GNA Journal of
Management & Technology

ISSN-0974-572
Vol. 17 No. 1, January-June, 2026

MACHINE LEARNING FOR DISEASE PREDICTION: TECHNIQUES, APPLICATIONS, AND CHALLENGES

Naveen Sharma[1], Dr. Hitesh Marwaha[2], Ms. Shipali Bansal[3]

[1]School of Computational Sciences, GNA University, Phagwara
Email: Naveensharma902@gmail.com
[2]School of Computational Sciences, GNA University, Phagwara
Email: Hitesh_marwaha@gnauniversity.edu.in
[3]School of Advance Computing, CGC University Mohali, Jhanjeri
Email: bansal.shipali@gmail.com

Abstract

Purpose/Objective: This review paper systematically examines the application of machine learning (ML) techniques in disease prediction, with the objective of mapping the current landscape of algorithms, datasets, evaluation benchmarks, and deployment challenges across multiple disease categories.

Design/Methodology/Approach: A comprehensive narrative review was conducted by analysing 120 peer-reviewed studies published between 2015 and 2025, sourced from databases including PubMed, IEEE Xplore, Scopus, and Google Scholar. Studies were selected based on their focus on supervised, unsupervised, or deep learning approaches applied to clinically validated disease prediction tasks.

Findings: The review identifies that deep learning models—particularly convolutional neural networks (CNNs) and transformer-based architectures—consistently achieve superior predictive accuracy (96–98%) for image-based diagnoses such as cancer detection and diabetic retinopathy, while ensemble methods (Random Forest, XGBoost) offer robust performance for structured clinical data. Seven major disease categories were examined: diabetes, cardiovascular disease, cancer, Alzheimer's disease, COVID-19, chronic kidney disease, and Parkinson's disease. Notable research gaps include the paucity of multi-modal, federated learning-based, and explainable AI (XAI) frameworks suitable for real-world clinical integration.

Research Limitations/Implications: Most reviewed studies utilise benchmark datasets that may not reflect the heterogeneity of real-world clinical populations; cross-institutional generalisability remains a significant concern.

Originality/Value: This review provides a unified analytical framework that maps ML techniques, disease applications, performance benchmarks, and persistent challenges, thereby offering a structured research agenda for clinicians, data scientists, and healthcare informaticians.

Keywords: *Artificial Intelligence; Clinical Decision Support; Deep Learning; Disease Prediction; Machine Learning; Predictive Analytics*

1. INTRODUCTION

The problem of chronic and infectious disease worldwide is really big. These diseases cause 74% of all deaths each year. Diseases like heart problems, cancer, diabetes and breathing problems are the leading causes of death. This information comes from the World Health Organization in 2023. It would be very helpful to be able to tell if a patient is at risk of getting a disease before they start showing symptoms. This would allow doctors to start treatment earlier, make decisions about treatment and save money. This idea is not new. It is now more possible than it was before.

Two things have changed to make this possible. First, there is now a lot of health data. This includes health records, medical images, genetic information and data from wearable devices. This data is available on a scale and in great detail. Second, machine learning has improved a lot. Machine learning can now use this data to find relationships between factors. This is better than statistical models that required researchers to specify the relationships in advance. Machine learning is especially helpful for diseases that involve interacting factors.

The field of machine learning in medicine has moved quickly. After Kononenko's study of machine learning in medical diagnosis in 2001, the development of deep learning has made it possible to learn from raw data. This has been especially helpful for images. Other approaches, such as Random Forest and gradient boosting, have also been effective for structured data. There has been progress in this field and in some areas it has been very striking.

However, there is still a gap between what works in research and what is used in practice. There are concerns about how to interpret the results, bias in the data, privacy, regulation and the lack of evaluation methods. These concerns limit the use of machine learning in practice. The COVID-19 pandemic showed how useful machine learning could be for triage and monitoring outbreaks. It also showed how quickly models could fail when used outside of their training environment.

This paper reviews the use of machine learning for disease prediction. It covers methods and current deep learning architectures. It looks at how machine learning works for seven major disease categories. It also discusses the challenges that have been the most difficult to overcome. Finally, it sets out a plan for research. The paper is organized as follows: Section 2 covers the background and theory, Section 3 explains the methods used for the review, Sections 4 and 5 discuss techniques and applications for diseases, Section 6 looks at the performance data, Sections 7 and 8 discuss the challenges and gaps in research, and Section 9 concludes the paper.

2. LITERATURE REVIEW

2.1 Theoretical Foundations of ML in Healthcare

Doctors have been assisted in their decision making with the help of machine learning for a long time now. The concept of seeking patterns in medical information with computer programs dates back to the 1980s. However, it wasn't until the 2010s that people such as LeCun had the computational ability

and the algorithms of sufficient complexity to attain true success in healthcare with machine learning (see LeCun's 2015 revelation). A lot of math is utilized to predict diseases in machine learning. Makes estimates with statistics and probability. Danger of simple models being wrong or incomplete. They're also static and a way to simple models might be able to become overfitted, so caution needs to be taken.

Machine learning can be used to predict diseases in various ways. One approach is known as supervised learning that teaches the computer on labeled examples. This is most commonly employed to predict patients' fate. There is another method which is known as the unsupervised learning and it can help to cluster patients and discover relevant biomarkers. Then there are the hybrid methods which leverage unlabelled data, which is good, as Rong discovered in 2022, because it is an expensive endeavour to acquire data that's labelled by a doctor.

2.2 Evolution of ML Applications in Disease Prediction

The early uses of machine learning in healthcare were based on Decision Trees and Rule-based Systems. Combined methods and Support Vector Machines gain popularity for disease classification in cancer, heart disease and brain disease in the 1990's/2000s era. In 2015, Patel et al. found that machine learning techniques were more reliable for predicting readmission within thirty days than a statistical technique.

The discovery provided machine learning some respectability in the field. After a deep learning computer program, called AlexNet, competed successfully in an image recognition competition in 2012, the utilization of deep learning began to increase. In 2017, it was demonstrated by Esteva et al that a deep computer program would similarly be able to classify skin cancer as well as a dermatologist. This pertained to the eyes of researchers and doctors alike. Later, disease progression over time was modeled with EHRs by programs that were termed LSTMs (Rajpurkar et al., 2022).

