajagopalan N, et al. Multi-modal deep learning models for alzheimer's disease prediction using MRI and EHR. In: *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2022, pp. 168–173.
216. Bhadra S and Kumar CJ. Enhancing the efficacy of depression detection system using optimal feature selection from EHR. In: *Computer methods in biomechanics and biomedical engineering*, 2023, pp. 1–15.
217. Gong J, Simon GE and Liu S. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. *PLoS One* 2019; 14(9): e0222665.
218. Geraci J, Wilansky P, de Luca V, et al. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evid Base Ment Health* 2017; 20(3): 83.
219. Wang L, Sha L, Lakin JR, et al. Development and validation of a deep learning algorithm for mortality prediction in selecting patients with dementia for earlier palliative care interventions. *JAMA Netw Open* 2019; 2(7): e196972.
220. Ford E, Rooney P, Oliver S, et al. Identifying undetected dementia in UK primary care patients: a retrospective case-control study comparing machine-learning and standard epidemiological approaches. *BMC Med Inf Decis Making* 2019; 19(1): 1–9.
221. Liang W, Zhang W, Jin B, et al. Learning to detect pathogenic microorganism of community-acquired pneumonia. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 969–972.
222. Siordia-Millán S, Torres-Ramos S, Salido-Ruiz RA, et al. Pneumonia and pulmonary thromboembolism classification using electronic health records. *Diagnostics* 2022; 12(10): 2536.
223. Hsu J-C, Wu F-H, Lin H-H, et al. AI models for predicting readmission of pneumonia patients within 30 days after discharge. *Electronics* 2022; 11(5): 673.
224. AlSaad R, Malluhi Q, Janahi I, et al. Interpreting patient-Specific risk prediction using contextual decomposition of BiLSTMs: application to children with asthma. *BMC Med Inf Decis Making* 2019; 19: 1–11.
225. AlSaad R, Malluhi Q, Janahi I, et al. Predicting emergency department utilization among children with asthma using deep learning models. *Healthcare Analytics* 2022; 2: 100050.
226. Gupta M, Phan T-LT, Bunnell HT, et al. Obesity prediction with EHR data: a deep learning approach with interpretable elements. *ACM Trans Comput Healthcare* 2022;3(3):1-19.
227. Xue Q, Wang X, Meehan S, et al. Recurrent neural networks based obesity status prediction using activity data. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Orlando, FL, USA, 2018.

228. Wu Z, Xu D, Hu PJ-H, et al. A hierarchical multilabel graph attention network method to predict the deterioration paths of chronic hepatitis B patients. *J Am Med Inf Assoc* 2023; 30(5): 846–858.
229. Koh E and Kim Y. Risk association of liver cancer and hepatitis B with tree ensemble and lifestyle features. *Int J Environ Res Publ Health* 2022; 19(22): 15171.
230. Reiser M, Wiebner B, Hirsch J, et al. Neural-network analysis of socio-medical data to identify predictors of undiagnosed hepatitis C virus infections in Germany (DETECT). *J Transl Med* 2019; 17: 1–7.
231. Lee LT-J, Yang H-C, Nguyen PA, et al. Machine learning approaches for predicting psoriatic arthritis risk using electronic medical records: population-based study. *J Med Internet Res* 2023; 25: e39972.
232. Norgeot B, Glicksberg BS, Trupin L, et al. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. *JAMA Netw Open* 2019; 2(3): e190606.

# Appendix A

## *Detailed search formulas and criteria*

This appendix provides the specific search formulas used in each database (i.e. ID 1 in Table 1) to ensure the retrieval of relevant literature on the application of deep learning in public health data.

## *PubMed*

To conduct a comprehensive search in PubMed, the following steps were taken. First, we navigated to the PubMed Web site. The search formula used was for ID 1 as follows:

("Public Health Data" [Title/Abstract] OR "Public Health Dataset" [Title/Abstract] OR "EHR" [Title/Abstract] OR "electronic health record" [Title/Abstract]) AND ("deep learning" [Title/Abstract] OR "representation learning" [Title/Abstract] OR "neural network" [Title/Abstract] OR "convolutional neural network" [Title/Abstract] OR "ConvNet" [Title/Abstract] OR "CNN" [Title/Abstract] OR "recurrent neural network" [Title/Abstract] OR "RNN" [Title/Abstract] OR "long short-term memory" [Title/Abstract] OR "LSTM" [Title/Abstract] OR "generative adversarial network" [Title/Abstract] OR "GAN" [Title/Abstract] OR "autoencoder" [Title/Abstract] OR "restricted boltzmann machine" [Title/Abstract] OR "deep belief network" [Title/Abstract] OR "DBN" [Title/Abstract])

## *Web of Science*

We accessed the Web of Science database. The initial search formula for ID 1 was:

TS=("Public Health Data" OR "Public Health Dataset" OR "EHR" OR "electronic health record") AND TS=("deep learning" OR "representation learning" OR "neural network" OR "convolutional neural network" OR "ConvNet" OR "CNN" OR "recurrent neural network" OR "RNN" OR "long short-term memory" OR "LSTM" OR "generative adversarial network" OR "GAN" OR "autoencoder" OR "restricted boltzmann machine" OR "deep belief network" OR "DBN")