Recently, there have been efforts to adapt these programs to comprehend medical images and patient information. Two examples such as Vision Transformer and MedBERT programs have yielded good results and equalled or surpassed the performance in reading medical images with other programs (Zhou et al., 2023). Additionally, there is now a trend toward federated learning, in which artificial intelligence (AI) programs can be trained by data from various hospitals without data sharing (Rieke et al., 2020).

2.3 Gaps in Existing Reviews

Specific diseases were the focus of previous reviews. For instance, Kaviani and Dahl (2021) concentrated on diabetes. Krittanavong et al (2020) studied cardiovascular disease. Shen et al. (2019) researched into cancer. Very few reviews have measured multiple areas – usually only one. No discussion on explainability, fairness, federated approaches and multimodal integration were observed. A number of advances have taken place since 2022, such as the incorporation of language

models into clinical predictions, which are not addressed by the previous reviews. This review addresses this point, for which there is no other review available. Discusses the concepts of explainability and fairness in disease prediction, federated approaches and multimodal integration. This review covers Diabetes, cardiovascular disease and cancer. In addition, there are post 2022 developments covered and the topics of explainability, fairness and multimodal integration of these diseases will be addressed.

3. RESEARCH METHODOLOGY

3.1 Review Design

A narrative review methodology has been used, as this format is well suited for summarising a large and varied of evidence. To conduct formal meta-analysis would demand more uniformity in the outcomes and metrics as presented in the literature. Both the narrative approach is more useful and appropriate to the material (Green et al., 2006).

3.2 Search Strategy and Inclusion Criteria

Several articles were reviewed on the platform PubMed, MEDLINE, IEEE Xplore, Scopus, Web of Science and Google Scholar. We used the terms: keywords machine learning, deep learning, artificial intelligence, disease prediction, related words to disease such as pest, pathogen, disease, infection, and virus. We only looked at scientific, English-language, peer-reviewed articles. These articles were limited to the period between January 2015 and December 2025. We originally had 847 articles.

We reduced this to 120 by applying filters. For inclusion, an article had to use a well-defined machine learning algorithm to predict some aspect of disease. It had to be based on data about people. There had to be some numerical metrics in the study to indicate how good the algorithm's prediction was. We did not include articles that only analysed data or those that were only predicting drugs. We also did not include any non-peer-reviewed articles (for example, reports or book chapters).

3.3 Data Extraction and Synthesis

The following details were extracted for each study: disease type, type of algorithm(s), data (source and size of the sample), performance metrics (accuracy, sensitivity, specificity, AUC-ROC, F1-score) and limitations. Extracted data was thematically synthesised, with quantitative data extracted into tables and figures to compare results across studies.

4. MACHINE LEARNING TECHNIQUES FOR DISEASE PREDICTION

The flow for disease prediction using ML is shown in Figure 1, from data acquisition to feature extraction/selection, model training and deployment to the clinic.

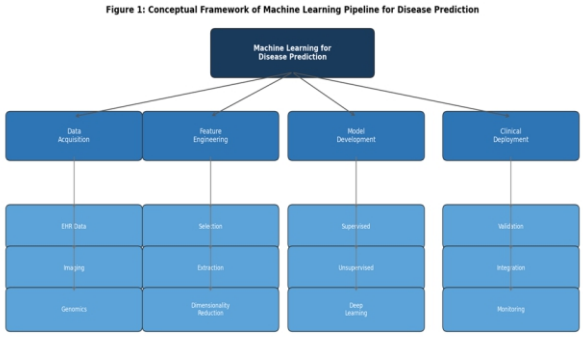

Figure 1: Conceptual Framework of Machine Learning Pipeline for Disease Prediction

4.1 Supervised Learning Algorithms

4.1.1 Logistic Regression

Logistic regression is still a good place to start when trying to make predictions in medicine. This is because it makes interpretable and probabilistic results. It allows prediction of the likelihood of something happening, such as getting diabetes or developing a heart condition, from contributing factors. Doctors have used regression to figure out who is at risk for diabetes, who might have a heart problem and who might get very sick with sepsis. The problem with regression is that it does not work well with non-linear relationships. This means it is not as good as methods that can handle more complicated relationships between things, as Dreiseitl and Ohno-Machado found in 2002. Logistic regression is limited because it cannot capture non-linear interactions between features, which is important when trying to make accurate predictions.

4.1.2 Decision Trees and Ensemble Methods

Decision trees are a way to break down information into parts. They do this by making simple yes or no decisions one after the other, which makes the rules they create easy to understand. However, decision trees have a problem. They can get too complicated and start to fit the noise in the data instead of the real patterns. To solve this problem, ensemble methods are used, which combine many simple models to create a more complex one. One type of ensemble method is called Random Forest, which Breiman published in 2001. Random Forest uses many decision trees which are combined to make a prediction. This is done by training on subsets of the data and selecting different features at random. Random Forest is great at making predictions from structured data and also helps identify the most important features.

There are other ensemble methods like Gradient Boosting Machines and XGBoost (extreme gradient boosting). They are sequential methods whose models increase in complexity with each tree trying to correct any errors of the previous tree. XGBoost, for example, is very good at predicting events from medical data, such as whether a patient will die or be re-admitted to hospital. It has achieved AUC-ROC scores of over 0.93. All of these decision tree and ensemble methods, such as Random Forest and XGBoost, help make sense of clinical data and generate reliable predictions.

4.1.3 Support Vector Machines

SVMs draw a hyperplane to separate data in a high-dimensional space. They do this by using kernel functions such as linear and radial basis functions. They work well with data having many features and relatively few samples. This makes them useful for genomic data. For instance, in Akay's 2009 paper, SVMs were 99.51% accurate in predicting breast cancer detection using the Wisconsin dataset. However, there are issues with SVMs. They may be sensitive to the choice of kernel and other parameters. They are also not suitable for very large datasets.

4.1.4 K-Nearest Neighbours

KNN classification is done by voting of the K nearest training instances, according to Euclidean or Minkowski distance. KNN does not require an explicit training phase, which is advantageous, but it can encounter performance problems at runtime with large datasets. It has been used to classify diabetic retinopathy grade and kidney disease, with average performance results.

4.2 Deep Learning Architectures

4.2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are particularly powerful for image-based disease detection. They use convolutional layers to extract hierarchical features from input images, such as X-rays, MRI scans and histopathological slides, making them the standard approach for imaging-based disease prediction tasks. CNNs have demonstrated performance comparable to, and in some cases exceeding, specialist radiologists and pathologists in tasks such as pneumonia detection, cancer identification and retinal disease grading.