## *Scopus*

In Scopus, the search was conducted using the following detailed ID 1 formula:

TITLE-ABS-KEY ((“Public Health Data” OR “Public Health Dataset” OR “EHR” OR “electronic health record”) AND (“deep learning” OR “representation learning” OR “neural network” OR “convolutional neural network” OR “ConvNet” OR “CNN” OR “recurrent neural network” OR “RNN” OR “long short-term memory” OR “LSTM” OR “generative adversarial network” OR “GAN” OR “autoencoder” OR “restricted boltzmann machine” OR “deep belief network” OR “DBN”))

### *Springer*

For Springer, the search was performed using the following ID 1 search formula:

(“Public Health Data” OR “Public Health Dataset” OR “EHR” OR “electronic health record”) AND (“deep learning” OR “representation learning” OR “neural network” OR “convolutional neural network” OR “ConvNet” OR “CNN” OR “recurrent neural network” OR “RNN” OR “long short-term memory” OR “LSTM” OR “generative adversarial network” OR “GAN” OR “autoencoder” OR “restricted boltzmann machine” OR “deep belief network” OR “DBN”)

## Appendix B

Appendix Table A1

**Table A1.** Top publishers This table lists the top publishers and their respective article counts.

| Publisher | Count |
| --- | --- |
| IEEE | 644 |
| Elsevier | 200 |
| Springer | 137 |
| ACM | 148 |

Appendix Table A2

**Table A2.** Deep Learning Models This table shows the count of various deep learning models used in the reviewed articles This table shows the count of various deep learning models used in the reviewed articles.

| Type | Count |
| --- | --- |
| LSTM | 224 |
| CNN | 174 |
| GAN | 150 |
| AE | 148 |
| RNN | 139 |
| GRU | 46 |
| GNN | 21 |
| DBN | 11 |

Appendix Table A3

**Table A3.** Top researchers names and number of publications per researcher This table lists the top researchers and the number of publications attributed to each.

| Researcher name | Number of occurrence |
|---|---|
| Sun, jimeng (9737233900) | 9737233900′ occurs 26 times |
| Wang, fei (56177292700) | 56177292700′ occurs 23 times |
| Xiao, cao (57199799682) | 57199799682′ occurs 22 times |
| Xu, Hua (55493876700) | 55493876700′ occurs 17 times |
| Liu, Hongfang (7409753328) | 7409753328′ occurs 16 times |
| Jiang, Xiaoqian (24479530900) | 24479530900′ occurs 15 times |
| Glicksberg, Benjamin S. (55845627200) | 55845627200′ occurs 15 times |
| Tang, Buzhou (35115621400) | 35115621400′ occurs 15 times |
| Ma, fenglong (57052032200) | 57052032200′ occurs 14 times |
| Qian, Buyue (36601594000) | 36601594000′ occurs 13 times |

Appendix Table A4

**Table A4.** Top Journals and the Number of Publications shown in each Journal This table lists the top journals and the number of publications in each journal.

| Journal name | Number of occurrence |
|---|---|
| 'Journal of Biomedical informatics' | Occurs 116 times |
| 'Lecture notes in computer science' | Occurs 73 times |
| 'BMC medical informatics and decision making' | Occurs 73 times |
| 'Journal of the American medical informatics association' | Occurs 68 times |
| 'JMIR medical informatics' | Occurs 61 times |
| 'IEEE journal of Biomedical and health informatics' | Occurs 56 times |
| 'Artificial intelligence in medicine' | Occurs 48 times |
| 'PLoS one' | Occurs 41 times |
| 'Studies in health technology and informatics' | Occurs 37 times |
| 'Scientific reports' | Occurs 37 times |

## Appendix Table A5

**Table A5.** Summary of common disease categories and techniques studied This table summarizes the common disease categories and the techniques used in the studies.

| Disease category | Common techniques | Samples of studies' references |
| --- | --- | --- |
| Cancer | Recurrent neural network, convolutional neural network | 63,105,172–180 |
| COVID-19/Corona virus | Convolutional neural network, Multilayer perceptron, transfer learning | 140,181–188 |
| Diabetes | Convolutional neural network, long short term memory networks (LSTM) | 189–196 |
| Cardiovascular | Convolutional neural network, long short term memory networks (LSTM), bidirectional long short-term memory (BiLSTM) | 197–203 |
| Stroke | Recurrent neural network (RNN), convolutional neural network (CNN), deep neural network (DNN), perceptron neural network | 204–208 |
| Hypertension | Convolutional neural network (CNN), long short-term memory (LSTM), bidirectional long short-term memory (BiLSTM), Autoencoder networks | 209–211 |
| Alzheimer | Convolutional neural network (CNN), long short-term memory, BERT, deep Auto-encoder | 212–215 |
| Depression | Artificial neural network, deep neural network | 216–218 |
| Dementia | Neural networks, recurrent neural networks | 219,220 |
| Pneumonia | Recurrent neural network, BiLSTM, deep neural network | 204,221–223 |
| Asthma | Recurrent neural network, long short-term memory, BiLSTM | 224,225 |
| Obesity | Recurrent neural network, long short-term memory, BiLSTM | 226,227 |
| Hepatitis | Artificial neural network | 228–230 |
| Arthritis | Convolutional neural network, longitudinal deep learning model | 231,232 |