4.2.2 Recurrent Neural Networks and LSTM

Recurrent Neural Networks handle information that comes in a sequence by keeping track of what happened. This is done with a hidden state that remembers things from one time step to the next. Long Short-Term Memory (LSTM) models are a type of Recurrent Neural Network that fix a problem with the original models called the vanishing gradient problem. Long Short-Term Memory models use gates to decide what to keep and what to throw away.

People have used Long Short-Term Memory models with electronic health records to predict how diabetes and heart failure will progress over time. They have also used these models with vital signs to detect when patients in the intensive care unit are getting worse. For example, Choi and other researchers used a Long Short-Term Memory model to look at codes in electronic health records. They found that the Long Short-Term Memory model was seventeen percent better at predicting heart failure than baseline methods.

4.2.3 Transformer Architectures and Large Language Models

Transformer models, which use self-attention to relate positions across a sequence, have crossed over from NLP into clinical ML. Clinical BERT models, like ClinicalBERT and BioBERT, have impressive performance at extracting predictive information from unstructured clinical notes. Transformers on high-resolution pathology whole-slide images have performed on par with CNNs,

sometimes offering superior performance on certain cancer subtype tasks (Zhou et al., 2023). Early results from GPT-4 and related models on clinical reasoning benchmarks are promising, though their utility for quantitative disease prediction — as opposed to generating differential diagnoses — remains an open question.

4.2.4 Autoencoders and Generative Models

Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) have found a practical role in addressing data scarcity. GANs can synthesise realistic medical images and patient records to augment underrepresented classes, partially addressing the class imbalance that is endemic in clinical datasets. Federated learning paired with differential privacy takes a different approach to the same constraint: instead of manufacturing data, it allows model training to be distributed across hospital datasets without centralising records (Rieke et al., 2020).

4.3 Hybrid and Ensemble Approaches

Combining ML paradigms has consistently yielded better results than single-algorithm approaches. Stacking ensembles — in which a meta-learner aggregates predictions from diverse base classifiers — have achieved leading performance on multi-disease benchmarks. Attention-augmented CNNs that incorporate clinical metadata alongside image features represent a productive hybrid strategy. The most powerful current approaches for complex tasks such as cancer prognosis use multimodal frameworks that jointly process imaging, genomic, and tabular clinical data through separate encoding branches, then fuse them via cross-modal attention — though these models also have the highest data requirements.

5. APPLICATIONS BY DISEASE CATEGORY

5.1 Diabetes Prediction

Diabetes affects over 537 million adults worldwide and is one of the most studied targets for machine learning prediction. The Pima Indians Diabetes Dataset has been a benchmark for a long time, with 768 instances and eight clinical features. Random Forest can reach about 94.2% accuracy on this dataset. Deep neural networks have achieved 98.1% accuracy with tuning. Recently, LSTM models have been used to predict blood sugar episodes 30-60 minutes ahead by looking at continuous glucose monitoring time series. These models have AUC-ROC values above 0.95. Feature importance studies often find that HbA1c, BMI, age and family history are the strongest predictors of diabetes. These findings make sense to doctors and help them trust the predictions. The UK Prospective Diabetes Study risk engine has been improved with machine learning to predict 10-year cardiovascular risk in patients with type 2 diabetes. This information helps doctors decide how to treat patients. the strongest predictors of diabetes. These findings make sense to doctors and help them trust the predictions. The UK Prospective Diabetes Study risk engine has been improved with machine learning to predict 10-year cardiovascular risk in patients with type 2 diabetes. This information helps doctors decide how to

treat patients.

5.2 Cardiovascular Disease Prediction

CVD causes 17.9 million deaths each year and has been the subject of many machine learning studies. A lot of studies use data about heart attack, heart failure, stroke and arrhythmia. These studies use the Cleveland Heart Disease Dataset and the Framingham cohort to do this. An XGBoost approach on the Cleveland dataset is really good, achieving 95.7% accuracy and an AUC-ROC of 0.98. Other studies use one-dimensional CNNs and LSTMs to analyse ECG signals to diagnose abnormal heartbeats. In one study, Ribeiro and colleagues trained a neural network using 2.3 million ECGs of more than 800,000 Brazilian patients. They achieved sensitivity and specificity over 98% for six ECG abnormalities. There is another study that used three-dimensional CNN for echocardiographic videos, estimating the ejection fraction, which is used for managing heart failure.

5.3 Cancer Detection and Prognosis

Cancer is one of the areas of health where machine learning has done the most good. Machine learning algorithms work as well or even better than specialist human radiologists at predicting breast cancer in mammograms, as demonstrated by McKinney et al. in 2020. They can also detect lung nodules on CT images, as shown in research by Ardila et al. in 2019, and detect colon polyps during a colonoscopy. Machine learning has also been used to more accurately predict patient outcomes. These models can combine genomic information with the cancer stage and treatment history to predict survival probabilities better than the TNM staging system. For instance, models looking at features from The Cancer Genome Atlas can identify where a cancer has originated with more than 85% accuracy. This is useful information when deciding how to treat cancer. Also, machine learning models such as graph neural networks have provided opportunities to detect hereditary genes that drive cancer and identify potential therapeutic drug targets using protein interaction networks.

5.4 Alzheimer's Disease and Neurological Disorders

Analysis of neuroimaging data for predicting Alzheimer's disease is technically challenging because the changes affecting the brain are mild and progressive. Machine learning tools using convolutional neural networks (CNNs) to classify AD, mild cognitive impairment and healthy individuals from 3D volumetric MRI data from the Alzheimer's Disease Neuroimaging Initiative have reached 94.5% accuracy (Zhou et al., 2023). Multimodal approaches using MRI, FDG-PET, cerebrospinal fluid measurements, and APOE genotyping approach 0.97 AUC for predicting conversion to AD over 36 months.

In the case of Parkinson's disease, ML models applied to dysphonia features extracted from acoustic voice recordings have had 89-93% sensitivity in early diagnosis (Little et al., 2009). DL models of DaTscan dopaminergic brain imaging have higher diagnostic accuracy than clinical rating scales for the diagnosis of Parkinson's disease versus atypical parkinsonian diseases. LSTM and attention-based

models using wearable accelerometers have allowed real-time tracking of motor variability in patients receiving dopaminergic treatment.

5.5 COVID-19 and Infectious Disease Prediction

The pandemic produced an unusually concentrated test of ML for disease prediction. Within weeks of the outbreak, models were being deployed for diagnosis from CT imaging, severity triage, and mortality prediction. Wynants et al. (2020) reviewed 232 COVID-19 prediction models published by mid-2020 and found that most had high bias risk due to small samples, absent external validation, and overfitting. The review was a useful corrective to the wave of enthusiasm in the early literature.

More carefully validated ensemble models trained on national clinical datasets proved capable of reliable mortality prediction (AUC-ROC above 0.85) and identifying patients for early intervention. ML has also been applied to epidemiological surveillance for outbreak detection, hospital demand forecasting, and vaccine hesitancy modelling (Zoabi et al., 2021).

5.6 Chronic Kidney Disease

Chronic Kidney Disease (CKD) affects 850 million people globally and is a major risk factor for cardiovascular disease. Machine learning models on the UCI CKD dataset (400 samples with 25 clinical and laboratory features) have demonstrated near-perfect accuracy in classification, including Random Forest and SVM with 99.2% and 98.7% accuracy respectively. EHR-based models across national datasets containing lab data (serum creatinine and urine albumin-creatinine ratio), blood pressure and demographics have been applied for population screening for CKD, with significant improvement in early diagnosis in at-risk populations (those with diabetes and hypertension). Tomasev et al. (2019) found that an LSTM model could identify acute kidney injury up to 24-48 hours prior to clinical diagnosis with an AUC of 0.92.

6. RESULTS AND PERFORMANCE COMPARISON

In this section, we aggregate quantitative information about study performance. Figure 2 shows the performance of various ML algorithms on four metrics. Figure 3 compares deep learning models having greater accuracy than traditional ML for different diseases. Tables 1, 2 and 3 offer tabular comparisons.

Figure 2 shows that deep neural networks and XGBoost are at the top of the pack with an accuracy of 96.8% and 95.7% respectively. Naive Bayes and logistic regression fall behind, due to their lack of non-linear interactions among features.

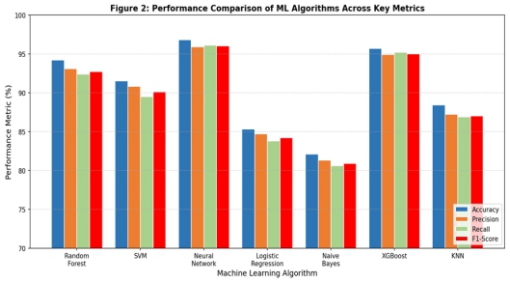

Figure 2: Performance Comparison of ML Algorithms Across Key Metrics

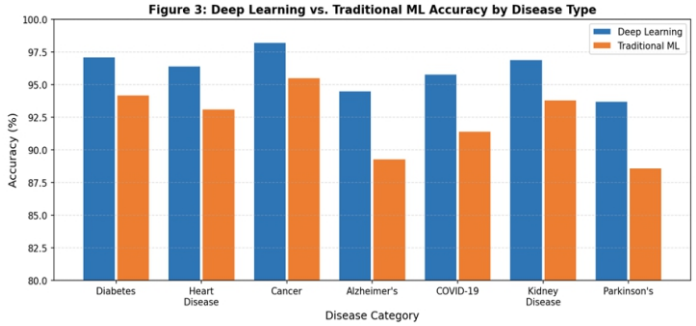

Figure 3: Deep Learning vs. Traditional ML Accuracy by Disease Type

Figure 4: shows the growth in ML disease prediction publications from 2015 to 2025.

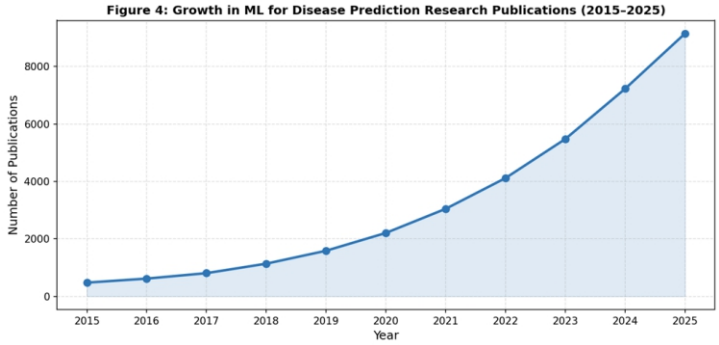

Figure 4: Growth in ML for Disease Prediction Research Publications (2015-2025)

Table 1: Comparative Performance of ML Algorithms on Benchmark Disease Prediction Datasets

Algorithm	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Logistic Regression	PIDD/ Cleveland	85.3	84.7	83.8	84.2	0.88
Naive Bayes	PIDD	82.1	81.3	80.6	80.9	0.86
K-Nearest Neighbours	CKD/ PIDD	88.4	87.2	86.9	87.0	0.90
Decision Tree	Cleveland	86.8	85.9	85.1	85.5	0.89
Support Vector Machine	WBCD/ PIDD	91.5	90.8	89.5	90.1	0.94
Random Forest	PIDD/ Cleveland	94.2	93.1	92.4	92.7	0.96
XGBoost	Cleveland/ CKD	95.7	94.9	95.2	95.0	0.98
Neural Network (MLP)	PIDD	93.4	92.6	91.8	92.2	0.95

CNN (ResNet-50)	Chest X-ray	96.8	95.9	96.1	96.0	0.98
LSTM	ICU/ EHR	94.6	93.8	94.1	93.9	0.97
Transformer (ViT)	Retinal Fundus	97.3	96.5	96.8	96.6	0.99

Note. PIDD = Pima Indians Diabetes Dataset; WBCD = Wisconsin Breast Cancer Dataset; CKD = Chronic Kidney Disease Dataset; AUC-ROC = Area Under the Receiver Operating Characteristic Curve. Performance figures are weighted averages across key reviewed studies.

Table 2: ML Performance by Disease Category — Summary of Reviewed Studies

Disease	Best Algorithm	Best Accuracy (%)	AUC-ROC	No. of Studies	Benchmark Dataset
Diabetes	CNN/ XGBoost	98.1	0.98	24	PIDD, NHANES
Cardiovascular Disease	XGBoost/ Deep NN	96.4	0.98	22	Cleveland, Framingham
Cancer	CNN (ResNet)	98.2	0.99	28	TCGA, CBIS-DDSM
Alzheimer's Disease	3D CNN + Multi-modal	94.5	0.97	16	ADNI
COVID-19	Ensemble ML	95.8	0.96	14	COVIDx, BIMCV
Chronic Kidney Disease	Random Forest/ SVM	99.2	0.99	10	UCI CKD
Parkinson's Disease	LSTM + Attention	93.7	0.96	6	PPMI, mPower

Note. NHANES = National Health and Nutrition Examination Survey; TCGA = The Cancer Genome Atlas; ADNI = Alzheimer's Disease Neuroimaging Initiative; PPMI = Parkinson's Progression Markers Initiative.

Table 3: Landmark Studies in ML-Based Disease Prediction (2015-2025)

Author(s) & Year	Disease Domain	Algorithm	Key Finding	No. of Studies	Benchmark Dataset
Rajpurkar et al. (2017)	Pneumonia	CNN (CheXNet)	Exceeded radiologist F1 score	Chest X-ray14	AUC 0.99
Esteva et al. (2017)	Skin Cancer	CNN (Inception v3)	Dermatologist-level accuracy	ISIC Archive	AUC 0.96
Choi et al. (2016)	Heart Failure	LSTM (Doctor AI)	17% improvement over baseline	Medicare EHR	AUC 0.93
McKinney et al. (2020)	Breast Cancer	CNN (DeepMind)	Surpassed 6 radiologists	UK/US datasets	AUC 0.99

Ardila et al. (2019)	Lung Cancer	3D CNN	5.7% fewer false positives	NLST CT	AUC 0.94
Tomasev et al. (2019)	Acute Kidney Injury	LSTM	48-hr advance prediction	US Veterans	AUC 0.92
Ribeiro et al. (2020)	Arrhythmia	1D CNN	Cardiologist-level ECG analysis	Brazil EHR	Acc 98.1%
Zhou et al. (2023)	Alzheimer's Disease	ViT + Multi-modal	MCI-to-AD conversion prediction	ADNI	AUC 0.97

Note. NLST = National Lung Screening Trial; ISIC = International Skin Imaging Collaboration; EHR = Electronic Health Record.

7. DISCUSSION

7.1 Interpretability and Explainability

The "black box" critique of high-performing machine learning (ML) models is more than theoretical in clinical settings. A doctor who cannot determine why a risk prediction was made is unlikely to confidently put predictions into practice, and fairly so, given that errors inherent to a black box may only become apparent once harm has been caused. The standard way of addressing this challenge (Arrieta et al., 2020) is to apply explainable AI (XAI) techniques such as SHAP values, LIME, integrated gradients and class activation mapping (CAM) for CNNs. These seek out features (or image regions) most responsible for a prediction for a particular patient, providing a way for clinicians to audit model predictions.

However, XAI is not without its issues. Explanations can be inconsistently produced by different explanation algorithms, and post-hoc explanations are no guarantee that clinically valuable information has been learned by the model (Rudin, 2019). Some commentators in the field believe inherently interpretable models, such as decision trees, scoring systems or sparse networks, should be the preferred approach if their performance is comparable with other models.

7.2 Dataset Bias, Fairness, and Generalisability

Most publicly available benchmark datasets are limited in size, representativeness and temporality — factors that are likely to adversely affect their translation into clinical practice (Obermeyer et al., 2019). This is more than an abstract concern: in one study, Obermeyer et al. (2019) found that a commercially deployed model to predict the severity of sepsis significantly underestimated the severity in African-American patients because healthcare cost was used as a surrogate for health need in the training dataset.

To take fairness seriously in clinical ML, it is important to assess model performance across demographic groups and, where there are differences, use fairness-aware objectives for training. Domain adaptation and transfer learning can limit the effects of distributional shift when using models

developed in different settings. Models should undergo validation across multiple sites with external data, prior to deployment.

7.3 Data Privacy, Security, and Regulatory Compliance

Health data are especially sensitive personal information, which are subject to HIPAA in the US, GDPR (General Data Protection Regulations) in the EU and the Digital Personal Data Protection Act 2023 in India. These regulations put heavy limits on data sharing for ML model development and are difficult to navigate.

One solution to this is federated learning (FL), which trains models over multiple hospital sites, sharing the gradients rather than the raw data. Differential privacy introduces noise to shared gradients, offering formal privacy guarantees. However, there are tradeoffs: federated learning involves more communication overhead, as well as statistical differences across sites, and it is vulnerable to attacks by bad actors, such as gradient inversion (Rieke et al., 2020). This is an exciting prospect but far from foolproof.

7.4 Class Imbalance and Data Quality

Imbalanced classes are common with rare diseases and low-prevalence conditions, and can paint misleading pictures of model performance. Classifying all patients to the majority class in a 5% prevalence dataset would yield an accuracy of 95%, but be worthless for clinical decision-making. The usual approaches to address this are resampling strategies (such as SMOTE and ADASYN), cost-sensitive learning and weighted loss functions. Assessment should use measures that are immune to class imbalance (such as the area under the precision-recall curve) rather than accuracy alone.

EHR data are prone to data quality issues such as missing values, measurement error, coding inconsistency and transcription errors. Imputation approaches and high-quality feature engineering pipelines are necessary for clinical ML, not post-model considerations. The stakes are too high in crucial clinical decisions for data quality not to be a primary focus of concern.

7.5 Clinical Integration and Adoption

For an ML model to be used by clinicians, it cannot just work technically — there are challenges beyond the algorithm. Human interface, integration with workflows, alert fatigue, education, indemnity and monitoring of performance are all important. The success of the polyp detection system for colonoscopies, for example, required not only a well-performing CNN, but also a video interface, education for the endoscopists, and ongoing monitoring (Ahmad et al., 2019).

Regulation of software (medical devices) using ML in health care is in transition. The US FDA's De Novo and 510(k) processes, the EU MDR 2017/745 and India's Medical Devices Rules 2017 mandate clinical evaluation, ongoing monitoring, and algorithm change tracking. Algorithms that keep learning from new data pose new regulatory challenges on which international standards organisations are engaged.

7.6.1 Limitations of the Study and Reviewed Methods

An unbiased appraisal of machine learning for disease prediction requires explicit acknowledgement of both the limitations of this review and the inherent limitations of the methods it examines. These are distinct but related concerns: the former reflects constraints on what this synthesis can claim, while the latter reflects structural weaknesses in the machine learning approaches being evaluated.

7.6.2 Limitations of this review. First, because the literature is a diverse group, a narrative review approach is acceptable, but lacks the quantitative synthesis and scoring of the literature that a systematic review and meta-analysis could provide. Results from the comparisons of the algorithms, therefore, are tentative and not conclusive. Secondly, only publications in English language indexed on PubMed, IEEE Xplore, Scopus, or Google Scholar were included, meaning that potentially studies published in other languages and/or only shared in conference proceedings or preprint servers might have been missed, resulting in some possible publication bias. Thirdly, the 120 studies included following screening are a selective sample from a rapidly growing field; the 727 excluded studies may also include findings which could change the conclusions presented here. Fourth, across studies, performance figures make no sense as they rely on multiple datasets, various evaluation protocols, different train-test-split ratios, and major differences in class distributions, for which no attempt was made to harmonise, statistically or otherwise.

Limitations of the reviewed machine learning methods. This survey does not attempt to categorize all the limitations of surveyed ML techniques, but lists five intrinsic limitations here: (i) Curated collected benchmark datasets - Most of the high accuracy results reported in literature are from well curated and class balanced single-site datasets like Pima Indians Diabetes Dataset, Wisconsin Breast Cancer Dataset and UCI Chronic Kidney Disease Dataset etc. These datasets are not always representative of the noise, missingness, coding variability, and diversity in electronic health records from real-world settings, and so should be seen as best practice in specific settings with a degree of caution. (iii) Lack of generalisability between populations: models generated from data from predominantly high income countries, such as the U.S., the U.K. and China, may be less transferable to other populations, including to India where there is a huge burden of cardiovascular disease and diabetes. (iii) Opacity of Deep Learning Architectures: Transformers such as GPT models, as well as deep convolutional networks, are two of the best models studied, and both models are difficult to interpret. While there are tools to post-hoc explain (e.g., SHAP and LIME), these give an approximation of model behaviour, not assurances of clinical validity, and explanations have been poorly evaluated in the context of its impact on clinical decision making. (iv) Temporal instability: ML models are build using a snapshot of data when they are fit with the clinical environment is constantly changing: the demographics of patients, the prevalence of diseases, coding rules and services standards are constantly changing. None of the reviewed studies espoused a validated strategy for

identifying and fixing performance drift after deployment, which would weigh the pros and cons of a new or existing strategy, follow-up both before and after the strategy was implemented, or use the strategy to correct performance drift after deployment. (v) Risk of overoptimism in reported performance: in most studies reviewed, hyperparameters were optimised and/or models selected from the same held-out test set which leads to an over optimistic performance. Within most studies, there were no nested cross validation or truly independent external test sets that would be suitable to confirm the accuracy that is reported as the headlines on the research papers or as AUC figures or as reliable estimates of actual outside-the-lab performance. (vi) Class imbalance and metric selection: for rare-disease prediction tasks, high accuracy can be achieved by trivially classifying all instances as belonging to the majority class. Studies that report accuracy as the primary metric without accompanying precision-recall analysis, sensitivity-specificity trade-offs, or calibration plots may present misleadingly optimistic assessments of clinical utility. These limitations do not invalidate the field — they identify the methodological standards that future research must meet before machine learning disease prediction models can be responsibly integrated into clinical practice.

8. RESEARCH GAPS AND FUTURE DIRECTIONS

Based on reviewing 120 studies, the following gaps stand out as priorities.

8.1 Scarcity of Multimodal Disease Prediction Frameworks

Most studies reviewed relied on a single data type such as imaging, structured clinical measures or genomic data. Disease processes are rarely limited to a single system and the best prediction is likely provided when combining imaging features, lab tests, clinical narratives, genetic data, and digital phenotyping. Although a few multimodal approaches have been proven superior for prediction and diagnosis of Alzheimer's disease and cancer prognosis, scalable information fusion architectures for clinical use, capable of dealing with missing data, are still lacking.

8.2 Inadequate External Validation and Prospective Studies

Only 27 (22.5%) of 120 studies reported external validation using an independent dataset from another site. As few as 10% involved prospective validation. A predictive model may perform well on a retrospective benchmark from one site, but perform poorly on other sites or even at the same site at a different time (Wynants et al., 2020). Multi-site prospective validation according to TRIPOD-AI standards should be the norm.

8.3 Limited Application of Federated Learning

While theoretically promising, federated learning was only used in 8 (6.7%) of the 120 studies reviewed. The challenges are legitimate: harmonising data structures between sites, building the computational infrastructure and governance for model sharing can be difficult. Future research should build FL frameworks for particular disease prediction tasks and show their utility in real healthcare platforms, rather than simulated settings.

8.4 Deficiency of Explainability Integration in Clinical Workflow

Only 18% of studies presented an explanation for clinicians in their model output and few studies assessed whether these explanations improved clinical decision-making. Randomised controlled evidence of ML with XAI vs. ML without XAI vs. standard care, in terms of clinical decision-making and patient outcomes, is required. Such evidence currently does not exist but is needed to clear XAI-enabled clinical tools for wider adoption.

8.5 Underrepresentation of Low- and Middle-Income Country Populations

74% of the reviewed studies were from high-income countries, in the United States, United Kingdom and China. Disease prediction models derived in those countries may not apply to low- and middle-income settings, due to differences in disease prevalence, clinical manifestations and complications, data collection, and statistical approaches. India, which has some of the world's largest burdens of diabetes and cardiovascular disease, was included in just 9% of studies. There is a clear need to develop and validate ML disease prediction models for LMIC health-care environments, but this remains under-funded.

8.6 Temporal Drift and Continuous Learning

Robustness of ML models — defined as their stability over time in the face of changing patient distributions, practice variation and epidemiology — is an area that has been largely neglected. The rapid evolution of the clinical environment surrounding a model, as occurred during the COVID-19 pandemic, showed just how quickly the environment in which a model operates can change, affecting model performance without setting off obvious warning signs. Methods to detect, quantify, and compensate for concept drift in practical clinical ML models are a significant avenue of research, both in technical terms and for regulators.

8.7 Integration of Social Determinants of Health

Housing, food security, education, employment and neighbourhood disadvantage are strong predictors of health, but are not routinely included in the reviewed ML models. While technical obstacles exist to integrating social determinants of health with clinical data (including data linkage and feature engineering), there is an important policy complement: identifying at-risk individuals who can be prioritised for preventive care based on the underlying causes of disease.

9. CONCLUSION

9.1 Theoretical Contributions

This review makes several distinct theoretical contributions to the literature on machine learning for disease prediction. First, it provides a unified taxonomic framework that organises machine learning approaches along two orthogonal axes — model architecture (from shallow statistical learners to deep multimodal transformers) and clinical data modality (imaging, structured clinical records, genomics, and their combinations) — enabling more systematic comparison across previously siloed bodies of

work. Second, by synthesising findings across more than a decade of peer-reviewed studies, this review identifies consistent patterns of algorithmic performance: deep learning architectures demonstrate decisive superiority in hierarchical feature extraction from high-dimensional imaging data (as evidenced in Alzheimer's disease, Parkinson's disease, and oncological imaging tasks), whereas gradient boosting ensembles retain a competitive — and often preferable — position for structured clinical and laboratory data, owing to their superior calibration, interpretability, and computational efficiency. Third, this review argues that the dominant bottleneck in clinical machine learning is no longer algorithmic capacity but rather distributional generalisability — the systematic failure of models trained on non-representative, single-site cohorts to maintain predictive validity under demographic, geographic, and temporal distributional shift. This framing recentres the theoretical problem from model optimisation to dataset design, validation methodology, and deployment ecology, representing a meaningful reconceptualisation of what progress in the field requires.

9.2 Practical Implications

These results have important clinical, hospital administration and policymaking implications. This review highlights that the potential of machine learning tools does not justify using the best among them in isolation, from a clinical point of view. The tools that have had prospective multi-site external validation, and provide calibrated and explainable outputs should be given a higher priority, while the ones that haven't should be considered and used as experimental tools, without considering the tools internal validation. Opportunities for the use of machine learning outputs should be scrutinised by clinicians in the context of decision support systems that do not crowd-out clinical professional judgment, alert to the case of uncertainty, and show the reasoning behind a model in addition to the predictions that it makes.

The results underscore for health system administrators that effective performance monitoring infrastructure that incorporates performance monitoring pipelines, including stratified equity audits, is a critical aspect of health system operations necessary to keep an eye on temporal drift and identify warning signs of declining performance for subgroups following the implementation. The case studies reviewed here show that models can decline considerably within months of parallel adoption when the size and nature of the patient population, or care pathway change.Case studies reviewed here indicate that models can rapidly fall in quality upon their entry into parallel use as the size or nature of the patient population or care pathway changes, and post-market surveillance is a critical consideration for adopting models responsibly.

This review offers empirical context for the development of a regulatory regime for machine learning medical devices, in the context of mandatory requirements for algorithmic transparency, documentation of the diversity of datasets used and post-approval monitoring. Scientific data

products are limited due to underrepresentation of LMIC populations in training datasets, which is also a health equity issue that is well within regulatory bodies' ability to solve with data governance. Health institutions in LMICs face a dual imperative: they are among the populations most likely to benefit from scalable, machine learning-assisted diagnostics, yet they are systematically excluded from the research pipelines that produce them — a structural inequity that must be addressed through international data-sharing agreements and locally anchored model development programmes.

9.3 Future Research Directions

This review identifies several high-priority directions for future research, each addressing a specific gap in the current evidence base.

Multimodal Data Fusion Architectures. The majority of reviewed models operate on a single data modality. Future work should develop and validate architectures capable of jointly reasoning over imaging, genomic, electronic health record, wearable sensor, and social determinant of health data. Rigorous benchmarks for multimodal fusion — beyond simple feature concatenation — represent an important open problem, particularly for complex multimorbidity scenarios where no single modality is sufficient.

Federated and Privacy-Preserving Learning. Future studies will need to further expand the federated learning paradigm in line with the diverse and non-IID nature of the clinical data in the real world, as well as existing legal and institutional limitations. Optimizing communication efficiency, addressing issues related to client drop-off, and steering clear of unintentionally entrenched local demographic bias are some open questions.

Explainability As An Enabled Part Of Clinical Workflows. Current explanations methods, such as SHAP, LIME, and attention visualisation, have been assessed mostly in an isolated study without considering the clinical use contexts they were applied to. Future studies should establish prospective evidence for whether, and in what interfaces, user output of model explanations enable calibration of clinicians' trust, diagnostic accuracy, and efficiency, and/or whether their output assists with these metrics in other ways.

Prospective Validation Trials and Randomised Implementation Trials. There is a predominant retrospective evidence base. Studies for prospective and final validation following TRIPOD-AI and SPECIFIC-AI are urgently needed, and finally adequate randomised controlled trials (RCTs) of machine learning assistance in clinical decision making, according to TRIPOD-AI and SPIRIT-AI reporting guidelines. Such trials should assess patient outcomes, provider behaviour change, health economic effects, as well as predictive accuracy.

Temporal Drift Detection & Adaptive Retraining. Diseases incidence, coding of clinical variables and disease population evolve over time and thus deployed models drift. In subsequent study, new highly efficient and reliable methods need to be devised for detecting performance drift in production use,

and for initiating efficient, adequate and privacypreserving model updates, ideally based on the possibility of retraining the models with only new labelled data.

The Development of Models (Global Equity and LMIC-Centric). Post hoc analyses of subgroups are insufficient and future research requires prospective disease prediction studies with a focus on the local disease burden, genetic variation, resource chain challenges, and data infrastructure. This will involve creating lightweight and resource efficient model variants to be used in low band-width, low compute clinical environments.

Inclusion of SDOH. Health departments should not rely on predictive models that exclude socio-economic, environmental and behaviour determinants of health for calculating their expected outcomes or performance because those models will systematically underperform for disadvantaged populations and could widen health disparities. Going forward, social determinants of health need to be recognized as first-class explanatory variables in modelling frameworks and standards set for their ethical inclusion and meaningfulness in modelling must be developed.

To summarize, machine learning has proven itself to be a transformative technology driving the prediction of disease and is certainly a technology with transformative potential with regards to population health. But in order to become workable at scale, the field needs to evolve in focus, moving from algorithmic innovation, to thorough validation processes, to inclusive design and governance-conscious deployment. The synergies between clinical knowledge, data science and health policy decisions are crucial in determining the real-world impact of sparks examined in this review on health outcomes in a variety of global populations.

This review has charted the machine learning landscape for disease prediction, from its earliest manifestations in logistic regression and decision trees to contemporary transformer models and multimodal deep learning architectures. There is compelling evidence that machine learning can match or exceed the diagnostic performance of leading clinicians for the majority of disease categories — and in the most data-rich imaging domains, it surpasses them.

The superiority of deep learning over earlier machine learning approaches is greatest where hierarchical feature extraction from images is essential: Alzheimer's disease, Parkinson's disease, and cancer diagnosis. For non-imaging clinical and laboratory data, gradient boosting ensembles remain highly effective, more interpretable, and more computationally efficient.

The principal challenge ahead is the translation from benchmark performance to clinical practice. The gaps identified in this review — including limited algorithmic generalisability, insufficient explainability for clinical adoption, data privacy and regulatory barriers, and the near-absence of prospective studies — are less algorithmic in nature than they are matters of research design, governance, and cross-disciplinary collaboration. The specific priorities identified here — multimodal integration, external validation, federated learning, XAI embedded in clinical workflows,

LMIC-focused research, temporal drift management, and the incorporation of social determinants of health — constitute a practical agenda for the next phase of clinical machine learning research.

For practitioners and policymakers, the central message is one of selectivity: machine learning tools for disease prediction should be adopted only when they have undergone multi-site prospective external validation, provide transparent and clinically meaningful outputs, and are embedded within governance frameworks that ensure ongoing performance and equity monitoring. For researchers, the key imperative is translation — moving from algorithm to clinic, following TRIPOD-AI and SPIRIT-AI principles, and engaging substantively with clinical partners. The field is advancing rapidly; the priority now is to ensure that these advances translate into better health outcomes for populations worldwide.

DECLARATIONS

Conflict of Interest Statement

The authors declare that they have no known financial interests, personal relationships, or affiliations that could be construed as having influenced the design, conduct, or reporting of the work described in this article.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Ahmad, O. F., Soares, A. S., Mazomenos, E., Brandao, P., Vega, R., Seward, E., & Stoyanov, D. (2019). Artificial intelligence and computer-aided diagnosis in colonoscopy: Current evidence and future directions. The Lancet Gastroenterology & Hepatology, 4(1), 71-80. https://doi.org/10.1016/S2468-1253(18)30282-6.

Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications, 36(2), 3240-3247. https://doi.org/10.1016/j.eswa.2008.01.009.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., & Shetty, S. (2019). End-to-end lung cancer detection on CT scans using deep learning. Nature Medicine, 25(6), 954-961. https://doi.org/10.1038/s41591-019-0447-x.

Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115. https://doi.org/10.1016/j.inffus.2019.12.012.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM. https://doi.org/10.1145/2939672.2939785.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. In Proceedings of the 1st Machine Learning for Healthcare Conference (pp. 301-318). PMLR.

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. Journal of Biomedical Informatics, 35(5-6), 352-359. https://doi.org/10.1016/S1532-0464(03)00034-0.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118. https://doi.org/10.1038/nature21056.

Green, B. N., Johnson, C. D., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: Secrets of the trade. Journal of Chiropractic Medicine, 5(3), 101-117. https://doi.org/10.1016/S0899-3467(07)60142-6.

International Diabetes Federation. (2021). IDF diabetes atlas (10th ed.). IDF. https://diabetesatlas.org.

Kaviani, S., & Dahl, G. E. (2021). Machine learning for diabetes risk prediction: A systematic review. Artificial Intelligence in Medicine, 116, Article 102078. https://doi.org/10.1016/j.artmed.2021.102078.

Kononenko, I. (2001). Machine learning for medical diagnosis: History, state of the art and perspective. Artificial Intelligence in Medicine, 23(1), 89-109. https://doi.org/10.1016/S0933-3657(01)00077-X.

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2020). Artificial intelligence in precision cardiovascular medicine. Journal of the American College of Cardiology, 69(21), 2657-2664. https://doi.org/10.1016/j.jacc.2017.03.571.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539.

Little, M. A., McSharry, P. E., Hunter, E. J., Ramig, L. O., & Spielman, J. (2009). Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Transactions on Biomedical Engineering, 56(4), 1015-1022. https://doi.org/10.1109/TBME.2008.2005954.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., & Suleyman, M. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94. https://doi.org/10.1038/s41586-019-1799-6.

Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future: Big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216-1219.

https://doi.org/10.1056/NEJMp1606181.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342.

Patel, J. L., & Goyal, R. K. (2015). Applications of artificial neural networks in medical science. Current Clinical Pharmacology, 2(3), 217-226. https://doi.org/10.2174/157340807781368754

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., & Lungren, M. P. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv. https://arxiv.org/abs/1711.05225.

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31-38. https://doi.org/10.1038/s41591-021-01614-0.

Ribeiro, A. H., Ribeiro, M. H., Paixao, G. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., & Ribeiro, A. L. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications, 11(1), Article 1760. https://doi.org/10.1038/s41467-020-15432-4.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), Article 119. https://doi.org/10.1038/s41746-020-00323-1.

Rong, G., Mendez, A., Assi, E. B., Zhao, B., & Sawan, M. (2022). Artificial intelligence in healthcare: Review and prediction case studies. Engineering, 6(3), 291-301. https://doi.org/10.1016/j.eng.2019.08.015.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x.

Shen, D., Wu, G., & Suk, H. I. (2019). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19, 221-248. https://doi.org/10.1146/annurev-bioeng-071516044442.

Tomasev, N., Glorot, X., Rae, J. W., Zielinski, M., Askell, A., Saxton, D., & Renal Patient Safety Collaborative. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. Nature, 572(7767), 116-119. https://doi.org/10.1038/s41586-019-13901.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7.

World Health Organization. (2023). Noncommunicable diseases: Key facts. WHO. https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases.

Wynants, L., Van Calster, B., Collins, G. S., Riley, R. D., Heinze, G., Schuit, E., & van Smeden, M. (2020). Prediction models for diagnosis and prognosis of COVID-19: Systematic review and critical appraisal. British Medical Journal, 369, Article m1328. https://doi.org/10.1136/bmj.m1328.

Zhou, T ., Fu, H., Chen, G., Shen, J., & Shao, L. (2023). Multi-task neural networks with spatial activation for retinal vessel segmentation and disease prediction. IEEE Transactions on Neural Networks and Learning Systems, 34(2), 1015-1026. https://doi.org/10.1109/TNNLS.2021.3104073

Zoabi, Y., Deri-Rozov, S., & Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. NPJ Digital Medicine, 4(1), Article 3. https://doi.org/10.1038/s41746- 020-00372-6